RedFame

Journal of Education and Training Studies
Vol. 1, No. 2; October 2013
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Test Anxiety, Computer-Adaptive Testing and the Common Core

Nicole Makas Colwell

Correspondence: Nicole Makas Colwell, Educational Psychology, MESA, College of Education, University of Illinois at Chicago, USA. E-mail: ncolwe2@uic.edu

Received: March 17, 2013 Accepted: April 3, 2013 Available online: May 18, 2013

doi:10.11114/jets.v1i2.101 URL: http://dx.doi.org/10.11114/jets.v1i2.101

Abstract

This paper highlights the current findings and issues regarding the role of computer-adaptive testing in test anxiety. The computer-adaptive test (CAT) proposed by one of the Common Core consortia brings these issues to the forefront. Research has long indicated that test anxiety impairs student performance. More recent research indicates that taking a test in a CAT format can affect the ability estimates of students with test anxiety. Inaccurate measures of ability are disconcerting because of the threat they pose to the validity of test score interpretation. This paper raises concerns regarding how the implementation of a computer-adaptive test for a large-scale common core assessment system could differentially affect students with test anxiety. Issues of fairness and score comparability are raised, and the implications of these issues are discussed.

Keywords: computer adaptive testing, test anxiety, common core assessments

1. Introduction

The goal of large-scale assessments is to improve the educational process by monitoring student achievement. In recent years, federal law has mandated large-scale assessments for the purpose of accountability in hopes of advancing student performance. There is growing concern that the increase in testing over the years has had a negative impact on student learning (Miller, Linn, & Gronlund, 2009). Research has demonstrated that some of the adverse effects of high-stakes testing on students include illness, anxiety, and heightened levels of stress (Triplett, Barksdale, & Leftwich, 2003). Many parents and educators believe that standardized tests are responsible for creating anxiety and tension in students (Mulvenon, Stegman, & Ritter, 2005). This is not an unreasonable speculation since there has been a steady increase in the prevalence of test anxiety among students over the decades. In the early 1980's, researchers studying testing anxiety reported that between 10% and 25% of the students in the US experienced test anxiety (e.g., Hill & Wigfield, 1984). Today, this number has increased to more than 33% of US students experiencing some form of test anxiety (Methia, 2004). Research has established that test anxiety has a negative impact on achievement motivation and results in an inadequate assessment of student ability (Hembree, 1988). This is a serious concern as inadequate assessments of ability ultimately undermine the validity and reliability of test score interpretability. According to the *Standards for Educational and Psychological Testing*, validity is the most fundamental consideration when developing and evaluating a test, and in order to provide information about student achievement adequate validity evidence is required to obtain useful and meaningful scores (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Increased testing and the high stakes decisions surrounding standardized test scores have also led researchers to find more accurate and reliable ways to assess student ability. Research efforts to maximize efficiency and optimize measurement in testing led to the development of computer-adaptive testing (CAT) (Reckase, 2011). Numerous studies have shown that using CAT methods results in more accurate and precise measurement of ability in comparison to conventional fixed item testing. Despite the improved measurement precision CATs offer, they are still susceptible to the effects of test anxiety (Pitkin & Vispoel, 2001). Since misrepresentation of true skill levels can result, test anxiety can affect the validity of test score interpretations from examinee scores on any assessment including CATs. Therefore, considering the effects of test anxiety on CATs is essential due to the negative effects associated with test performance. This is especially important when CATs are part of high-stakes testing.

The recent adoption of the Common Core State Standards by numerous states brings this issue to the forefront. The Common Core State Standards Initiative was established to provide a consistent, clear understanding of what students are expected to learn and yield well-constructed tests that include tasks with real world relevance. Two consortia were funded to develop rigorous assessments, aligned to the new standards that effectively and efficiently measure student performance, accountability, and support better instruction (Reckase, 2011). One of the consortia, the Smarter Balanced Assessment Consortium (SBAC), plans to develop a computer-adaptive system for their set of assessments. Some experts have voiced concerns over SBAC's decision to implement a CAT, because there are many challenges to successfully implementing a computer-adaptive system (Reckase, 2011; Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). However, these concerns do not address student factors, such as test anxiety, computer anxiety or poverty. The individual differences within these student factors contribute to issues of fairness when only one of the consortia plans to use a computer-adaptive assessment. Additionally, score comparability across the Consortia is a challenging issue with two different modes of testing. As high-stakes decisions are being based on the test scores obtained from these Common Core Assessments, it is imperative that these issues are addressed in to order to ensure that the tests provide fair representations of all student ability. Therefore, the purpose of this paper is to examine the comparability of paper based tests versus computer-adaptive tests when assessing kids with test anxiety and highlight areas that the common core assessments need to address in regards to this issue. The first part of the paper will focus on defining test anxiety and computer-adaptive testing and summarize the research relating the two research areas. The second part will focus on the issues regarding test anxiety that are likely to impact the Common Core assessments since the consortia are constructing two different test formats.

1.1 What is Test Anxiety?

Anxiety is commonly referred to as an unpleasant emotional state characterized by excessive degrees of fear, worry and apprehension without a specific object or cause (Putwain, 2008). It is initiated by internal feelings, as a response to a perceived threat (Casbarro, 2005). Test anxiety is considered a special case of anxiety that occurs in an assessment context or evaluative situation. It is considered to be a multi-dimensional construct that consists of two major factors: a cognitive dimension and an emotionality dimension. The cognitive dimension refers to the mental activity that revolves around the testing situation and encompasses worry and irrelevant thinking or negative thoughts coupled with emotional discomfort. The emotionality dimension refers to the physiological component that includes tension, bodily reaction, and perceived arousal (Cassady, 2004; Zeidner, 2007). Zeidner (1998) defined test anxiety as a set of phenomelogical, physiological, and behavioral responses that accompany concern about negative consequences or failure in an evaluative situation. In educational settings test anxiety is common, where the demands from a testing situation can incite a fear of failure, threat to self-esteem and worry over how the performance will be judged by others (Putwain, 2008). Test anxiety largely depends on the extent to which students perceive assessments as threatening, and both personal and environmental characteristics can influence the onset (Putwain & Daniels, 2010).

When students believe the evaluative situation taxes or exceeds their intellectual, motivational, and social capacities, test anxiety is elicited (Putwain, Woods, & Symes, 2010). Skinner, Furrer, Marhcund and Kindermann (2008) reported that anxiety is strongly related to perceived control, where students low in perceived control are more at risk for escalating anxiety. In addition, the effort applied by a child is associated to his or her perceived ability of achieving success and control (Nicholls, 1990). This means students with high test anxiety would apply little effort when their perception of success on the test and control over the situation is low. Thus, the lack of control in computer-adaptive testing is disconcerting for test anxious students, because control is directly linked to their test performance.

In addition, children suffering from test anxiety are more sensitive to failure and feelings of judgment (Hill & Wigfield, 1984). This is another cause for concern, because not only does the adaptive algorithm behind CAT only allow examinees to answer about 50% of items correctly, it also bases the next question on the performance of the previous (Olea, Revuelta, Ximenez, & Abad, 2000). Therefore, it is possible that students will experience increases in their anxiety levels, if they expect to perform better than 50%, or they feel judged by their performance on each item.

1.2 What is Computer-adaptive Testing?

As opposed to a linear test with fixed test items, computer-adaptive testing is a special approach to the assessment of latent traits, where the test is specifically matched to the needs of each examinee (Davey & Pitoniak, 2006). The test taker's individual trait level is iteratively estimated during the testing process. The test is assembled as the examinee works, and selection of test items is based on examinee response given on items

presented previously. If the first item is answered correctly, the next item will be more difficult. If the item is answered incorrectly, then the next item will be less difficult. The examinee's current ability estimate is continually calculated and adapted based on the items that have been answered thus far. This process continues until a test termination criterion is reached, such as a pre-specified level of measurement precision or a fixed number of items have been administered (Davey & Pitoniak, 2006).

Administering an adaptive test consists of two steps: item selection and score estimation (Davey & Pitoniak, 2006). The first step involves administering appropriate items given the examinee performance level. The primary item selection criterion is to maximize the test information function and minimize measurement error in the examinee's score (Dragsaw, Luecht, & Bennett, 2006). Selecting items involves three considerations: optimizing test efficiency, properly balancing the test, and protecting items from overexposure.

In the second step, after administering the item, the computer uses the response to refine the score or estimate so that the next item it selects is even more appropriate. CAT employs item response theory (IRT) to estimate proficiency and calculate measurement errors. The IRT model relies on the principles of maximum likelihood estimation and uses trait levels and item parameters to calculate examinee proficiency estimates. An IRT model is often characterized by the mathematical function for the probability of observing a particular item response given the examinee's trait level and the particular item parameters. In this regard, it is a form of item-free measurement because it is conditional on ability. Examinees do not need to be administered the same items to obtain their proficiency estimates. This is what allows score comparisons even though individual tests will often be composed of different items.

CAT offers improved testing efficiency. That is, one can obtain examinee ability estimates of greater precision using fewer items than are required when using non-adaptive tests (Dragsaw et al., 2006). The item selection algorithm is the mechanism that makes this possible: items that are too easy or too hard for an examinee are not administered. To obtain the same measurement precision, a CAT needs to be only half as long as a parallel non-adaptive test (Dragsaw et al., 2006). Although achieving optimal efficiency is rare due to practical constraints (e.g., content balancing), CATs are still significantly shorter than fixed-item tests. The administrative efficiency and measurement precision that CATs offer has therefore led many measurement professionals to consider them superior to conventional tests (Zwick, 2006).

2. What role Does Test Anxiety Play in CAT?

Numerous studies have examined the psychometric and technical aspects of CAT, but research investigating the psychological effects of CATs on test takers is limited (Ortner & Caspers, 2011). Among these, some have looked at the effects of test anxiety in relation to CAT. Although the research is sparse, the findings are compelling.

2.1 Fixed-item Testing versus Computer-adaptive Testing

When comparing fixed-item tests (FIT) to a CAT, a common finding is that test takers dislike certain features of adaptive tests such as the inability to review or skip items (Tonidandel & Quinones, 2000; Vispoel, 1993). Test takers feel they are at a disadvantage and perceive the test as unfair (Vispoel, 1993). Additionally, a decrease in test-taker motivation and self-confidence is associated with adaptive testing versus conventional testing (Frey, Hartig, Moosbrugger, 2009; Hausler & Sommer, 2008). This results from being exposed to items that have higher difficulty levels more quickly in CAT than in FIT. The ability to answer only about half of the items correctly in an adaptive test has also been linked to lower motivation. Tonidandel, Quinones, and Adams (2002) found a significant negative relationship between a measure of self-motivation and the average probability of answering items correctly. As noted earlier when students believe their motivational capabilities are taxed by a given situation or their self-confidence has been questioned, test anxiety likely results. Thus, these findings suggest a greater chance for students to experience test anxiety in a CAT than in a FIT.

Ortner and Caspers (2011) examined the fairness of CAT when assessing persons who had high trait test anxiety in comparison to persons who had low trait test anxiety. Comparing CAT to FIT, they investigated whether test anxiety had differential effects on test performance depending on test mode. In addition, they examined whether telling test takers about the mechanisms of item selection in CAT might impact their subsequent test scores. The analysis revealed a significant interaction between test mode and test anxiety. Test scores of high and low test-anxious examinees were similar when taking the FIT; but in the CAT condition, high test-anxious examinees had considerably lower scores than low test-anxious examinees. This result was associated with a medium effect size and indicates that higher test-anxious students taking a CAT were more likely to obtain biased test scores than lower test-anxious students. This finding has practical implications when implementing a CAT in a high-stakes setting, such as the Common Core assessments. They also found a significant main effect for

information given about the testing and item selection procedures of CAT. Test takers who received information had higher test scores compared to test takers who did not receive that information (Ortner & Caspers, 2011), which suggests that providing information to test takers about CAT prior to taking the exam can decrease their anxiety.

2.2 Item Review

Another study examined the effects of item review (allowed vs. disallowed) on performance in both CAT and FIT and its relation to test anxiety (Olea et al., 2000). The researchers randomly assigned examinees to one of four conditions (FIT-review, FIT-non-review, CAT-review, and CAT-non-review), and they assessed the examinees' state-anxiety and degree of comfort levels. They reported a significant negative relationship between anxiety level and the review conditions, where examinees allowed to review their answers showed a decrease in anxiety levels, and examinees not allowed to review showed an increase in anxiety levels. Examinees in the FIT conditions also exhibited more calmness than examinees in the CAT conditions. The researchers also examined the effects of revision on the psychometric properties of the CAT. They found that although review significantly increases the number of correct responses, which ultimately results in increased ability levels of the examinee, it did not impact the level of precision in ability estimation (Olea et al., 2000). Thus, review contributes to more accurately estimated ability levels, and not allowing item review can lead to potentially underestimating the ability of examinees. This holds important implications for both high and low test-anxious examinees.

Researchers investigating gender differences in test anxiety and CAT report inconclusive results. In a study investigating gender, anxiety levels and CAT, the interaction of gender, test anxiety and test modality was not significant (Fritts & Marszaleck, 2010). This is surprising, as test anxiety and gender differences in conventional testing are well established in the test anxiety literature (Hembree, 1988). However, the results from the path analysis they conducted did show females to have higher trait test anxiety overall, which leads to higher state test anxiety. This is more consistent with previous findings regarding differences in gender and anxiety levels. The researchers cautioned against using the findings from their study, since few researchers have investigated gender differences in relation to CAT and anxiety levels. They suggested that researchers conduct more studies in order to confirm (or disconfirm) their findings (Fritts & Marszalek, 2010).

2.3 Self-adaptive Testing versus Computer-adaptive Testing

Researchers investigating the effects of different modes of adaptive testing on test anxiety have concluded that self-adaptive testing (SAT) is better suited than CAT for examinees with test anxiety (Shermis, Mzumara, & Bublitz, 2001; Wise, Roos, Plake, & Nebelsick-Gullett, 1994). SAT is a technique in which the examinee chooses the difficulty of the administered items, as opposed to a computer algorithm like CAT. Before responding to an item, the individual is asked how difficult an item he or she would prefer, where the levels of difficulty vary, ranging anywhere from three to eight levels (Shermis et al., 2001). The SAT algorithm then uses the chosen difficulty level and selects items tailored to the present ability level. Numerous studies have shown that SATs are associated with higher test scores than either CAT or FIT and lower anxiety levels compared to CAT (Wise et al., 1994). Other research, however, has shown that CATs are superior to SATs in terms of measurement precision and efficiency (Pitkin & Vispoel, 2001; Shermis et al., 2001).

The reduced anxiety levels resulting from SATs have largely been explained by the perceived control hypothesis (Wise, 1994). Research shows that in many different contexts people better manage a stressful situation when they perceive they have some control over their environment (Wise, 1994). In educational settings this has been particularly true for testing, where research shows that giving examinees the perception of control over certain testing situations has decreased their feelings of anxiety (Glass & Singer, 1972). Wise et al. (1994) investigated the perceived control hypothesis further. They examined the effect of providing examinees with additional control by allowing them to choose the test mode they would be administered (CAT or SAT). Results indicated that for examinees exhibiting high levels of test anxiety, those given a choice of test type performed better than examinees assigned to a test type. They also found that higher anxiety was associated with an increased proportion of examinees choosing the SAT. Taken together these results indicate that both control over item difficulty and control over choice of exam format have the ability to decrease anxiety in examinees. Perceived control is threfore more important for examinees that are highly anxious, because lowering their anxiety often results in increased test performance (Wise, 1994).

In a study by Shermis et al. (2001), the researchers used four test conditions to examine the differences between CATs and SATs in regards to test anxiety and measures of efficiency. These included (1) a traditional CAT; (2) an individual SAT, where examinees select the difficulty of each item; (3) a global SAT, where examinees select the difficulty range of the entire test at the beginning; and (4) a placebo SAT, where examinees were asked to

select the difficulty of each item, but items were actually selected based on a CAT algorithm. Additionally, the researchers examined whether item feedback (if the item was answered correctly or not) had an effect on anxiety and efficiency when taking a CAT or a SAT. They assessed differences in test condition in terms of feedback, test length, test time, ability, test anxiety and test satisfaction. The results indicated that the only significant difference between the conditions was test length. The tests taken under the global SAT condition proved to be statistically significantly longer compared to the tests taken under the other three conditions. However, this effect was modest, as the test included on average only about two more items. The researchers found no significant differences between the individual SAT, placebo SAT, and CAT conditions. These results lend support for the use of SATs, since measurement precision and efficiency do not appear to be compromised. Despite this encouraging finding, the researchers failed to replicate some of the benefits of SAT compared to CAT found in previous research (Shermis et al., 2001). They found no significant differences between the conditions for test anxiety or satisfaction with the testing situation. The only significant result was higher test anxiety scores for females compared to males. This result is most likely explained by the fact that the examinees were taking a math test, where differences in anxiety levels are commonly associated with stereotype threat (Shermis et al., 2001). The researchers found no significant interaction effects between gender, test mode and anxiety in this study either. One last result of notable interest from this study was a lack of significant feedback differences across conditions. This is inconsistent with previous research that shows item feedback is associated with decreased administration time, which holds especially true with increased levels of test anxiety (Shermis et al.). However, the non-significant results do provide evidence that item feedback has little impact on the efficiency of adaptive testing.

In an attempt to obtain more stable estimates of the differences found in the various studies of self-adapted and computer-adaptive tests, Pitkin and Vispoel (2001) conducted a meta-analysis. The goal was to evaluate the magnitude of mean differences in both proficiency estimates and anxiety levels across the 15 studies to date. These studies included 19 experiments reporting proficiency estimates and 9 experiments reporting post-test anxiety levels. The results indicated that examinees taking SATs had proficiency estimates that were on average 0.12 standard deviations higher than examinees taking CATs, but this effect is modest. Pitkin and Vispoel hypothesized that the higher proficiency estimates from SATs were due to bias; that is, SATs yielded more positively biased estimates than CATs because examinees had the ability to control item difficulties. However, just because examinees can obtain more biased estimates does not mean that they necessarily do (Pitkin & Vispoel, 2001). An alternative explanation is that examinees taking the SATs experienced reduced negative effects of test anxiety, and that affected their test performance (Rocklin, 1996). The results of the meta-analysis also supported this hypothesis (i.e., SAT post-test anxiety scores were lower than those associated with CATs by 0.18 standard deviation units).

These results have led some experts to conclude that SAT is a better model choice for estimating ability levels of examinees with test anxiety (Rocklin, 1996; Wise et al., 1994). However, the benefits of SAT's higher test scores and decreased test anxiety come at a price, and they might not be practical to implement operationally. Both efficiency and measurement precision are significantly lower for SATs in comparison to CATs (Wise, 1994). The extra time required for examinees to choose their item difficulty levels makes the SAT longer to administer. This is due to the additional directions required for SATs and the additional time examinees spend on choosing and verifying the difficulty level of each item (Pitkin & Vispoel, 2001). Previous studies have found that SATs can take up to 45% longer to administer than CATs when both tests are terminated at the same level of precision (Pitkin & Vispoel, 2001).

The less precise measurement of ability estimates is another disadvantage of SATs. CATs provide more reliable proficiency estimates because the algorithm chooses items at each iteration that maximally reduce measurement error for the examinee's estimated ability level. The reliability of SAT is lower because examinees do not choose items that are as optimally targeted to their ability level. Therefore, SATs typically yield estimates with higher standard errors (Wise, 1994). In addition, because examinees do not always select difficulty levels best matched to their proficiency levels, biased or inflated estimates can result (Pitkin & Vispoel, 2001). CATs use item-selection algorithms designed to match item difficulties to each examinee's trait level; therefore, they provide greater protection against biased estimates compared to SATs (Pitkin & Vispoel, 2001). One way to address this issue of SAT's biased estimates is to force examinees to choose appropriately higher or lower item-difficulty categories. However, some research has shown that this might actually undermine the positive effects of SATs (Pitkin & Vispoel, 2001). Pitkin and Vispoel (2001) reported a study that restricted item selection in a SAT and found post-anxiety scores that were significantly higher than post-anxiety scores from an unrestricted item selection SAT. Additionally, the scores were not significantly different from the scores

obtained from a CAT. Thus, it appears that the anxiety reducing benefits of SATs are eliminated if a restricted item-selection procedure is implemented.

The issues raised here indicate that implementing an SAT on a large-scale basis does not appear to be practical, despite the potential benefit for some test takers. Weighing the time and cost to create the algorithms and sufficient sized item banks for these tests against the modest benefits is an unrealistic undertaking for most educational purposes (Pitkin & Vispoel, 2001). This is a real shame when considering high-stakes testing, because an assessment that reduces examinee anxiety and subsequently provides a more accurate estimate of ability is advantageous to use when making high-stakes decisions.

3. CAT and the Common Core

3.1 What Impact will the Implementation of a CAT for the Common Core have on the Assessment of Student Ability?

One of the primary purposes of the Common Core Standards Initiative is to develop a set of standards that are common across states to help ensure that students are receiving a consistent education from state to state (Common Core State Standards Initiative). The development of the new Consortia assessment designs, therefore, is intended to provide a common and consistent measure of student performance across states (Common Core State Standards Initiative). The goal is to have a common metric that allows states to more accurately compare the abilities of their respective students. While this is a novel and ideal concept, the fact that two different assessment systems are being developed raises many questions and concerns. Namely, how does this affect the comparability of student scores across the two Consortia?

With the advancement of technology and the role computer-adaptive testing has played in assessment, more and more studies have investigated the comparability of computer- and paper-administered tests and explored the possibility of differential effects. Wang and Kolen (2001) noted that establishing score comparability between CAT and paper test versions is challenging for two reasons. First, there are potential differences between items presented online versus on paper. Second, there are differences arising from the adaptive nature of CAT tests versus the non-adaptive nature of paper-and-pencil administrations. The second issue is more likely to be the major concern with the Consortia, as both assessments are computer based, but only one is adaptive. In a review of K-12 comparability studies since 1993, Paek (2005) reported that the majority of studies have indicated that computer-based tests are equivalent in difficulty or slightly easier than paper-based tests. However, of the 21 studies reviewed, only four investigated the test performance of students in grades 1-5, with the majority investigating the test performance of high school students. Of these four studies looking at younger students, two indicated that the computer version was more difficult; one indicated that the paper test was more difficult, and one concluded that the two tests were comparable. The varying findings with a younger population are concerning for the Common Core assessments, as students are tested beginning in grade 3.

More recently, Kingston (2009) conducted a meta-analysis to compare the performance of students on computer- and paper-administered multiple-choice tests in K-12 populations. He synthesized the results of 81 studies performed during the period of 1997-2007 and looked at whether grade or subject had an impact on comparability. The results confirmed that small differences exist between the two modes of administration, but the effect sizes were small (-.01) across the 81 studies. However, two additional findings raised some concerns. First, although effect sizes were small, the subject matter of the test did appear to affect the comparability of scores. CAT was more advantageous for English Language Arts (ELA) and Social Studies, but the paper administration was more advantageous for Math. This difference was attributed to the constant switching of focus between two and three dimensions when given math problems in a computer format. Although minimal, these differences are alarming because of their potential to influence ability measures. Since both Consortia have proposed computer-based formats, subject matter differences are an issue that each must address.

The second concern involves the impact of socioeconomic status (SES) on test performance. The few studies analyzed in the meta-analysis that examined SES showed inconsistent findings. This has the potential for serious consequences, as SES varies greatly across the United States. SES is a well-established subgroup difference that can impact the validity of test score interpretation (Wang & Kolen, 2001). Students lacking fewer opportunities to be tested with CAT might experience higher trait and computer anxiety levels. Some empirical evidence supports this idea. For example, Fritts and Marszalek (2009) found a significant correlation between test modality and computer anxiety scores. Students from the district taking the paper-and-pencil test had higher computer anxiety compared to the students from the district taking the CAT. Their explanation for this finding was that students from the district taking the paper-and-pencil tests were of lower SES than students from the district taking the CAT. Also, the students taking the paper-and-pencil test had had fewer opportunities to work

on a computer than students taking the CAT. Therefore, in order to ensure that assessments provide accurate estimates of ability for low SES students, researchers need to conduct additional studies of these students' performances on CATs. Kingston (2009) concluded that the sparse and inconsistent research on this particular subpopulation warrants further examination into the differences between paper- and computer-administered tests.

Another consideration is that these studies have largely been about investigating the comparability of transferring an already established paper-and-pencil assessment system to a computer-based test (CBT) and arguing that measures of student ability using both versions of the test are comparable. Administrations of CBTs are roughly comparable to administrations of traditional paper-and-pencil tests, but how do score and ability measure comparability translate when comparing ability levels from two independent assessments given in two different test formats? This is the issue that the Common Core Consortia must attend to. It is somewhat comforting to know that both proposals included some discussion about the need to investigate score comparability, stating their intent to examine the extent of score comparability across the states and Consortia. However, neither Consortium has provided any details as to how they might go about this process (PARCC; SBAC). Rather, they both identify the establishment of comparability as an issue and simply state that the assessments will provide comparable and consistent measures of student ability within each Consortium of states and across the two Consortia. The research indicating that ability estimates vary across FIT and CAT conditions, however, suggests that it won't be easy to obtain score comparability between the assessments that the Consortia administer. A number of comparability studies will be needed to ensure that students taking the SBAC assessment are not unfairly advantaged, since research indicates that adaptive tests provide more accurate measures of ability.

Despite the findings from numerous studies that score comparability between computer- and paper-administered tests can be achieved, some professionals in the measurement field still have concerns about using adaptive testing in high-stakes assessment settings. A paper sponsored by the Rennie Center for Education Research and Policy outlined the practical challenges that exist when implementing a CAT for large-scale assessment: significant time and cost to develop a well constructed item pool, sufficient computer access, complex administration software, and resources for technical support and maintenance (Reckase, 2011). The challenges are substantial and raise the issue of the practicality of successfully implementing a large-scale CAT. Reckase (2011) recommended strategic and careful planning in the development process and pilot testing to ensure proper implementation. However, there is a concern that the extent of the challenges posed by CAT could lead to failure to address psychometric complexities such as comparability (Reckase, 2011).

In 2010, the Educational Testing Service (ETS), Pearson and the College Board collaborated to prepare a paper on adaptive testing in a Common Core assessment system (Way et al., 2010). They discussed considerations related to the implementation and use of adaptive testing and supplied recommendations on the best practices for creating a common assessment system. They asserted that the technology of CAT holds significant promise for the Common Core assessments, but the process must be pursued cautiously and deliberately (Way et al., 2010). Few adaptive testing models have thus far been used in K-12 for a large-scale summative assessment purpose, and limited research is available on the impacts of testing in this regard. Although little evidence suggests that negative effects are continuing concerns for current adaptive testing programs, the criticism that they could affect the scores for individual test takers has not been sufficiently addressed (Way et al., 2010). This becomes especially important when considering adaptive testing for high-stakes assessments, because there is still a concern about threats to valid scores and interpretations (i.e., test anxiety). The concerns from Pearson and The College Board discussed above suggest that the SBAC faces considerable difficulties when implementing its CAT, and that it is imperative to overcome the challenges, given the lasting implications that the state assessments could have on students.

3.2 How Might the Implementation of the Common Core Assessments Affect Test Anxiety?

The SBAC will not only encounter normal and technical testing logistics such as those discussed above, but students' individual differences, such as test anxiety, gender and SES, are likely to present problems as well. Both areas are of concern because of their potential impact on obtaining accurate and reliable measures of student ability. The Rennie Center, ETS, Pearson, and The College Board have all raised concerns, providing information that the Consortia need to take into consideration when developing the tests and identifying technical constraints that the Consortia will face that may be problematic during implementation. However, these organizations have failed to address the issue of how individual differences are likely to impact the assessment process. Not only are the constraints that these variables could place on the implementation largely absent from the discussion, but also missing is how the implementation of a CAT for this purpose will impact *students*.

Statements such as "CBT benefits English Language Learners (ELL)" and "more accurately assesses student ability" have been thrown out on the respective websites, but there is no mention of research plans by either consortium to look into the role that individual differences play in performance on their planned assessments (PARCC; SBAC). According to the *Standards for Educational and Psychological Testing*, this is a critical component in the test development process for establishing validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

The lack of attention to these variables and the compelling research on CAT's role in test anxiety raises two questions. First, will we see a rise in the prevalence of test anxiety among students? As mentioned earlier, test anxiety rates have been steadily on the rise over the past decade, perhaps due to the increasing numbers of assessments administered. Additionally, research supports the claim that CAT's heighten the effects of test anxiety among students. Therefore, it is plausible to suggest that implementing a CAT assessment on this large of a scale will influence the prevalence of test anxiety across the country. Second, how will the implementation affect the measurement of ability for children suffering from test anxiety? The research discussed above suggests that CAT does impact their test performance. Will student's with test anxiety receive a fair representation of their ability level? Going back to the idea of perception of control, students with test anxiety are more adversely affected by lack of control in terms of test performance. Perceived control raises another concern regarding fairness comparability. Will students with test anxiety who are administered the CAT be at a disadvantage compared to those students with test anxiety who are administered the FIT? Test anxious student's taking the CAT test might receive an inaccurate measure of ability and be unfairly compared to all non-test anxious students as well as other test-anxious students that take the fixed-item test.

The concern of unfairly representing test-anxious student's ability in comparison to other test anxious students is highlighted when considering research findings regarding the interaction between test mode and test anxiety. The research implies that student's with test anxiety perform better with FIT. Thus, not implementing a CAT for all students might discriminate against students suffering from test anxiety. Since only the SBAC plans to use adaptive testing, test anxious students taking the SBAC exam are unfairly disadvantaged when compared to test anxious students taking the PARC exam. Will issues of fairness arise for those states signed on for the SBAC assessment?

Fairness in testing is a serious issue, because it can undermine the validity of test score interpretations. According to the *Standards for Educational and Psychological Testing*, test developers are obligated to adhere to strict guidelines to assess bias and fairness and ensure that testing outcomes for examinee subgroups are comparable (AERA et al., 1999). Standard 7.1 states:

> When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. (AERA et al., 1999, p. 80)

The lines of credible research regarding differential impact of test scores from test-anxious examinees taking CATs are well established. Thus, the standard requires SBAC to address test anxiety and provide two sets of validity evidence to investigate test score meaning for subgroups of test-anxious students.

> Additionally, according to Standard 7.11,
>
> When a construct can be measured in different ways that are approximately equal in their degree of construct representation and freedom from construct-irrelevant variance, evidence of mean score differences across relevant subgroups of examinees should be considered in deciding which test to use. (AERA et al., 1999, p. 83)

In regards to this standard, fairness issues could become a problem for states using CAT versus FIT assessments because CATs and FITs are designed to measure the same construct of student achievement as defined by the common core standards. A major problem arises here if children with test anxiety who take CATs are differentially impacted, but children with text anxiety in other states that aren't using CATs are not similarly impacted, since research suggests that use of other assessment formats does not result in these children receiving lower scores. Students with test anxiety living in the SBAC region may be unfairly disadvantaged because of their inability to access the equivalent test form provided by states using the assessment designed by PARCC.

These concerns emphasize the importance for the SBAC to not only conduct its own research but also utilize the current CAT and test anxiety research. Previous studies provide results and suggestions that could help the SBAC effectively deal with these concerns. Despite the promising research regarding the benefits of SAT for

test-anxious examinees, the capability of the SBAC to undertake implementing an SAT is highly unlikely. However, Ortner and Caspers's (2011) finding that possibly familiarizing students with the adaptive testing procedures might decrease the negative impact of adaptive testing on their test scores deserves further examination. If the consortium does develop an appropriate and effective adaptive testing model, this simple procedure could offer a plausible solution to the problem of differential effects for test-anxious students that CAT poses.

4. Concluding Thoughts

The issues raised in this paper make it apparent that more research is warranted concerning the use of a high-stakes summative assessment system for assessing the Common Core standards. Ensuring that the measures of student ability obtained across the states are comparable in meaning is one of the core rationales behind the development of these assessments. Therefore, the Consortia's lack of attention to issues of differential effects and individual differences is surprising, because those issues could have a direct bearing on the comparability of scores. Unfortunately, there have been few studies investigating the comparability of scores from CATs and FITs, and the studies that have been published report inconsistent results, which is disconcerting. Currently, there is not enough evidence to determine whether differences in the modes of test administration will have a major impact on the comparability of student scores. Another issue is the samples used in the studies that have examined test anxiety and CAT. The majority of these studies used undergraduate students and did not examine these effects in K-12 populations. The effects of test anxiety on scores from conventional modes of testing have largely been shown to be similar in school-age children (Hembree, 1988). However, since the Common Core assessments are geared for K-12 students, it is important to investigate how the role of CAT in test anxiety impacts students at each grade level because the impact might be different for younger students.

In the end, the million-dollar question still remains: Can the two different Common Core assessments proposed by the Consortia really provide common measures of student performance? Until more research is provided, a definitive answer to this question is difficult. One thing is certain: When policy makers and school administrators are basing more and more high-stakes decisions on the results from assessments, we are obligated to explore and find a solution to the growing problem of test anxiety for all modes of testing administration.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Casbarro, H. (2005). *Test anxiety and what you can do about it: A practical guide for teachers, parents, and kids.* Port Chester, NY: Dude Publishing.

Cassady, J. C. (2004). The influence of cognitive test anxiety across the learning-testing cycle. *Learning and Instruction, 14*, 569–592. http://dx.doi.org/10.1016/j.learninstruc.2004.09.002

Common Core State Standards Initiative. (n.d.). *About the standards*. Retrieved from http://www.corestandards.org

Davey, T., & Pitoniak, M. J. (2006). Designing computer-adaptive tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 543-573). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Dragsaw, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-515). Westport, CT: Praeger Publishers.

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effects of adaptive testing on test taking motivation. *Diagnostica, 55*, 20-28. http://dx.doi.org/10.1026/0012-1924.55.1.20

Fritts, B. E., & Marszalek, J. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13*(3), 441-458. http://dx.doi.org/10.1007/s11218-010-9113-3

Glass, D. C., & Singer, J. E. (1972). Behavioral aftereffects of unpredictable and uncontrollable aversive events. *American Scientist, 60*, 457-465.

Hembree, R. (1988). Correlates, causes, effects and treatment of text anxiety. *Review of Educational Research, 58*, 47–77.

Hill, K. T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *The Elementary School Journal, 85*(1), 105-126. http://dx.doi.org/10.1086/461395

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22,* 22-37. http://dx.doi.org/10.1080/08957340802558326

Methia, R. A. (2004). *Help your child overcome test-anxiety and achieve higher test scores*. College Station, TX: VBW.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Mulvenon, S. W., Stegman, C. E., & Ritter, G. (2005). Test anxiety: A multifaceted study on the perceptions of teachers, principals, counselors, students, and parents. *International Journal of Testing, 5*, 37–61. http://dx.doi.org/10.1207/s15327574ijt0501_4

Nicholls, J. G. (1990). What is ability and why are we mindful of it? A developmental perspective. In R. J. Sternberg & J. Kollogian (Eds.), *Competence considered* (pp. 11-40). New Haven: Yale University Press.

Olea, J., V., Revuelta, J., Ximenez, M.C., & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicolog á , 21*, 157-173.

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment, 27*(3), 157-163. http://dx.doi.org/10.1027/1015-5759/a000062

Paek, P. (2005). *Recent trends in comparability studies*. Austin, TX: Pearson Educational Measurement.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2010, August). *On the road to implementation: Aligning assessments with the Common Core State Standards.* Retrieved from http://www.parcconline.org

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement, 38*(3), 235-247. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01125.x

Putwain, D. W. (2008). Deconstructing test anxiety. *Emotional and Behavioral Difficulties, 13*(2), 141-155. http://dx.doi.org/10.1080/13632750802027713

Putwain, D. W., & Daniels, R. A. (2010). Is the relationship between competence beliefs and test anxiety influenced by goal orientation? *Learning and Individual Differences, 20,* 8–13. http://dx.doi.org/10.1016/j.lindif.2009.10.006

Putwain, D. W., Woods, K. A., & Symes, W. (2010). Personal and situational predictors of test anxiety of students in post-compulsory education. *British Journal of Educational Psychology, 80,* 137-160. http://dx.doi.org/10.1348/000709909X466082

Reckase, M. D. (2011, May). Computerized adaptive assessment (CAA): The way forward. In *The Road Ahead for State Assessments, Policy Analysis for California Education and Rennie Center for Education Research & Policy.* Cambridge, MA: Rennie Center for Education Research & Policy.

Rocklin, T. R., (1996). Self-adaptive testing: Improving performance by modifying tests instead of examinees. *Anxiety, Stress, and Coping, 10,* 83-104. http://dx.doi.org/10.1080/10615809708249296

Shermis, M. D., Mzunara, H. R., & Bublitz, S. T. (2001). On test and computer anxiety: Test performance under CAT and SAT conditions. *Journal of Educational Computing Research, 24*(1), 57-75. http://dx.doi.org/10.2190/4809-38LD-EEUF-6GG7

Skinner, E., Furrer, C., Marhcund, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology, 100*(4), 765-781. http://dx.doi.org/10.1037/a0012840

Smarter Balanced Assessment Consortium. (2010). *Smarter Balanced Assessment Consortium: A summary of Common Core components.* Retrieved from http://www.k12.wa.us/smarter/

Tonidandel, S., & Quiñones, M.A. (2000). Psychological reactions to adaptive testing. *International Journal of Selection and Assessment, 8*, 7-15. http://dx.doi.org/10.1111/1468-2389.00126

Tonidandel, S., Quiñones, M. A., & Adams, A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology, 87,* 320-332. http://dx.doi.org/10.1037/0021-9010.87.2.320

Triplett, C. F., Barksdale, M. A. & Leftwich, P. (2003). High stakes for whom? Children's perceptions of high stakes testing. *Journal of Research in Education, 13*(1), 15–21.

Vispoel, W.P. (1993). Computerized adaptive and fixed item versions of the ITED vocabulary subtest. *Educational and Psychological Measurement*, 53, 779-788. http://dx.doi.org/10.1177/0013164493053003022

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01115.x

Way, W., Twing, J., 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 www.ets.org/s/commonassessments/pdf/ AdaptiveTesting.pdf

Wise, S. L. (1994). Understanding self-adapted testing: The perceived control hypothesis. *Applied Measurement in Education, 7*(1), 15-24. http://dx.doi.org/10.1207/s15324818ame0701_3

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational measurement: Issues and practice, 8*(3), 5-10. http://dx.doi.org/10.1111/j.1745-3992.1989.tb00324.x

Wise, S. L. Roos, L. L., Plake, B. S., & Nebelsick-Gullett, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. *Applied Measurement in Education, 7*(1), 81-91. http://dx.doi.org/10.1207/s15324818ame0701_6

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York, NY: Plenum Press.

Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, and future directions. In M. Zeidner, P. A. Schutz, & R. Pekrun, (Eds.), *Emotion in education, Educational psychology series* (pp. 165-184). San Diego, CA: Elsevier Academic Press. http://dx.doi.org/10.1016/B978-012372545-5/50011-3

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647-679). Westport, CT: Praeger Publishers.