Listening. Learning. Leading.®

Research Report
ETS RR–11-32

Computer-Adaptive Testing for Students With Disabilities: A Review of the Literature

Elizabeth Stone

Tim Davey

August 2011

Computer-Adaptive Testing for Students With Disabilities: A Review of the Literature

Elizabeth Stone and Tim Davey
ETS, Princeton, New Jersey

August 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Daniel Eignor

Technical Reviewers: Frederic Robin and Cara Cahalan Laitusis

Abstract

There has been an increased interest in developing computer-adaptive testing (CAT) and multistage assessments for K-12 accountability assessments. The move to adaptive testing has been met with some resistance by those in the field of special education who express concern about routing of students with divergent profiles (e.g., some students with math-based learning disabilities may have difficulty with basic computation but not high level problem solving) and poor performance on early test questions. This paper consists of a literature review focusing on adaptive testing issues for students with disabilities in the K-12 sector. While it is clear that there are issues that will present obstacles to administering accountability tests adaptively to students with disabilities, this synthesis of research and policy developments with respect to this topic will be useful both for development of research agendas and to inform states that are currently using or are considering moving to CAT.

Keywords: accountability, computer-adaptive testing, CAT, disabilities, K-12

Table of Contents

A Brief Background on CAT 3
Who is Using CAT? Who Plans to Use CAT? 5
Challenges of CBTs and CATs Specific to Students With Disabilities 10
Administration Issues 10
Technical Issues 11
Utility Issues 15
Appropriateness Issues 15
Possible Ways to Address These Issues, and Considerations for Future Work 15
References 18

Accountability testing in the K-12 sector is in the process of undergoing a dramatic change. The Elementary and Secondary Education Act of 1965 and its successors require states to evaluate student proficiency and identify achievement deficits (primarily at the sub-group, school, and district levels) via approved assessments aligned with specific content-based standards. Policies related to the use of standardized testing for accountability have been criticized, particularly in the context of evaluating the proficiency of students with disabilities. Often, the criticisms have focused on exclusion of students with disabilities from state testing and the inappropriate difficulty level of the general test for some students. Students with disabilities typically have individualized education programs (IEPs) that describe under what specific conditions a particular student will be assessed. These conditions may involve accommodations on the general test form or administration of an alternate assessment of below-grade-level content (National Center for Learning Disabilities, 2009). Other approaches to assessing the proficiency and academic progress of students with disabilities have included growth models (Buzick & Laitusis, 2010) and a modified (2%) assessment that measures grade-level content in a modified format (Burling, 2007). For example, modified assessments may have a reduced number of distractors per item and more "white space" (i.e., less density of text) on each page.

It has been argued that many state tests do not appropriately measure the proficiency of students with disabilities. Achievement gaps between these and other students would correspondingly be misrepresented as well. These arguments can be summarized as follows: First, state tests are not often deliberately assembled in accordance with principles of universal design (Johnstone, Altman, & Thurlow, 2006) that incorporate accessibility into the test development process from the start. Although improvements have been made in this area, it remains the case that tests are most often made accessible to students with disabilities retrospectively, through the use of accommodations that may affect measurement of the construct and alter the meaning of resulting test scores. Second, a significantly higher proportion of students with learning disabilities fall into the lower tail of the proficiency distribution. Because state tests are usually designed to provide the most accurate measurement in the midrange of the ability scale, where the bulk of students are located, students with proficiencies on the low end of the scale may have no alternative but to guess randomly on many questions (Abedi, Leon, & Kao, 2007; Minnema, Thurlow, Bielinski, & Scott, 2000). Empirical evidence

of this was provided by Laitusis, Buzick, Cook, and Stone (2011, p. 293), in which the percentage of students with learning disabilities scoring at chance level on 4th- and 8th-grade mathematics and English-language arts assessments from one state ranged from 12–22%, while the percentage for students without disabilities ranged from 1–3% on the same tests. An assessment for which this amount of guessing exists for a subgroup of students provides inadequate information about proficiency for those students, and may induce anxiety in and decrease the engagement of students who are unable to demonstrate proficiency.

In addition, established fairness measures may not be as appropriate or useful for students with disabilities. For example, differential item functioning (DIF) methods may not be able to identify problematic items as successfully because they require that differentially functioning items be outliers rather than make up a large portion of the items on the test (i.e., that the total test score be relatively DIF-free). When there are construct-irrelevant factors preventing a student from demonstrating his true level of proficiency with respect to test items—a scenario that can result in a test score that has an altered meaning—the identification of items displaying DIF may be impeded. Comparing two groups with very different ability distributions (e.g., students without disabilities and students with disabilities) often leads to the statistical exclusion of the students in the tails, leading to even smaller effective sample sizes and little information about the students with disabilities who often are the direct focus of the study.

One way to address the misfit of test items to students in the tails of the proficiency distribution is to adapt the test to their proficiency levels. This would involve, at particular points in the test, selecting future items or item sets based on previous performance. The simplest scenario would be to include one adaptive point in the test, creating what is usually referred to as a *two-stage test*. In many cases, the first set of items functions as a routing test and determines to which second-stage set of items the test-taker is routed. The second-stage tests have different ranges of difficulty, and test takers scoring highly at the first stage are routed to a second-stage test of relatively high difficulty. Adding additional adaptive points leads to additional stages and a general multistage testing paradigm (see, e.g., Hendrickson, 2007). If this process is continued until the test adapts after every item, item-level adaptive testing is the result. Although adaptive testing can be performed on paper-based tests, it is often (especially in the case of item-level adaptive testing) implemented in a computer-based test setting and is referred to as a *computer-adaptive test* or *computerized adaptive test* (CAT). It is item-level CAT that we

consider for our discussion here. Other reviews have dealt thoroughly with the general feasibility of using item-level CAT in K12 testing (see, e.g., Accountability and Curriculum Reform Effort [ACRE], 2010; Data Recognition Corporation [DRC], 2007; Peterson, 2005; Reckase, 2011; REL West/WestEd, 2008; Way, 2005; Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010); this review focuses more directly on issues related to test takers, particularly test takers with disabilities.

This study consists of a review of the existing literature on CAT for students with disabilities to address the following questions:

- What are the potential psychometric and practical benefits of using CAT to assess students with disabilities in accountability testing?
- What are the potential psychometric and practical drawbacks of using CAT to assess students with disabilities in accountability testing?
- How can the potential drawbacks be addressed to ensure fair and valid assessments for all students?

In this review of the literature on CAT for students with disabilities (SWD), we begin by outlining the history of CAT and describing the procedure in a general technical sense. We then consider who is currently using CAT and who plans to use it in the future for K-12 testing. We discuss some of the concerns that have been or might be raised, as well as potential advantages, for the use of CAT to test SWD. Finally, we suggest ways to address these issues and discuss future research to be done in this important area.

A Brief Background on CAT

Adaptive testing is not a new idea. The idea of tailoring questions to the responder is a key element of oral comprehensive exams and intelligence tests (van der Linden & Glas, 2010). However, widespread adoption of CAT would not have been practical without two developments: the use of computers in testing (with improved processing capabilities and speed) and item response theory (IRT) procedures. While the requirement of a computer to individualize test administration to each examinee is fairly clear, the importance of IRT cannot be overestimated as a way of providing comparable scores for students who have taken very different tests (thus eliminating the need for post-test equating to deal with tests of varying

difficulty). Whereas classical test theory conflates item difficulty with test taker proficiency, IRT allows item characteristics and test taker characteristics to be parameterized separately. Because of this separation, actions on items (e.g., item parameter calibration, item selection) can take proficiency into account, and actions on test takers (e.g., ability estimation, scoring) can take item characteristics into account. One of the most commonly used IRT models is the three-parameter logistic model (3PL; Lord, 1980, p. 12). This model has separate parameters for the item difficulty, discrimination (how well it separates higher-ability examinees from lower-ability examinees in terms of answering the item correctly), and pseudo-guessing probability, which takes into account the idea that for multiple-choice items there is a probability greater than zero of answering the item correctly due to chance. The parameters define an item characteristic curve (ICC) which relates the probability of answering an item correctly to an ability parameter θ on a continuous scale. Each item has its own ICC, and summing the ICCs in a test produces the test characteristic curve (and the number-right true score).

Adaptive testing uses IRT in a variety of capacities. Beginning with a reasonable ability estimate (usually mid-range or based on preliminary data), item selection methods attempt to capitalize on the information about the examinee ability estimate that a potential item provides. Being highly informative means that the item information function is strongly peaked at a particular point on the ability scale; in other words, the item provides a lot of information at that point but less information at other points on the ability scale. Information is a function that is directly related to the square of the discrimination parameter; however, highly informative items are valuable assets and would be over-selected and over-exposed if no additional constraints were added to the selection process. In addition, it does not necessarily make sense to select a highly informative item at the beginning of the test where the error in the ability estimate is large. After each item is responded to by the test taker, a new interim ability estimate is calculated. Based on the updated ability estimate and any other constraints imposed on the test (e.g., content specifications that must be met, item exposure controls, or sets of items that are prohibited from being administered together), the next item is chosen. This process continues until a fixed number of items have been administered or a convergence criterion for the ability estimate has been reached, at which point a final ability estimate is computed for use in scoring. See van der Linden and Pashley (2010) for a more detailed description of adaptive test item selection and ability estimation.

Who is Using CAT? Who Plans to Use CAT?

Recently, there has been increased movement toward computerized testing in general (CTB McGraw-Hill, 2003, Scheuermann & Björnsson, 2006;) and toward the use of CAT in the K-12 sector specifically. According to a questionnaire sent to all state education departments to collect information for a study examining the feasibility of moving South Carolina's state test to computer (DRC, 2007), of the states who replied only Idaho, Oregon, and South Dakota used adaptive tests at the time, although not necessarily for accountability purposes. Delaware noted at the time of the survey that their forthcoming Delaware Comprehensive Assessment System (DCAS) request for proposals (RFP) would be geared toward computer-based testing. When the RFP was released in 2009, it contained a requirement that the reading and mathematics summative assessments for NCLB in grades 3–8 be computer-adaptive and web-based. The science (grades 5 and 8) and social studies (grades 4 and 7) portions could consist of multiple fixed forms initially, but the goal was for adaptive tests in those areas as well. The proposed DCAS system would also include (possibly adaptive) benchmark and end-of-course (EOC) assessments. The move toward adaptive assessments was designed to "produce the most precise estimate of student achievement and growth, and greater detail in diagnostic feedback" (p. 14). Minnesota noted in response to the survey that "[u]sing computer-adaptive testing for statewide accountability testing is problematic, although there has been some push for this" (p. 1–51). North Carolina wondered, "[i]f you do CAT, are you going item-wise or with 'chunks' of items? We tried chunks (boil/freeze CAT) and saw no increase in reliability or decrease in [standard error of measurement] as would have been theoretically expected" (p. 1–57). In fact, North Carolina had implemented CAT tests in 2000–01 (prior to NCLB) with the express goal of providing an accommodated test for SWD, but had to discontinue their use of them because of the requirement that NCLB test items be on grade level (ACRE, 2010). Idaho and South Dakota had similar issues with meeting grade-level requirements that led them to move to a hybrid test (fixed for accountability purposes, adaptive for fine-tuning ability estimation), and a fixed form (with a voluntary adaptive test), respectively (Olson, 2003).

In 2008, the North Carolina State Board of Education released a framework of recommendations for moving the state toward *next generation* assessments. The report that resulted from that impetus (ACRE, 2010) categorizes the advantages to using CAT as relating to administration, technicality, utility, and appropriateness. It should be noted that many of these

advantages are true of computer-based tests (CBTs) in general. The four categories are discussed here.

Administration. CATs can be administered in less time, primarily because fewer items need be administered than in a conventional test to achieve the same measurement precision. Way (2010) notes that CATs that use the one-parameter Rasch model can be about 20% shorter, and 3PL CATs can be 40–50% shorter. In addition, if test termination is based on measurement precision (i.e., a variable-length, rather than fixed-length, test), increased efficiency can be attained. Because the tests are administered via computer, there is little in the way of staff time or resources needed to prepare and ship testing materials. Tests can be scheduled more flexibly, and test security is enhanced both by having the test stored on computer rather than on paper (Wainer, 2000, p. 11) and by administering different items to different test takers (Thompson, 2010).

Technicality. As mentioned previously, one major benefit associated with adaptive testing is the ability to more precisely target test-item difficulty to estimated test-taker ability. Kingsbury and Hauser (2004) compared test information and measures of score accuracy and classification between fixed and adaptive forms of 4th and 8th grade reading and mathematics tests and noted that information at the extremes of the ability distribution was three times greater in the adaptive setting. ACRE (2010) includes an excerpt from Alpert (2010) in which research on the Oregon Assessment of Knowledge and Skills (OAKS) was demonstrated to have lower standard error than the relevant paper test at the tails of the proficiency distribution. Phillips (2009) describes the move to adaptive testing for accountability as providing "[b]etter reliability and more accurate measurement for high and low achieving students and better measurement for SWD and English language learners."

Utility. The computerized format of most adaptive tests allows additional test taker information to be obtained. For example, the computer can keep track of when and how often help was accessed and can also record item latencies (i.e., the time it takes a test taker to respond to an item). It may also be possible to track which accommodations are used when (Thurlow, Lazarus, Albus, & Hodgson, 2010). This would be useful for analyzing student performance after the test. Disabilities research is often a challenge

because specific accommodation types may not be noted in the data file and whether or not the accommodation was actually used (e.g., to answer a specific item under review) is unclear. Novel item types can be explored in a computer-based environment to allow test takers to display various facets of their knowledge. Some newer item types involve multiple-selection multiple-choice choices, selection of text within a passage, or even a chemistry experiment or architectural design simulation. Computerized testing allows the test taker to interact more directly with the test. For example, typical sentence correction items (in which the test taker must identify whether or not something is grammatically incorrect in the proposed sentence; if so, they must choose the option that has the corrected sentence, and if not, they must choose the original sentence) often include what might be considered superfluous reading load. The student is really reading slightly different versions of the same sentence four times. If, instead, the student were able to manipulate the one original sentence on screen, that would cut the reading load down dramatically. The computer could also keep track of whether students were performing any or all of the appropriate steps to solve a problem, and could provide partial scores or a form of scaffolding based on progress on an item. These interactions represent some of the possibilities of "complex" CBT described in Luecht and Clauser (2002).

The use of computers in testing has been shown to provide additional motivation to and engagement of students taking a test; a CAT, specifically, can provide an increase in motivation due to less time required to test and more-appropriate targeting of items to test takers (Clark, 2004; Thompson, 2010). Further, typical CAT item selection algorithms can be modified to include intermittent easier items to further increase test taker motivation and self-confidence (Hausler & Sommer, 2008). Parshall, Spray, Kalohn, and Davey (2002) also note the possibility of targeting items so that test takers have a probability of answering correctly that is greater than the 0.50 value typically used (if no pseudo-guessing probability is assumed). This suggestion and others for addressing examinee affective reactions to adaptive testing are summarized on p. 44 of that text. Verschoor and Straetmans (2010) administered an adaptive placement test to adult mathematics learners with varied levels of previous education in the subject. The study's general examinee population had inadequate mathematics education, did not remember

their school years fondly, and had a tendency toward test anxiety. In many ways, the test anxiety and observed below-average performance make this population similar in some respects to SWD taking tests as part of the general K-12 population. Thus, it is interesting to note that the authors began the adaptive test not with items with difficulties geared toward the average population, but rather with one or two of the easier items in the pool. This approach may ameliorate some of the anxiety that an adaptive test may cause any student, not just SWD. It should be noted that the placement test is cut-score based; in other words, the greatest precision is required around the cut points. For K-12 and other achievement testing, precision is required at multiple score points that categorize scores based on proficiency level.

Appropriateness. By presenting items that are (theoretically) tailored to individual student ability level, an adaptive test can challenge a student at the upper end of the proficiency range while not discouraging students at the lower end of the ability range (Wainer, 2000, p. 11). Additionally, a wide variety of test accommodations can be implemented on the computer, including those involving modified color contrasts, read-aloud accommodation or sign language interpretation, changes in visual representation, text highlighting, and content filters. Many accommodations could be provided via computer much more cheaply than for a paper and pencil test, and alternate test formats could be provided on demand rather than requiring advance notice (Thurlow, Lazarus, Albus, & Hodgson, 2010). The use of a computer-provided read-aloud accommodation has been found to cause less embarrassment and intimidation over having the test read aloud by a human reader, in addition to increasing access (Abell & Lewis, 2005; New England Compact, 2005).

While there are clearly advantages to administering tests via computer, there are also possible drawbacks that affect the general test-taking population. For example, technology comes with a price, and not all school districts will be able to provide the same technology. This socioeconomic factor would have to be addressed in order to provide a fair testing situation, were computer-based tests to be required. In order to receive the benefits afforded by computer-based testing for all students, standardization would have to be ensured, and this may not be feasible. Further, technological literacy, as a construct-irrelevant factor, should have no bearing on performance, and some students are necessarily more tech-literate than are others. In addition to

the typical technological literacy required to take a computer-based test, SWD may have specialized technology that they must make use of to access the test, leading to additional threats to comparability and fairness (e.g., students who are blind and use assistive technologies such as screen readers and refreshable braille displays to access text electronically). Adaptive testing, in theory, provides a way to more accurately, precisely, and efficiently determine the proficiency of a test taker. The precision and accuracy are a result of the test pool having items at all difficulties, including at the tails. However, the accountability system requires tests to be aligned with content standards. Adapting the test at the testlet level, rather than at the item level, may allow strict content standards to be met (Folk & Smith, 2002). Way (2005) notes that the need to meet these content requirements may make a fixed-length CAT more feasible than a variable-length CAT.

One CAT-related issue that has been cited as problematic is that it is often infeasible to allow test takers to review or revisit items and change their responses. The usual reason for not allowing item review is that the CAT algorithm selects each item in sequence depending on the current ability estimate; therefore, returning to an item that was administered previously and changing the response would change the ability estimate one way or the other and could add instability to the estimate. In addition, it could be possible for test takers to use review to game the system (Wise & Kingsbury, 2000). One such scenario that has caused concern would involve a test taker answering items incorrectly, receiving easier and easier items, then returning and answering correctly for all items. However, some items of a higher difficulty would need to be answered correctly to end up with a higher ability estimate. Gershon and Bergstrom (1995) showed that while it would be nearly impossible to significantly increase an artificially reduced ability estimate, it would be more likely that a test taker of truly high ability who attempted to game the system would neglect to change an incorrect answer to the correct answer (thereby sinking the estimate below the true value). Any attempt to use this strategy would likely lead to an ability estimate with increased error. Way (2005) notes that because of the minor impact on measurement that would occur were review enabled, it could be a feature worth allowing. If review over the whole test is not desirable, even giving test takers the ability to review within smaller blocks of items might be useful. Wise and Kingsbury (2000) cite studies in other areas of psychological research that indicate that individuals may feel less anxiety and may have improved performance if they have control in a situation. However, allowing item review may

impact testing time or the time available per item.

ACRE (2010) note that public perception can be negatively affected when using CAT because (a) the public may be skeptical about the fairness of different students taking different test items and (b) the movement away from a number-correct raw score may remove some of the transparency that most linear tests enjoy. It may be difficult to understand both the mechanism behind and the scoring of a CAT.

Challenges of CBTs and CATs Specific to Students With Disabilities

Many of the challenges that will arise for testing SWD in an adaptive setting fall into the categories just stated.

Administration Issues

Kamei-Hannan (2008) investigated the accessibility of CATs for students who are blind or visually impaired and who require braille or large print accommodations, focusing on the reading and language portions of the Measure of Academic Progress (MAP) at a school for the deaf and blind. The author found that testing time was a consideration due to unfamiliarity with some aspect of technology or with braille. In addition, the use of magnification tended to increase testing time because of the need to scroll or scan through documents. For students using refreshable braille, approximately 21% of reading items and 13% of language items were found to have accessibility issues. For the reading test, these were mostly related to long reading passages that required scrolling and pictures associated with word attack skill questions. The author notes that these issues depend, for a large part, on grade level: at lower grade levels, a student will be presented more often with items associated with pictures, and at advanced grade levels students will be presented with longer reading passages. For the language test, underlining was often a cause of accessibility issues for braille readers because either it wasn't present, or it was represented in an unfamiliar way (using an eight-dot braille cell). Thurlow, Lazarus, Albus, and Hodgson (2010) also noted that SWD may require more preparation to use some technology-enhanced features (e.g., they specifically mention measurement tools such as rulers and compasses). Additionally, they caution that using text-to-speech renderings rather than human voicing for a read-aloud accommodation may be different than what students expect to hear, and that oral presentation of pictures or mathematical statements may be problematic. These issues speak more to computerized tests in general, but they take on greater importance in a CAT

because the reliance on computer-based accommodations (e.g., text-to-speech, refreshable braille) becomes critical for every test item in the pool rather than a subset of items that are "easily" rendered in an alternate format. In addition, it should be noted that one threat to proper CAT functioning occurs if items that were estimated in the general population to be of low difficulty are actually harder because of some unforeseen locus of variance such as presentation via text-to-speech. This idea will be developed more fully in the section on technical issues.

The general advantage of a shorter testing time in CAT would be even more beneficial for students who have problems with eye strain, particularly if they will be testing exclusively on a computer. However, this advantage would be nullified if situations such as those explored in Kamei-Hannan (2008) arise. A shorter testing time might help students who, when taking the test with an extended-time accommodation, suffer fatigue from the longer testing session (Cahalan Laitusis, Morgan, Bridgeman, Zanna, & Stone, 2007).

While computer-based tests can increase accessibility by allowing the implementation of a variety of accommodations (Phillips, 2009), DRC (2007) notes several areas to consider for SWD when moving to computerized testing: (a) familiarity with and ability to use technology (which also holds for students without disabilities); (b) use of innovative item types that may not be accessible (e.g., items that are not braillable); and (c) allowing multiple options for selecting responses (e.g., mouse, keyboard, touchscreen) (p. 24). Additionally, implementing accommodations that satisfy student IEPs (e.g., separate testing location, required software) can be challenging if schools have all computers in grouped lab areas. Another issue with accommodations is that different SWD use different accommodations and it might be difficult to enable all accommodation types to interact well with the accessible computer platform. A pressing limitation is that no platforms currently have refreshable braille, which raises concerns about participation of and fairness for students who are blind. For systems that must score on the fly, a similar constraint is imposed upon any student who uses an accommodation that requires an alternative method of scoring (e.g., when responses are hand-signed or oral).

Technical Issues

Adaptive tests are more efficient than linear tests. Fewer items are required to be administered to reach a particular measurement precision because the items are better targeted to hone in on the proficiency of each test taker. This is an advantage for all students, especially those who may experience fatigue effects, but face validity may decrease--- there may be

concern from parents and students if the majority of students only have to take a reasonable number of items while others have to take additional items to reach the desired level of precision of the ability estimate. This could be a particular concern for SWD, who could have test lengths quite different from students without disabilities, especially if idiosyncratic responses cause the estimation algorithm to destabilize. Alternatively, students who have to answer fewer items may question whether the shorter test was truly able to estimate their ability level (i.e., whether they had enough of a chance to show what they know). However, two tests with the same number of questions and same time limit may not measure students to the same precision and may not be equitable in terms of time required. Difficult items may require more time to respond to, so two tests of the same length may have different inherent time requirements, introducing differential speededness (see, e.g., Bridgeman, Laitusis, & Cline, 2007). There may also be convergence issues if the proficiency estimate fails to stabilize and the standard error does not shrink.

As noted in Parshall (2002), adaptive testing usually uses an underlying IRT model, so calibration of items will likely be necessary. It will be important to note which examinees will be used to form the item parameters. Because almost all students will be tested with the same item bank (although they should receive different items), with the possibility that the item response functions will differ for some SWD, this situation needs to be explored in detail. Karkee, Lewis, Barton, and Haug (2003) compared results from including and excluding SWD on state standards-based assessments in several subjects (reading and writing, math, and science) in grades 4, 7, and 8. The most frequently occurring accommodations were extended time and oral presentation. They found that although there were some significant differences in some IRT parameter estimates, there was minimal effect on student scores or DIF between nonaccommodated and accommodated groups. However, it should be noted that the overall sample was very large and the original number of DIF items was small. If examinees with disabilities are not used at the calibration stage, it would be crucial to implement subgroup analyses to verify the appropriateness of the resulting item parameters for these groups. Parshall also notes that some testing programs simply move items from paper to computer, and this may have a different mode effect for SWD than for students without disabilities because they may have more or less access, or may have a different interaction with computers. An issue related to using θ-based scores under the 3PL model is that, viewing the θ-based score as a weighted sum of item scores, the weights in that model will depend on the (possibly unstable or misleading, in

the case of SWD) θ estimate (Lord, 1980, p. 75). If test scoring is accomplished via a different model, some of the problematic aspects may be mitigated.

A potential disadvantage of adaptive testing that requires additional research is the implication of divergent knowledge patterns in students with specific disability subtypes. Recently, researchers have questioned if SWD are more likely to exhibit idiosyncratic knowledge patterns within a content domain and, if so, the implications of this for adaptive testing algorithms (Buzick & Laitusis, 2010; Laitusis, Cook, Buzick, & Stone, 2011). For example, in recent testimony to the Senate HELP committee, Martha Thurlow from the National Center for Education Outcomes (NCEO) commented that "Even when constrained to grade level, adaptive testing practices must be transparent enough to detect when a student is inaccurately measured because of splinter skills common for some SWD, for example, with poor basic skills in areas like computation and decoding, but with good higher level skills, such as problem solving, built with appropriate accommodations to address the barriers of poor basic skills" (ESEA Reauthorization: Standards and Assessments, 2010). This divergent profile is most likely to occur in students with learning disabilities because classification of students is heavily influenced by divergent cognitive profiles (IQ-achievement discrepancy) or lower achievement levels in specific academic knowledge areas. Fletcher, Lyon, Fuchs, and Barnes (2007) define five broad learning disabilities characterized by deficits in decoding, reading fluency, comprehension, math fluency, or writing. The implication for this is, for example, that students with learning disabilities defined by deficits in math fluency, *dyscalculia,* may perform poorly on relatively easy test items that measure basic calculation but perform well on relatively difficult items that measure higher-level mathematical knowledge. The consequences of such idiosyncratic responding in an adaptive setting can be disastrous in terms of arriving at a stable and accurate proficiency estimate. For example, memorization of a particular item type led to this situation on a high-stakes CAT. One particular examinee responded correctly to the memorized items, which were calibrated as being difficult compared to those to which the examinee had responded incorrectly. This created a scenario for which determination of the "best" ability estimate was unsuccessful.

The need for detection of aberrant response patterns in IRT is not a new idea (Kingsbury & Houser, 1993), and general procedures have been in place for years (e.g., Drasgow, Levine, & McLaughlin, 1987; Tatsuoka, K. K., 1984). Detection statistics were used by Meijer, Egberink,

Emons, and Sijtsma (2008) and Egberink, Meijer, Veldkamp, Schakel, and Smid (2010) to investigate aberrant response patterns on a self-perception inventory and a career-development inventory, respectively, with a focus on interpretability of the resulting scores in the presence of aberrance. The use of CAT in the presence of idiosyncratic knowledge patterns has been studied by Kingsbury and Houser (2007) and it has been shown that scoring of adaptive tests can be problematic when a test taker responds to items in an unpredictable way (e.g., correctly answers more difficult items while incorrectly answering easier items). Some early research raising the specter of idiosyncratic response patterns by SWD took place in the context of reading for students with learning disabilities. Cromer and Wiener (1966) administered a four-passage oral story-reading task to fifth-grade students in order to determine whether passage tense and content differentially affected responses for students in a remedial reading group versus students not in a remedial reading group. The researchers found that students with learning disabilities had significantly poorer performance (in terms of uncorrected errors) on affective stories than on neutral stories, while there was no significant difference for the students without disabilities. In addition, although both groups struggled with present-tense stories as compared to past-tense stories, the students with learning disabilities had significantly more of a struggle with the present-tense stories than did the students without disabilities.

Another example of unpredictable item difficulty discrepancy was identified by Stone, Cook, Cahalan Laitusis, and Cline (2010) for students who are blind or visually impaired. In that study, which used DIF to identify items for which the students who are blind or visually impaired performed differently than the reference group of students without disabilities, a panel of experts provided insights about potential reasons for the DIF that was found. During that discussion, a teacher of the visually impaired noted that although some test items require students to examine and interpret material from alternative document formats (e.g., a flyer or a poster), students who are blind or visually impaired may not have had as much experience with those formats because they are not always used in the classroom. There are also questions about how the brailling of math item features affects the way students respond to items. Item response theory models that often underlie adaptive testing may assume that the higher a student's proficiency level, the more likely the student is to answer the item correctly. However, it seems clear that there may be cases in which a student with a disability would respond in a way contrary to this general model. One area in need of additional research is to further examine

situations in which discrepant responding by SWD may occur. For example, previous studies (e.g., Stone, 2008) have found that some math item types (e.g., four-quadrant graph items) consistently displayed DIF in a comparison of students without disabilities and students taking the test with math modifications (arithmetic tables, calculator, and/or math manipulatives). Other testing areas (e.g., English language arts) for students with other types of disabilities (e.g., visual disabilities, reading-based learning disabilities) may display similar potential patterns worth investigating. From a fairness standpoint, if SWD are administered easy items based on divergent incorrect responding to the first few items, they may be deprived of the opportunity to demonstrate their proficiency on items deemed more challenging that are actually within their proficiency range (Thurlow, Lazarus, Albus, & Hodgson, 2010).

Utility Issues

As mentioned previously, certain types of accommodations may make on-the-fly scoring (and, hence, rapid score reporting) infeasible. Additionally, now that SWD will be participating in CATs in greater numbers, research should be done to find evidence to support or refute the hypothesis that these students will find the test more motivating and engaging than its linear counterpart, an area that appears to be lacking in the literature.

Appropriateness Issues

While adaptive testing may be better able to target students who would normally be ill-measured by a typical state test, there is some concern that moving to easier items to assess the students with poorest performance may lead to an out-of-level test (Consortium for Citizens with Disabilities, 2009; Kingsbury & Hauser, 2004; Thompson & Way, 2007; Thurlow, Elliott, & Ysseldyke, 2003). The requirement for accountability tests to contain on-grade-level items is a possible conflict noted by many disabilities researchers and may mean that a student with disabilities will not receive an accurate score estimate if the true proficiency is below grade level. In addition, on-grade-level items may not be informative if they are still too difficult for a student (Trotter, 2003).

Possible Ways to Address These Issues, and Considerations for Future Work

There are test development approaches, hybrid CAT/linear testing methods (e.g., Clark, 2004), multistage or two-stage testing methods, and alternative scoring approaches that may

address some of the concerns raised in this paper (see Davey, 2011 for a summary).

Multistage testing allows the difficulty of the test to be adjusted for different examinees, but also provides additional benefits. For example, it allows items to be presented in a particular order to examinees. This could be useful for students in K-12 who are used to linear tests with items that increase in difficulty as the test progresses. An item-level CAT should, theoretically, get harder with each correct answer and easier with each incorrect answer, but the dynamic may be off-putting. Item context may also be controlled through the use of testlets or multiple stages of linear testing. One possible risk with two-stage testing is that the routing test could route a student to the wrong second-stage test, something that could be adjusted for were there additional stages (Folk & Smith, 2002, p. 48). The requirement that tests be on grade level might be addressed by using a hybrid test with a fixed portion for accountability and an adaptive portion to fine-tune the proficiency estimate. Another alternative would be to use IRT-based growth models to take into account out-of-level status (DRC, 2007).

Scaffolding could be used instead of adaptive testing to allow students to respond to items with more challenging content (Almond et al, 2010). Scaffolds are meant to bridge the gap between a student's actual and possible proficiency, and they are typically used in an instructional setting. Some scaffolds mentioned by Almond et al. are (a) allowing students additional chances to respond after answering an item incorrectly on the first try, and (b) presenting a different item that provides a clue about appropriate strategy for the initial item and then providing (for response) a third item that is of similar content and difficulty to the initial item. While accommodations are not meant to change the construct being measured, scaffolds may alter the item so that measurement is of a different knowledge, skill, or ability. Therefore, the use of scaffolds in summative assessment would require the creation of an appropriate scoring design that would take this into account.

Problematic use of IRT pattern-based scores (e.g., flat likelihoods) and lack of transparency in IRT scoring may be addressed through the use of *estimated number-correct scores* or *IRT-equated number-correct scores* (Stocking, 1996). Estimated number-correct scores are IRT-based scores that are obtained by summing the probabilities of correct responses, given the examinee ability estimate, for all items in a reference item set. For example, the reference test could be the conventional linear test to which the adaptive test's θ metric is linked, and the estimated number-right score would then be the expected number correct if the test taker (or any

test taker with the same ability estimate) had been given the linear form of the test. An even more familiar number-correct approach could be applied using IRT true score equating. These IRT-equated number-right scores would be obtained by counting up the number right on the adaptive test and using the test characteristic curve for the administered items to convert the number right score to an IRT true score. Specifically, one would find the ability value that the observed number right maps to through the test characteristic curve. This procedure is the method of scoring for at least one high-stakes multistage test. While the scoring is still based on IRT θ estimates, the use of a number-correct approach may be more palatable to test takers and score users.

Research currently underway in this area includes an analysis of a current state test to determine if SWD respond to specific item types differently than students without disabilities (i.e., divergent response profiles). If significant differences are detected, one way to address problematic item types or content would be to include constraint codes that would allow for control of item selection based on those characteristics in a similar way as content overlap and gender/ethnicity issues have been addressed (Way, 2005). Differential speededness issues could also be addressed through the use of constraints (van der Linden, Scrams, & Schnipke, 1999). Possible multidimensionality of item responses may make the use of other response models, such as multidimensional IRT, worth investigating (Wise & Kingsbury, 2000).

Clearly, the move toward adaptive testing in the K-12 setting will require that a significant amount of thought be given to how such testing conditions would work for students who may not respond to or interact with the test in the same way as the majority of the testing population. The goal of providing fair and valid assessments for all students requires that we make this close inspection of our testing practices.

References

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential distractor functioning in reading assessments for students with disabilities*. Minneapolis: University of Minnesota, Partnership for Accessible Reading Assessment.

Abell, M., Bauder, D., & Simmons, T. (2004). Universally designed online assessment: Implications for the future. *Information Technology and Disabilities Journal, 10*(1). Retrieved from http://people.rit.edu/easi/itd/itdv11n1/abell.htm

Accountability and Curriculum Reform Effort (2010). *Computerized adaptive testing: How CAT may be utilized in the next generation of assessments*. Retrieved from http://www.ncpublicschools.org/docs/acre/publications/2010/publications/20100716-01.pdf

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke, J., Torres, C., Haertel, G., Dolan, B., Beddow, P., & Lazarus, S. (2010). *Technology enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research.* Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Alpert, T. (2010, April). *A coherent approach to adaptive assessment*. Presentation at Best Practices for State Assessment workshop for the National Academy of Science Board on Testing and Assessment, Washington, DC. Retrieved from http://www7.nationalacademies.org/bota/Best_Practices_for_State_Assessment_presentation_Alpert.pdf

Bridgeman, B., Laitusis, C. C., & Cline, F. (2007). *Time requirements for the different item types proposed for use in the revised SAT®* (ETS Research Report No. RR-07-35). Princeton, NJ: ETS.

Burling, K. (2007). NCLB regulations for modified achievement standards (2%). Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/680128F1-B412-4A20-9A20-D4FD6AC613D6/0/wp0702.pdf.

Buick, H. M., & Laitusis, C. C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39*(7), 537–544.

Cahalan Laitusis, C., Morgan, D. L., Bridgeman, B., Zanna, J., & Stone, E. (2007). *Examination of fatigue effects from extended-time accommodations on the SAT Reasoning Test* (College Board Research Report No. 2007-1). New York, NY: The College Board.

Clark, L. (2004).Computerized adaptive testing: Effective measurement for all students. *Technological Horizons in Education Journal*, *31*(10), 14–16.

Consortium for Citizens with Disabilities. (2009). *Race to the Top Assessment Program*. Retrieved from http://www.c-c-d.org/task_forces/education/CCD%20R2T%20Assessment%20program%20letter%20final.pdf.

Cromer, W., & Wiener, M. (1966). Idiosyncratic response patterns among good and poor readers. *Journal of Consulting Psychology*, *30*(1), 1–10.

CTB McGraw-Hill (2003). *Computer-based testing – Issues and considerations* [Abstract]. Retrieved from http://ccsso-cbtool.com/pdf_files/Design%20Considerations%20Article%201.pdf

Data Recognition Corporation (2007). *Study on the feasibility and cost of converting the state assessment program to a computer-based or computer-adaptive format*. Retrieved from http://eoc.sc.gov/NR/rdonlyres/CAEF9136-26CB-421D-80E3-D5B35B72CE76/5535/SCFeasibilityFinalReport.pdf.

Davey, T. (2011) *Practical considerations in computer-based testing*. Retrieved from http://www.ets.org/Media/Research/pdf/CBT-2011.pdf

Drasgow, F., Levine, M. V., McLaughlin, M. E. (1987) Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*(1), 59–79.

ESEA Reauthorization: Standards and Assessments: Hearings before the Health, Education, Labor, and Pensions Committee (HELP), United States Senate, 111th Cong., 2nd Sess. (2010) (testimony of Martha Thurlow).

Egberink, I. J. L., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in a computerized adaptive test: An empirical example using the CUSUM. *Personality and Individual Differences*, *48*(8), 921–925.

Fletcher, J. M., Lyon, G. R., Fuchs, L., & Barnes, M. (2007). *Learning disabilities: From identification to intervention*. New York, NY: Guilford Press.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Erlbaum.

Gershon, R. C., & Bergstrom, B. A. (1995, April). *Does cheating on CAT pay: NOT!* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hausler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly*, *50*(1), 75–87.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*, 44–52.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Kamei-Hannan, C. (2008). Examining the accessibility of a computerized adapted test using assistive technology. *Journal of Visual Impairment & Blindness*, 102*(5)*, 261–271.

Karkee, T., Lewis, D. M., Barton, K., & Haug, C. (2003, April). *The effect of including or excluding students with testing accommodations on IRT calibrations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kingsbury, G. G., & Hauser, C. (2004, April). *Computer adaptive testing and the No Child Left Behind Act*. Paper presented at the annual meeting of the American Educational Research Association, San Diego CA. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ki04-01.pdf.

Kingsbury, G. G., & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice*, *12*(1), 21–27.

Laitusis, C. C., Cook, L. L., Buzick, H. M., & Stone, E. (2011). Adaptive testing options for accountability assessments. In M. Russell (Ed.), *Assessing students in the margins: Challenges, strategies, and techniques*. Charlotte, NC: Information Age Publishing.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Luecht, R. M., & Clauser, B. (2002). Test models for complex CBT. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments*, (pp. 67–88). Mahwah, NJ: Erlbaum.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self perception profile for children. *Journal of Personality Assessment*, *90*(3), 227–238.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum.

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis* (Out-of-Level Testing Report 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

National Center for Learning Disabilities (2009). *No Child Left Behind Act (NCLB): An overview*. Retrieved from http://wwworg.ncld./on-capitol-hill/federal-laws-aamp-ld/no-child-left-behind-act/no-child-left-behind-act-nclb-an-overview

New England Compact Enhanced Assessment Project (2005) *Computer-based read-aloud tools: An effective way to accommodate students with disabilities*. Available at http://www.necompact.org/NECompact%20Summaries_0205.pdf

Olson, L. (2003, May 8). Legal twists, digital turns. *Education Week*, *22*(35), 11–15.

Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing* (pp. 119–141). Mahwah, NJ: Erlbaum.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. C. (2002). *Practical Considerations in Computer-Based Testing*. New York, NY: Springer-Verlag.

Peterson, Kristin A. (2005). C*omputerizing statewide educational assessments in Minnesota: A report on the cost and feasibility of converting the Minnesota comprehensive assessments to a computerized adaptive format*. Minneapolis, MN: Office of Educational Accountability, College of Education and Human Development, University of Minnesota.

Phillips, G. (2009, December). *Race to the Top Assessment Program: A new generation of comparable state assessments*. Presentation at the United States Department of Education Public Hearings, Denver, CO.

Reckase, M. (2011). Computerized adaptive assessment (CAA): The way forward. In Policy Analysis for California Education and Rennie Center for Education Research & Policy (Eds.), *The road ahead for state assessments* (pp. 1–12). Cambridge, MA: Rennie Center for Education Research & Policy.

REL West/WestEd. (2008). *Considerations in statewide implementation of computer-adaptive testing*. Retrieved from http://relwest.wested.org/system/memo_questions/11/attachments/original/Computer_20adaptive_20testing_20June_202008_1_.pdf.

Scheuermann, F., & Björnsson, J. (Eds.). (2006). *The transition to computer-based assessment. Lessons learned from the PISA 2006 computer-based assessment of science (CBAS) and implications for large scale testing.* JRC Scientific and Technical Reports. Retrieved from http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/8713/1/reqno_jrc49408_final_report_new(1)%5B1%5D.pdf

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics 21*(4), 365–389.

Stone, E. (2008, April). *Examining the scores of students with disabilities on state standards-based math and science tests: A differential distractor functioning analysis.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Stone, E., Cook, L. L., Cahalan Laitusis, C., & Cline, F. (2010). Using differential item functioning to investigate the impact of testing accommodations on an English language arts assessment for students who are blind or visually impaired. *Applied Measurement in Education 23*(2), 132–152.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*(1), 95–110.

Thompson, N. A. (2010). *Adaptive testing: Is it right for me?* Retrieved from www.assess.com/docs/Thompson_(2010)_-_Adaptive_Testing_Right.pdf.

Thompson, T., & Way, W. D. (2007, June). Investigating CAT designs to achieve comparability with a paper test. Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements.* Thousand Oaks, CA: Corwin Press.

Thurlow, M., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility Principles for Reading Assessments.* Minneapolis, MN: National Accessible Reading Assessment Projects.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report No. 78). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Trotter, A. (2003, May 8). A question of direction. *Education Week*, *22*(35), 17–21.

van der Linden, W. J., & Glas, C. A. W. (2010). Preface. In W. van der Linden & C. Glas (Eds.), *Elements of Adaptive Testing* (pp. v–vii). New York, NY: Springer.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. van der Linden & C. Glas (Eds.), *Elements of Adaptive Testing* (pp. 3–30). New York, NY: Springer.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210.

Verschoor, A. J., & Straetmans, G. J. J. M. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. van der Linden & C. Glas (Eds.), *Elements of Adaptive Testing* (pp. 137–149). New York, NY: Springer.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd edition). Mahwah, NJ: Lawrence Erlbaum.

Way, W. D. (2010, June). *Some perspectives on CAT for K-12 assessments.* Presented at the National Conference on Student Assessment, Detroit, MI.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Practical questions in introducing computerized adaptive testing for K-12 assessments.* Retrieved from http://www.pearsoned.com/RESRPTS_FOR_POSTING/ASSESSMENT_RESEARCH/AR6.%20PEM%20Prac%20Questions%20in%20Introl%20Computer%20Test05_03.pdf

Way, W. D., Twing, J. S., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved from http://www.ets.org/s/commonassessments/pdf/AdaptiveTesting.pdf

Wise, S. L., & Kingsbury, G. G. (2000) Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*, *21*, 135–155.