A Monte Carlo Simulation Investigating the Validity and Reliability of Ability Estimation in Item Response Theory with Speeded Computer Adaptive Tests

T. A. Schmitt  
*Eastern Michigan University, Michigan, USA*

D. A. Sass and J. R. Sullivan  
*University of Texas at San Antonio, Texas, USA*

C. M. Walker  
*University of Wisconsin-Milwaukee, Wisconsin, USA*

Imposed time limits on computer adaptive tests (CATs) can result in examinees having difficulty completing all items, thus compromising the validity and reliability of ability estimates. In this study, the effects of speededness were explored in a simulated CAT environment by varying examinee response patterns to end-of-test items. Expectedly, ability estimates became increasingly negatively biased as the CAT became more speeded, with the bias magnitude depending on the speededness condition. Ability estimates for higher-ability examinees were also influenced more than estimates for low- and middle-ability examinees with the realistic item pool. This finding is likely linked to test operational characteristics since these results were not replicated with an ideal item pool. In any case, the CAT is relatively robust to the speededness assumption, assuming that only a few items are speeded. Results also suggested that if test developers score the test at the point of speededness, ability estimates remain unbiased. This study’s findings help inform test developers and test takers about the effects of test speededness within a CAT environment. Additional implications for educators and researchers are discussed.

---

This research was partially funded by the 2004 Graduate Student Research Program sponsored by the Association of American Medical Colleges.

Correspondence should be sent to Thomas A. Schmitt, Eastern Michigan University, Department of Psychology, 537K Mark Jefferson, Ypsilanti, MI 48197. E-mail: tschmit1@emich.edu
Computer adaptive tests (CATs) offer several advantages over traditional paper-and-pencil (P&P) tests (Meijer & Nering, 1999), which has resulted in the increased popularity of CATs for assessing general ability1 or aptitude in such areas as certification, licensure, professional school examinations, and k-12 educational settings (see van der Linden & Glas, 2000; Wainer et al., 2000; Weiss, 2009). Some of the most notable examples of CATs measuring aptitude include large-scale testing programs such as the Graduate Record Examinations (GRE) (van der Linden & Glas, 2000) and the Armed Services Vocational Aptitude Test Battery (ASVAB) ( Sands, Waters, & McBride, 1997). CATs have also been developed to evaluate psychological disorders, such as depression (e.g., Fliege, Becker, Walter, Rose, Bjorner, & Klapp, 2009; Simms & Clark, 2005; Walter & Holling, 2008). According to Weiss (2009), numerous CATs are operational or in development throughout the world. Because of the potential benefits of CATs, policymakers have become excited about the potential CATs have in the k-12 educational setting (Way, 2005) for assessing formative learning objectives (see Chalhoub-Deville, 1999). For example, Cito and the National Institute for Educational Measurement in The Netherlands developed CATs designed for ages 4–12 to assess basic arithmetical skills, spelling ability, reading comprehension, vocabulary, and world orientation. Other notable k-12 tests are the Learning Potential CAT (LPCAT) and the Measures of Academic Progress (MAP) (Weiss, 2009). Given the increased use of CATs for broader purposes in a wider variety of educational contexts and settings, there is a more urgent need to investigate the validity and reliability of scores provided by CATs.

Many educational assessments are intended to be power tests, but when tests have imposed time limits, they can become speeded tests (Evans & Reilly, 1972; Schnipke, 1995; Lu & Sireci, 2007). Gulliksen (1950) was one of the first to delineate speed and power tests. He defined pure speed tests to have many items of low difficulty (i.e., easy items) that examinees rarely answer incorrectly, but the test length is such that no examinee can complete the test in the allotted time. Thus, a pure speed test simply measures how many items the examinee can answer correctly within a limited time-period. Pure power tests allot ample time for all examinees to respond to the same number of items (Gulliksen, 1950), with the final score being estimated with models from classical test theory (CTT) or item response theory (IRT). Power tests measure an examinee’s ability without time constraints, whereas speed tests measure the ability to answer items quickly. The most appropriate type of test depends on the assessment purpose. If interest were in constructs such as processing speed or ability to reason quickly, a speeded test would be more appropriate. However, if constructs measuring content mastery or
aptitude were of interest, where decision speed is not a factor, a power test would be more suitable. For practical purposes, many adaptive assessments combine both speed and power.

Due to the combination of speed and power (Rindler, 1979), many examinees, such as English language learners (ELLs), do not have sufficient time to complete the assessment (Bejar, 1985), resulting in speeded tests. Being aware of the causes of speededness and how speeded tests can affect examinees is particularly important from both fairness- and score-oriented perspectives (Lu & Sireci, 2007), since a test may become biased against certain disadvantaged groups due to examinee characteristics (e.g., ELLs, students with non-diagnosed disabilities, etc.), test operational characteristics (e.g., time-limit induced speededness, item characteristics, item pool characteristics, the IRT model, etc.), or a combination of both. When the majority of examinees are hurried, speededness becomes an operational issue and making operational adjustments becomes necessary (Bridgeman & Cline, 2004). However, there are instances when only certain groups of examinees are disadvantaged. For example, ELL students may experience speededness because English is not their native language and they cannot read test items quickly enough. Of course, students can improve their reading speed and learn to answer items more quickly, but test developers should also empirically evaluate examinee response times. If speededness is present, they should consider making operational adjustments (e.g., increase testing time, expand the item pool, model speededness) to mitigate potential speededness effects and better estimate ability. If a test is to be fair and unbiased, examinees should not be subject to unfair test operational characteristics, such as time-limit induced speededness (Bejar, 1985; Lu & Sireci, 2007; Schnipke, 1995).

Because test bias and fairness are linked to validity, tests that produce biased or unfair proficiency estimates may lead to invalid interpretations of examinee test scores and incorrect decisions on behalf of the test-taker, such as exclusion from graduate school (Wright, 2008). As put forth by the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), validity is central to having test scores that are interpretable for their intended use. In other words, if test scores and assessment results are to be valid indicators of the underlying construct of interest (e.g., ability), they must measure what they are purported to measure and facilitate accurate decision making. Further, the construct measured by educational achievement tests is usually not defined to include response time as part of the measured construct, and test administrators usually do not study or report empirical results from studies of speededness to ensure it is not negatively impacting student performance. Therefore, if certain groups are disadvantaged because of test speededness, the factor structure and construct validity of the test is in question (Lu & Sireci, 2007).
Because CATs are grounded within a theoretical framework known as IRT, one of the underlying assumptions of IRT, and consequently, CAT, is that of local item independence. For this assumption to hold, item responses must be determined by ability and be independent of other factors. In addition, if local item dependency (LID) occurs, ability estimates and their respective standard errors may be invalid and unreliable (Goegebeur, De Boeck, Wollack, & Cohen, 2008). One of the major causes of LID is test speededness because item responses are then correlated due to test speededness rather than examinee ability (Yen, 1993). Speededness commonly affects responses to end-of-test items because examinees frequently randomly guess on or omit end-of-test items (Bejar, 1985; Boughton & Yamamoto, 2007; Bridgeman & Cline, 2004, 2000; Lu & Sireci, 2007; Oshima, 1994; Schnipke & Scrams, 1997; Yamamoto, 1995), which produces biased and underestimated ability estimates when using common IRT models (Bridgeman & Cline, 2004; Goegebeur et al., 2008). To avoid LID and prevent the negative effects of test speededness, examinees must have adequate time to complete the entire test, including end-of-test items, or testing administrators must make other operational adjustments to lessen the effects of speededness (Bridgeman & Cline, 2004; Goegebeur et al., 2008).

Several research studies have implicated speededness in affecting the test factor structure with findings that racial minority examinees often take longer to complete the SAT-Verbal sections compared to White examinees (Dorans, Schmitt, & Bleistein, 1992; Lawrence, 1993; Schmitt & Dorans, 1988). Speededness is also problematic for ELLs (Grabe, 1999; Larson, 1999) who commonly score lower on achievement measures than students with English as their primary language (Abedi & Lord, 2001; Durán, 2008). Under these circumstances, justifiable differences between the groups cannot be discerned as they may emerge due to language barriers and not true differences in ability. Moreover, this problem may be exacerbated within a CAT environment for groups (e.g., low socioeconomic status, etc.) unfamiliar with the computerized administration system. Since the goal of testing is to produce unbiased ability estimates and research has shown speededness can disadvantage certain groups of examinees because of incorrect score interpretation (e.g., Sireci, 2005; Sireci, Scarpatici, & Li, 2005), it is important to assess how speededness affects the accuracy of examinee proficiency estimates in an effort to provide direction for future development and operational implementation of CATs.

Although the detection of speededness is impractical in P&P tests, computer-based adaptive tests often record examinee response times, making the detection of speededness possible (Bridgeman & Cline, 2000, 2004; Schnipke, 1995; Schnipke & Scrams, 1997; van der Linden & van Krimpen-Stoop, 2003; Lu & Sireci, 2007). For example, Bridgeman and Cline’s (2004) examination of examinee response times on the GRE analytic section revealed that more than half of the examinees were hurried on the final six questions. Although the analytic GRE was discontinued in 2002, hybrid power CATs with time limits continue to be...
used and are expanding into the k-12 setting. As Schnipke (1995) and Lu and Sireci (2007) described, time limits do not necessarily mean a test is speeded, but if a significant portion of examinees are hurried or certain examinees are forced to rapidly guess on items, then the test is speeded and operational adjustments should be made. Thus, speeded tests usually result from power tests with strict time limits (Lu & Sireci, 2007). Maintaining the quality and practicality of an adaptive test is costly, and if examinees are allotted unlimited time to complete adaptive tests the statistical benefits would be negated by added impracticalities, increased administrative costs, and possible threats to test security. Therefore, time limits are routinely imposed when administering CATs, even if some examinees do not have sufficient time to complete the test. Since one advantage of CAT over conventional P&P tests is that fewer items can be administered and still result in accurate ability estimates (Wainer et al., 2000; Weiss, 1982), it could be argued that the time saved should be extended to examinees. Indeed, a balance is needed to ensure tests are unspeeded, but are still completable in a reasonable amount of time.

Modeling Speededness within CAT

Although there have been studies examining the robustness of IRT to assumption violations, few studies have assessed how speededness affects an examinee’s performance within a CAT environment. Because the local independence assumption is often violated by speeded CATs, thus impacting construct validity, most educational assessments are not meant to incorporate speededness as a part of the construct. Moreover, many CAT developers do not empirically evaluate their CAT to determine whether it is speeded, and if so, how examinees are affected. Additional research is needed on how operationally induced speededness can interact with other operational factors, such as item pool characteristics, to bias examinee test scores. This research is particularly important for marginalized groups of test takers whose test scores may be more affected by speeded CATs and prone to invalid ability estimates.

Many promising CAT models and methods are available that model and/or account for speededness. Response-time (RT) modeling began with Rasch’s (1960/1980) models for misreading in a text and reading speed. Roskam (1997), Schnipke and Scrams (2002), van der Linden (2009a), and van der Linden (2009b) give a thorough review of RT models. Van der Linden has also conducted extensive and valuable research in an effort to control for or model speededness. Van der Linden, Scrams, and Schnipke (1999) proposed one of the first realistic models in an effort to account for the effects of speededness within CAT. Other models/methods proposed to account for speededness include a lognormal model of response times (van der Linden, 2006), a hierarchical model for modeling speed and accuracy (van der Linden, 2007), a model that uses response times for item selection in CATs (van der Linden, 2008), predictive control of speededness in
CATs (van der Linden, 2009b), and using response times as collateral information to improve parameter estimation (van der Linden, Klein Entink, & Fox, 2008).

Another line of speededness models began with Yamamoto and Everson’s (1997) hybrid model, which used multiple IRT models to model examinee behavior on end-of-test items. Bolt, Cohen, and Wollack (2001) extended Rost’s (1990) multi-class mixture Rasch model to a two-class mixture Rasch model in an effort to distinguish multiple latent classes of examinees in terms of their speeded response patterns to end-of-test items. This model was later extended to the 3-PL (3-parameter logistic) model (Bolt, Mroch, & Kim, 2003). Although these models performed well when examinees changed their response strategy at a single point in the test, they failed to model speededness gradually. In an effort to model this gradual process, Wollack and Cohen (2004) formulated a gradual process change IRT model. This model consists of a random problem-solving process (e.g., 3-PL) and a random guessing process, with the latter response strategy gradually supplanting problem solving for end-of-test items. The current study will use Wollack and Cohen’s (2004) model to simulate speededness, but Goegebeur et al. (2008) have also expanded their model to be estimable for test data.

Study Significance

As Wise and Kingsbury (2000) discussed, both examinee factors and operational test factors are central to implementing and maintaining successful adaptive tests. This is especially critical as CATs become more common in formative educational assessment settings and as examinee diversity increases. The goal of the current study was to investigate how CAT operational factors (i.e., time limit–induced speededness, item characteristics, and item pool characteristics, the IRT model) influence ability estimation across a range of abilities. Although test speededness can be considered both an examinee and operational factor, in the current study all examinee abilities were equally influenced by test speededness within each experimental condition. Thus, operationally induced speededness was simulated to result from strict time limits by having all simulees or examinees gradually increase their response rate on end-of-test items. This simulation complements the work of Bridgeman and Cline (2004) who studied examinee response times and found that most examinees rushed when responding to end-of-test items. Of course, test speededness is not a pure operational issue since examinee characteristics may also affect how quickly test-takers respond to items. Regardless of the cause, the bias estimates should remain constant whether speededness is an operational and/or examinee artifact, and test administrators should be aware that certain examinees can be disadvantaged by operational factors of CATs (Lu & Sireci, 2007).

Another important operational consideration is the IRT model utilized. Although research investigating speededness with common IRT models exists (e.g.,
Oshima, 1994), little to no research has examined the effect of test speededness on ability estimation within a CAT environment using standard IRT models (e.g., 3-PL) with realistic speededness conditions (Wollack & Cohen, 2004). Given the 3-PL model’s difficulty of handling examinee guessing at the end of CATs, and that it is commonly used by test developers (Bridgeman & Cline, 2004), it is important to evaluate the 3-PL model’s operational effectiveness in a simulated CAT environment.

As discussed, Bridgeman and Cline (2004) provided evidence that a large number of examinees are affected by extreme, but detectable, speededness on end-of-test items, and that potential bias resulting from speededness was greater for high-ability examinees. Their findings contributed to operational changes to the analytic GRE CAT; thus their research provides a real-world example of the practical implications of the current line of research. As Bridgeman and Cline (2004) stated, testing programs that employ CATs with strict time limits need to make operational adjustments to the CAT if significant portions of examinees are speeded or certain groups of examinees are affected by the speeded nature of the CAT. Thus, the main goal of the current study was to explore how CAT operational factors affect ability estimation across different abilities within a more controlled study that simulates a gradual increase in speededness at the end of the test (Goegebeur et al., 2008; Wollack & Cohen, 2004).

To understand these factors, three research questions were considered. First, as a test becomes more speeded and examinees are forced to answer end-of-test items more quickly, how robust is the 3-PL model? Stated differently, at what point should examinees, test developers, and/or researchers be concerned that the results are too biased due to speededness? Second, does the amount of bias vary as a function of ability level? If so, what ability levels are affected most (i.e., display the greatest bias) by speeded tests with typical item pools, and what CAT operational factors, such as characteristics of the item pool, may be affecting bias in the context of speeded CATs? Third, will the estimation stability (i.e., standard error of estimate) remain constant across speededness conditions and ability values? This study hopes to (1) bolster the understanding of how test speededness affects estimation error across the ability spectrum, (2) elucidate potential operational factors that test developers can modify to reduce estimation error, and (3) help test developers and test takers better understand how test speededness influences ability estimation error.

**METHOD**

**Item Parameters**

A realistic item pool consisting of 400 items was generated by reviewing the item pool characteristics of several CAT studies in an effort to represent a realistic
SPEEDEDNESS IN COMPUTER ADAPTIVE TESTING

CAT item pool (see Oshima, 1994; Papanastasiou & Reckase, 2007; Vispoel, 1998; Wang & Vispoel, 1998). A second ideal item pool of 400 items was also generated as a control condition to help describe the results generated from the realistic item pool. Item parameters for the 3-PL model (a-, b-, and c-parameters) were generated for both the realistic and ideal item pools. Table 1 describes the targeted distributional characteristics of the simulated realistic and ideal item pools.

The discrimination \((a)\) parameters for the realistic and ideal item pool were randomly drawn from a normal distribution with a mean of 1.10 \((SD = 0.10)\) and 1.90 \((SD = 0.10)\), respectively. The \(a\) parameter value of 1.90 was chosen for the ideal item pool because they are highly discriminating items that could be obtained in practice (Hambleton, Swaminathan, & Rogers, 1991; Wang & Vispoel, 1998). The difficulty \((b)\) parameters for the realistic and ideal item pool were randomly drawn from a normal distribution with a mean of 0.00 \((SD = 1.00)\) and 0.00 \((SD = 1.73)\), respectively. The \(b\) parameters for both realistic and ideal item pools were allowed to vary from \(-3.00\) to \(3.00\) to provide adequate coverage for the ability levels from \(-2.00\) to \(2.00\). Further, \(b\) parameters for the ideal item pool had an equal number of items in each interval to ensure consistent coverage across abilities. The guessing \((c)\) parameters for both the realistic and ideal item pools were fixed at 0.15.

Data Generation

The advantage of using simulated data is that the true test structure is known, as are the true ability \((\theta)\) and item parameters \((a, b, \text{ and } c)\). This allows for the systematic evaluation of different amounts of speededness on end-of-test items. The first step in the CAT algorithm was to order the items into an information table based on the amount of maximal item information for a particular ability, \(I_i(\theta_j)\) (Segall, Moreno, Bloxom, & Hetter, 1997; Thissen & Mislevy, 2000). For the 3-PL model, Birnbaum (1968) showed that \(I_i(\theta_j)\) can be written as (see Hambleton &
Swaminathan, 1985)

\[ I_i(\theta_j) = \frac{D^2a_i^2(Q_i)}{P_i(\theta_j)} \left[ \frac{(P(\theta_j) - c_i)^2}{(1 - c_i)^2} \right], \tag{1} \]

where \( Q_i = 1 - P_i(\theta_j) \). To ensure adequate item coverage, items were sorted at a particular ability from \(-2.0\) to \(2.0\) in increments of \(0.2\) resulting in \(21\) ability intervals. In other words, items were sorted so that low-ability examinees received less difficult items and high-ability examinees received more difficult items. Thus, items were selected based on the maximum amount of information they provided at a particular \(\theta\).²

The next step in the CAT simulation was to select an examinee’s true ability level, \(\theta\). A total of \(21,000\) simulees (i.e., simulated examinees) were generated by randomly selecting \(21\) \(\theta\)s from a uniform distribution of \(-2.0\) to \(2.0\) in increments of \(0.2\). Because the evaluation of the accuracy of estimation should not be assessed without carrying out replications (Harwell, Stone, Hsu, & Kirisci, 1996), each \(\theta\) was assigned \(1,000\) simulees.

The third step in the adaptive algorithm was to select the first item based on \(I_i(\theta_j)\) to administer to an examinee. It is common in a CAT context to assume that examinees come from a standard normal distribution with \(\mu = 0\) and \(\sigma = 1\). Therefore, the first item administered in a CAT is often an item of moderate difficulty. Although the first administered item was selected to be moderately difficult, research has shown that the difficulty of the first item has little effect on estimation accuracy (Lunz, Bergstrom, & Gershon, 1994).

Once an item was selected and an examinee with a known \(\theta\) was generated, the next step in the CAT simulation was to generate a response based on the selected \(\theta\) and item. Using the 3-PL model, the probability of the examinee answering the item correctly was generated and then compared to a random number drawn from a uniform distribution \([0,1]\). If the probability of a correct response was greater than or equal to the uniform random number, a correct response was recorded; otherwise, an incorrect response was recorded. Note the process of generating data was the same regardless of whether the test was speeded. However, the model to generate data did depend on whether or not the CAT was speeded as indicated below.

**Unspeeded CAT (Control condition).** The 3-PL model (Birnbaum, 1968) was used to generate unspeeded examinee response patterns. The 3-PL model is defined as (see Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991)

\[ P_i(\theta_j) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}, \tag{2} \]
where $P_i(\theta_j)$ is the probability of examinee $j$ ($j = 1, \ldots, n$) with ability $\theta_j$ responding to item $i$ ($i = 1, \ldots, n$) correctly, $a_i$ is the discrimination parameter, $b_i$ is the difficulty parameter, $c_i$ is the guessing or pseudo-chance level parameter, and the constant $D$ is a scaling factor set to 1.7.

**Speeded CAT (Conditions 1–9).** To generate speeded data, this study simulated speededness as a gradually occurring process. This is consistent with the notion that as the sum of response times of past items approaches the total time allotted for the test, examinees will begin to gradually reduce their item response times. To simulate this gradual proliferation of speededness the following model was used (Goegebeur et al., 2008; Wollack & Cohen, 2004)

$$P_i^*(\theta_j) = c_i + (1 - c_i) \left[ \frac{e^{D a_i (\theta_j - b_i)}}{1 + e^{D a_i (\theta_j - b_i)}} \times \min \left\{ 1, \left[ 1 - \left( \frac{i}{n} - \eta_j \right) \right]^{2 \lambda_j} \right\} \right], \quad (3)$$

where $P_i^*(\theta_j)$ is the probability of examinee $j$ with ability $\theta_j$ responding to speeded item $i$ ($i = 1, \ldots, n$) correctly, $\eta_j$ ($0 \leq \eta_j \leq 1$) is the speededness point parameter where examinee $j$ first experiences an effect due to speededness, $\lambda_j$ ($\lambda_j \geq 0$) is the speededness rate of examinee $j$, and $\min \{x, y\}$ is the smaller of the two values $x$ and $y$. To the left of the min term is the 3-PL model described in Equation 2, which reflects unspeeded response probabilities, and to the right is the speeded portion of the model. So with unspeeded CATs Equation 3 is simply the 3-PL model.

As $\lambda_j$ increases and/or $\eta_j$ decreases, the test becomes more speeded and the probability of answering correctly approaches $c_i$. Alternatively, as $\lambda_j$ decreases and/or $\eta_j$ increases, the test becomes unspeeded or approaches the 3-PL model. This model then consists of a random problem-solving process (3-PL model) and a random speededness process. So for end-of-test items, as more items are administered the CAT gradually becomes more speeded, thus simulating the realistic conditions of timed CAT.

Because examinees are likely to experience a gradual increase in speededness on the end-of-test speeded items, the probability of a correct response was decreased by varying the speededness parameters $\eta_j$ and $\lambda_j$ in Equation 3 according to the following distributions: $\eta_j \sim \text{Beta}(\alpha, \beta)$ and $\lambda_j \sim \log N(\mu_\lambda, \sigma^2_\lambda)$. For the speededness point parameter $\eta_j$, this was done by varying the amount of time left to respond to varying points near the end of the test, resulting in moderate, severe, and extreme levels of examinee speededness. This was simulated by randomly sampling from beta ($\beta$) distributions of (28.00, 2), (9.96, 2), and (5.50, 2). From $E(\eta_j) = \alpha / \alpha + \beta$, the simulation conditions were $E(\eta_j) = 0.93$, $E(\eta_j) = 0.83$, and $E(\eta_j) = 0.73$. The speededness rate parameter $\lambda_j$ was also varied to reflect three rates of speededness (low, medium, and high) by randomly sampling from normal distributions of (1.5, 1), (2.5, 1), and (3.5, 1) and then transformed to a lognormal
distribution by $\exp(\lambda_j)$. From $E(\lambda_j) = \exp(\mu + \sigma^2 / 2)$ the simulation conditions were $E(\lambda_j) = 54.60$, $E(\lambda_j) = 20.09$, and $E(\lambda_j) = 7.39$. Using this procedure, nine main speeded conditions and three secondary speeded conditions were simulated.

**Ability Estimation**

All ability estimates were computed using the Maximum Likelihood Estimation (MLE) procedure. MLE was selected given that Wang and Vispoel (1998) showed that CATs using MLE consistently produced less bias than Bayesian-based approaches, especially for extreme ability levels. Since an ML estimate of ability cannot be obtained until an examinee obtains at least one correct and one incorrect response, a method of trying to maximize the probability of obtaining a correct and incorrect response was written into the CAT algorithm. If a correct response was recorded, then a more difficult item was selected from two $\theta$ intervals higher in the information table. Alternatively, if an incorrect response was recorded, then a less difficult item was selected from two $\theta$ intervals lower in the information table.

After a useable response pattern (i.e., at least one correct and one incorrect) was obtained, estimated abilities ($\hat{\theta}$) were obtained using a Newton-Raphson iterative procedure to find the ML estimate. A new item, based on $\hat{\theta}$, was then selected from the information table in the range of the updated $\hat{\theta}$. Thus, as each item was selected, an estimate of the simulee’s ability was calculated from the current responses, and then the next optimal item was chosen.

The above procedure was carried out for a 30-item CAT given that MLE results in valid ability estimates when test lengths are greater than 20 items (Hambleton & Swaminathan, 1985; Lord, 1980; van der Linden & Pashley, 2000; Wang & Vispoel, 1998). It should be noted that when the Newton-Raphson iterative procedure was unable to converge on an ML estimate of ability, a new 30-item CAT was simulated. This procedure eliminated estimation bias due to non-convergence of ability estimates.

**Experimental Conditions**

The simulated 30-item CAT was used to create several experimental conditions: a single control condition (unspeeded test), nine main speeded conditions (see Table 2), and three secondary conditions. The 30-item unspeeded CAT acted as the control condition and provided an assessment of the algorithm’s fidelity. Because there is commonly little variation in average response time latencies at the beginning of tests (Bridgeman & Cline, 2004), this study simulated end-of-test items to be speeded for the speeded conditions (e.g., Boughton & Yamamoto, 2007; Bridgeman & Cline, 2004). Thus, each CAT was simulated to be speeded based on Bridgeman and Cline’s (2004) empirical observations and Wollack and Cohen’s (2004) simulation study of moderate, severe, and extreme levels of examinee speededness. These conditions provided a realistic proliferation of speededness,
as examinees of all ability levels experience different amounts of speededness (see Bridgeman & Cline, 2004).

In addition to the nine main speeded conditions, this study examined three conditions (Conditions 10–12) for treating end-of-test items at the different speededness points of moderate, severe, and extreme speededness. Thus, conditions 10–12 (moderate-ignore, severe-ignore, and extreme-ignore) ignored speeded end-of-test items. More specifically, Conditions 10–12 did not score the speeded items but instead provided an ability estimate at the last unspeeded item.

Across the conditions, item responses and ability estimates corresponding to the proportion of test items that were completed under normal conditions were identical for each simulated examinee. Therefore, only the items and the responses to test items simulated to be speeded were different for each examinee across the conditions. For example, for the three speeded conditions in which examinees did not have time to complete the last 9 test items, the first 21 item responses for each of the speeded conditions were identical to those that were utilized for the 30-item unspeeded CAT. The ability estimate obtained from these initial responses was also the same. This method of simulating the CAT algorithm allowed the estimation differences between the unspeeded and speeded conditions to be solely dependent on the speeded conditions.

**Evaluation of Estimation Accuracy**

This study evaluated the amount of estimation bias and stability for the control condition, the nine speeded conditions, and the three secondary conditions. Mean plots were used to describe and compare how different conditions of speededness affected estimation error. For each condition, bias, SE, and RMSE served as the dependent variables and were calculated as:

\[
Bias(\hat{\theta}_j) = \frac{1}{N} \sum_{k=1}^{N} (\hat{\theta}_{jk} - \theta_j),
\]  

(4)
\[ SE(\hat{\theta}_j) = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} \left( \hat{\theta}_{jk} - \frac{\sum_{k=1}^{N} \hat{\theta}_{jk}}{N} \right)^2}, \quad (5) \]

\[ RMSE(\hat{\theta}_j) = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (\hat{\theta}_{jk} - \theta_j)^2}, \quad (6) \]

where \( \theta_j \) is the true ability for person \( j \), \( \hat{\theta}_{jk} \) is the ability estimate for true ability \( j \) in the \( k \)th replication, and \( N \) is the number of replications (\( N = 1000 \) in this study).

Each dependent variable is commonly used to assess parameter recovery in IRT and CAT simulation studies (see Gifford & Swaminathan, 1990; Harwell et al., 1996; Wang & Vispoel, 1998). RMSE is a measure of overall estimation accuracy consisting of two components: bias (or systematic error) and SE (or random sampling error). Given that bias consists of only systematic error, it is valuable for three reasons: (1) comparing group mean parameter estimates when assessing the effect of the studied conditions, (2) to reference ability estimates from different tests to a common scale, and (3) in making classification decisions. This is because bias can result in systematic shifts of group means, individual estimates, and classification cutpoints (Wang & Vispoel, 1998). SE functions as an empirical error variance with smaller estimates indicating greater stability and reliability. Each of these components is essential in assessing parameter recovery in simulation studies (Harwell et al., 1996). SAS (SAS, 2010) was used for all simulations and analyses.

RESULTS

Bias

On average, bias statistics indicated that ability was underestimated for both item pools across each of the nine main speeded conditions (Conditions 1–9) (see Figures 1a and 1b). A direct relationship also existed between the speededness transition point and bias as well as the speededness rate and bias. Results for the realistic item pool became more concerning as the speededness point moved from moderate to extreme and the speededness rate moved from low to high. Expectedly, the least amount of bias (< |−.10|) for the realistic item pool occurred when examinees had moderate speededness transition points at low, medium, and high speededness rates (Conditions 1–3). When the speededness transition point was severe and the rate was low, bias remained less than |−.10| for most ability estimates (Condition 4). When the speededness transition point was extreme and the rate was medium or high (Conditions 8 and 9), bias was between −.10 and
FIGURE 1a
Bias for realistic item pool across 21 $\theta$ s for the nine main experimental conditions (1–9), three end-of-test items conditions (10–12), and the unspeeded control condition.
Bias for ideal item pool across 21 $\theta$'s for the nine main experimental conditions (1–9), three end-of-test items conditions (10–12), and the unspeeded control condition.
SPEEDEDNESS IN COMPUTER ADAPTIVE TESTING

−.30 for most abilities, but was between −.30 and −.60 for high-ability examinees. When examinees were given ability estimates at the moderate, severe, and extreme speededness transition points (i.e., the speeded items were not scored, Conditions 10–12) bias was nearly identical to the unspeeded 30-item CAT. Although differences in bias across the ability spectrum were generally consistent for all conditions, the realistic item pool bias was larger for high-ability examinees as compared to low-ability examinees. Not only was bias larger for high-ability examinees, it became more pronounced as the speededness transition point moved from moderate to extreme and the speediness rate went from low to high. For Conditions 1–9, this can be observed at abilities −2.0 and 2.0, which had bias statistics ranging from −.03 to −.18 and −.08 to −.52, respectively.

Because realistic item pools are not ideal, and thus do not have a flat information function with uniformly high discrimination parameters across ability levels and uniform difficulty parameters across the range of ability levels, the response probabilities are usually less than optimal (i.e., .50 for the 1- and 2-PL and .58 for the 3-PL). In other words, CATs with realistic item pools have more items targeted at average ability levels than at low and high ability levels; thus low-ability examinees will likely receive items that are more difficult and high-ability examinees will likely receive less difficult items. This results in response probabilities that are larger for high-ability examinees and smaller for low-ability examinees. Since response probabilities are already small and closer to zero for low-ability examinees, but larger and inflated for high-ability examinees, a speeded CAT can result in greater drops in response probabilities for high-ability examinees as compared to low-ability examinees. As evidenced for high-ability examinees in Figure 1a, with larger drops in response probabilities comes greater bias. Therefore, as the test becomes more speeded, the bias difference between low- and high-ability examinees becomes larger, with high-ability examinees experiencing the largest amount of bias.

The bias results from the ideal item pool also indicated that speededness not only interacts with examinee ability but also with item pool characteristics (Figure 1b). The ideal item pool bias was generally constant across abilities for the unspeeded condition, and remained constant through the Extreme-Ignore condition (Condition 12). It can also be seen that bias is generally smaller across all conditions, which indicates that speededness does interact with the characteristics of the item pool to affect ability estimation across the conditions differentially.

Standard Error

Regardless of whether a point estimate is biased, it is critical to produce stable point estimates (i.e., small SEs). Figures 2a and 2b report the SEs for the realistic and ideal item pools across ability levels for each condition. For the realistic item pool, the SEs around ability estimates remained less than .35. Unlike the bias
FIGURE 2a
Standard error for realistic item pool across 21 $\theta$s for the nine main experimental conditions (1–9), three end-of-test items conditions (10–12), and the unspeeded control condition.
FIGURE 2b
Standard error for ideal item pool across 21 $\theta$s for the nine main experimental conditions (1–9), three end-of-test items conditions (10–12), and the unspeeded control condition.
results, SEs were larger than the unspeeded 30-item test ability estimates when speeded items were ignored (Conditions 10–12). So although estimating ability at the speededness transition point results in valid ability estimates (i.e., small bias), they are less reliable (i.e., large SEs). The reason for this is that since SE is a function of sampling error, the sampling of items and examinees can have a significant effect on SEs. In other words, depending on the distributional parameters of the item pool, the SE size may look significantly different across the range of abilities. This is apparent when comparing the realistic item pool (Figure 2a) to the ideal item pool (Figure 2b). The SEs for the conditions in the ideal item pool are less than the SEs for the realistic pool, but the ideal item pool also shows fluctuations in SE at different ability levels as the test becomes more speeded. These random fluctuations in SEs are also present in the results across item pools of Wang and Vispoel (1998). This demonstrates that ability can interact with characteristics of the item pool, resulting in random fluctuations in SE.

Consequently, these results imply that when the speededness transition point is moderate, severe, or extreme and examinees do not finish the test and the remaining items are scored incorrect, their ability estimates can become unstable resulting in a larger confidence interval around the point estimate (i.e., their scores become less reliable). Unlike the bias results, SEs were smaller for moderate transition points (Condition 1–3) than for when speeded items were not scored (Conditions 10–12). Collectively, this indicated that while not scoring speeded items resulted in less biased scores, these ability estimates were less stable.

Another interesting observation is that ability estimates were less stable for lower-ability examinees than higher-ability examinees on average. This is actually a characteristic of the 3-PL model. A well-known phenomenon of realistic item pools is they often have less-than-ideal items for examinees with low and high abilities. As previously mentioned, this is because the realistic item pools do not have flat information functions that result from ideal discrimination and difficulty parameters across the ability continuum. This results in less item information for low- and high-ability examinees, and because the SEs are inversely proportional to the amount of item information at different ability levels, inflated standard errors occur at low and high abilities. With the guessing parameter set equal to zero and assuming appropriately targeted difficulty parameters for both low- and high-ability examinees, the SEs will be the same for low- and high-ability examinees. However, when the guessing parameter is greater than zero, there is less information for low-ability examinees. Since less information results in larger SEs, lower-ability examinees will have larger SEs than higher-ability examinees. Hambleton and Swaminathan (1985) have also demonstrated this finding. Thus, as CATs becomes more speeded, characteristics of the item pool and chosen model (i.e., 1-, 2-, or 3-PL) interact with ability resulting in larger and potentially unstable SEs.
Root Mean Square Error

Because RMSE is a function of bias and SE, the results of the realistic and ideal item pools reported in Figures 3a and 3b were similar to those already reported for bias and SE. With the exception of low- and high-ability levels, results from the realistic item pool for moderate and extreme speededness transition points (Conditions 1–6) possessed RMSE values less than .30 for most other ability levels (see Figure 3a). Examinees with extreme speededness transition points (Conditions 7–9) commonly had RMSE values decidedly higher than moderate and severe speededness transition points. Expectedly, the SEs in estimates at the speededness transition points (Conditions 10–12) were reflected in the RMSEs, and thus, the RMSE was slightly higher than the unspeeded condition. In general, for the realistic item pool, there was an increase in RMSE as the simulated conditions became more speeded or examinees were not given sufficient time to even make an educated guess. Consistent with the bias and SE results, RMSEA for the ideal item pool remained relatively constant across ability levels with random fluctuations in RMSEA increasing with increased speededness.

Also similar to the bias results of the realistic item pool, high-ability examinees had the largest RMSE values regardless of the testing condition. Consistent with the SE results, there were extreme fluctuations in RMSE at various ability levels, which as discussed later was purported to be due to measurement error and imperfections in the CAT estimation procedure. In general, the RMSE results concurred with the bias and SE results.

DISCUSSION

The primary purpose of this study was to consider the effects of speededness on ability estimation within a simulated CAT environment and to discern the degree to which speededness affects the estimation of examinee ability. Moreover, this study sought to understand how speededness induced by strict time limits interacts with other operational factors (i.e., item characteristics and item pool characteristics) to affect ability estimation at various ability levels. Results revealed that when examinees are speeded on a significant portion of end-of-test items, their scores suffer from invalid (larger bias) and unreliable (larger SEs) ability estimates. The amount of bias can be relatively small if only a few items are speeded, thus suggesting the 3-PL is robust to small amounts of speededness.

From the perspective of a test developer, the degree of bias and SEs is an artifact of the speededness severity as well as other operational characteristics of the items, the item pool, and the 3-PL model. This study also revealed that although every examinee experienced the same degree and type of speededness, the results were not identical across the ability spectrum. In fact, estimates for examinees with higher ability levels displayed more bias than estimates for examinees with lower
Root mean square error for realistic item pool across 21 θs for the nine main experimental conditions (1–9), three end-of-test items conditions (10–12), and the unspeeded control condition.
FIGURE 3b
Root mean square error for ideal item pool across 21 \( \theta \)s for the nine main experimental conditions (1–9), three end-of-test items conditions (10–12), and the unspeeded control condition.
abilities with the typical item pool. However, high-ability examinees also tended to have more stable ability estimates or lower SEs. As mentioned in the Results section, both of these effects were artifacts of the item pool (Thissen, 2000) and item characteristics (Hambleton & Swaminathan, 1995) interacting with varying degrees of speededness across the range of abilities. Thus, these findings are consistent with those of Bridgeman and Cline (2004) showing that the 3-PL model has difficulty dealing with guessing on end-of-test items. Overall, the CAT appears reasonably robust to test speededness, assuming the degree of speededness is minor, but bias becomes a concern when examinees cannot make educated guesses on several end-of-test items. From an operational standpoint, these results suggest that CAT developers should have relatively few (< 3) speeded items, more items targeted at high-performing examinees, and consider implementing methods/models to deal with end-of-test speededness when examinee’s responses are speeded due to strict time limits.

The studied conditions attempted to simulate time-limit induced speededness as reported in previous studies of speededness with real examinees (e.g., Bridgeman & Cline, 2004) and to help provide possible explanations for some findings by using a controlled simulated CAT. Ideally, the results will provide researchers and test developers with a starting point for understanding conditions that might result in biased ability estimates with speeded CATs. For example, if test developers learn, perhaps by studying examinee response times, that end-of-test items are speeded and thus ability estimates are likely inaccurate and inconsistent, they may elect to either shorten the test or increase the testing time. These results also provide some guidelines for how biased ability estimates may be depending on the degree of speededness, the item pool, and the ability of the examinee.

Since most CATs are designed to be power tests, as opposed to speeded tests, it is important for the majority of examinees to finish the test without guessing or omitting end-of-test items in order to gain a valid assessment of their true ability. Bridgeman and Cline (2004) indicated that approximately half of the examinees guessed on the final six questions of the GRE-A in order to complete the section within the allotted time limit, which resulted in approximately a 70-point reduction in GRE-A scores for those examinees. In extreme cases, they found that GRE-A scores were often 200 points less if high-ability examinees got easy and discriminating items incorrect on speeded tests. This displays the significance of speeded CATs when related to ability estimation, as the test is also likely measuring processing speed.

As indicated, test speededness can result from both examinee and operational factors, so it is important that both test takers and test developers are aware of the implications of speeded CATs. These results are important for groups of examinees who may be more prone to having difficulty completing the test in the allotted time. With the large influx of ELL students at both the k-12 level and the university level,
these findings are particularly troubling. This is especially true given that most mandated statewide academic tests are required to be taken in English (Young, Cho, Ling, Cline, Steinberg, & Stone, 2008), even when English is not the student’s native language. This deficit, which may result in speeded tests, may partially contribute to the large difference found between ELL and non-ELL students in past research of academic testing (Abedi, 2002; Abedi & Lord, 2001; Duran, 2006). For example, even though Asian American k-12 math scores are significantly higher than Caucasian k-12 math scores, Asian American students score lower on reading and writing than their Caucasian counterparts (College Board, 2006). Further, Asian American k-12 students are more likely than African American and Caucasian students to attend schools dominated by ELLs (Ruiz-de-Velasco, Fix, & Clewell, 2000). In the context of a speeded CAT, high-ability Asian American examinees may be more prone to inaccurate estimates of their ability or proficiency. This is especially important considering the results of this study are consistent with those of Bridgeman and Cline (2004) that show high-ability examinees are affected the most by speeded CATs. Although testing accommodations are often provided for ELLs, the role or contribution of speededness is currently unclear. This is especially true given that most states and test developers do not provide consistent accommodations, empirically evaluate the extent that their CATs are speeded, or assess how speededness may affect different groups of examinees. As indicated by Solorzano (2008), who provided a very detailed literature review related to testing ELLs, testing these students is a complex task that requires careful consideration regarding test fairness.

Additional research is needed to consider the effects of speededness within CATs and P&P tests on ability estimation, with focus on at-risk examinees. This is especially true for students with learning disabilities since these students comprise about 90% of the examinees who require test accommodations on the SAT (Camara & Schneider, 2000). This need becomes even more evident as students with disabilities are applying to higher education at record numbers and are often receive testing accommodations (Cahalan, Mandinach, & Camara, 2002). Given the processing speed difficulties are often associated with learning disabilities, one of the most common testing accommodations for these students is additional time to complete the test (Ofiesh, Mather, & Russell, 2005). Cahalan et al. (2002) found that allowing students with learning disabilities extra time results in a significant increase in SAT scores. Camara, Copeland, and Rothschild (1998) also reported that extra time on the SAT I: Reasoning Test seemed to increase the number of items reached and overall test scores. However, the question remains as to the validity of examinee test scores under different testing time conditions. Studying the effects of providing additional time for unique groups of examinees (especially those groups that tend not to respond to items at a higher rate; see Grima & Liang, 1992) would help to clarify the relationships between speededness and ability estimation error with real data.
The empirical results of Bridgeman and Cline (2004) support the conclusions of the current study and emphasize the need for unspeeded tests, especially when factors of the CAT affect ability estimation. These results also speak to the potential difficulty of common IRT models in handling speededness. It should be recognized that this is not a limitation of the IRT model, but the practical necessity of time constraints. Ideally, all time constraints would be removed, nevertheless this is practically and economically infeasible because unlimited time and resources are not realistic. Fortunately, methods have been developed to model test speededness (see Goegebeur et al., 2008; van der Linden, 2009a), but they have yet to be implemented in operational CATs. As found in the current study, an alternative method is to score the test at the point of speededness. This procedure produces reasonably unbiased ability estimates and presents a reasonable alternative for speeded tests. However, for test fairness reasons (e.g., examinees should be administered the same number of items) this practice is unlikely to be adopted (see Mills & Steffen, 2000).

Limitations

A potential limitation of this study was that the numerous models used to correct for speededness were not applied. Although these models exist, the difficulty of applying them in live testing situations remains and, to our knowledge, they have not been used in live CATs. For this reason, the standard IRT model (e.g., 3-PL) was used to replicate real testing conditions. Another potential limitation is that not all possible types of speeded tests were simulated. This study focused on end-of-test speededness given that past research has found this to be most common (see Bridgeman & Cline, 2004), whether it is due to time constraints or fatigue (Wollack, Wells, & Cohen, 2003). In any case, future research could evaluate other test-taking strategies that examinees might use under speeded test conditions. For example, some examinees may hurry on all the items knowing they will not be able to finish the exam in the allotted time.

Test length and item pool characteristics might also be considered a limitation, as not all test lengths and item pools were evaluated. However, recall that although tests vary in length, ability estimates within a CAT environment are often stable and relatively unbiased after about 20 items. Consequently, these results should generalize to different test lengths in terms of how much bias might be expected for different degrees of speededness assuming at least 20 non-speeded items are present. A similar argument could be made for item pool characteristics, as researchers could calculate the test information to determine what examinees will be most influenced by test speededness. For example, tests with fewer easy items that are not discriminating (less test information for low-performing examinees) should produce elevated bias for low-performing examinees, but research is need for different test lengths and item pools.
Another generalizing limitation of the current study is the conceptualization of a true score, as there exist other true score conceptualizations (Lord & Novick, 1968; Sutcliffe, 1965). Similar to past simulation research on test speededness (e.g., Oshima, 1994), this study assumed that one’s true ability is constant over the testing period and independent of test speededness. The limiting factor in this assumption is that it ignores the speed-accuracy tradeoff and fails to acknowledge that one’s true ability may change during the testing period. For instance, it is feasible that as test speededness increases, ability also decreases because processing speed is now incorporated into the ability construct. From one perspective, processing speed is often integrated into general intelligence/ability estimates, but the contribution (based on the factor loadings) to the overall score is often smaller than the other components (see Lichtenberger & Kaufman, 2009). The idea of whether processing speed should be incorporated into ability estimates or modeled separately is certainly controversial and an operational issue, so including the processing speed component rests on whether the test was designed to be a speed, power, or combination of the two. The assumption of a known and unchanging ability (i.e., Platonic true score) is critical here because without it Equations 4 (bias) and 6 (RMSE) cannot be utilized. The reason is that one’s true ability would fluctuate throughout the test, thus making the concept of “bias due to speededness” relatively meaningless. Thus, we assumed a constant true ability in each simulated CAT.

A final limitation was the lack of a consistent smooth curve under some testing conditions when looking at bias, SE, and RMSE, despite the large number of replications. This finding is not unique to this study since other researchers have produced similar results within a CAT environment (e.g., Wang & Vispoel, 1998). For low- and high-ability examinees, the systematic measurement error is due to the interaction between the item pool characteristics and speededness. For other abilities along the continuum, it is simply a result of random measurement error. Although an ideal item pool did reduce the amount of error, aspects of the 3-PL model and speededness continue to result in fluctuations in bias, SE, and RMSE across the ability range. Therefore, fixed-length CATs are unlikely to produce consistent estimates across abilities (Thissen, 2000; Wang & Vispoel, 1998).

Conclusions

This study demonstrated that speededness affects the estimation of ability within CATs and that speededness can result in invalid and unreliable ability estimates. Further, it indicated the extent to which operational factors (i.e., time-limit induced speededness, item characteristics, item pool characteristics) and examinee characteristics (i.e., ability level) moderate the accuracy and consistency of ability estimates. This was evident by higher-ability examinees having greater bias than lower-ability examinees and the realistic item pools resulting in greater bias across
abilities than ideal item pools. In any case, examinees with insufficient time on the last few items should be aware that under most conditions their ability should not be considerably underestimated. At the same time, speededness does produce dependency, which violates an important assumption underlying the test theory. This of course influences estimation validity and blurs the operational definition of ability (i.e., the test score becomes an index of ability and processing speed, as opposed to an index of ability alone).

From an operational standpoint, it can be difficult to detect a point of speededness, but testing administrators can provide reasonable estimates of test speededness by tracking the response times for each item. Test constructors need to be aware that if the response time for examinees accelerates near the end of a test, this may be indicative of a speeded test and biased ability estimates for specific examinees. They then need to consider if operational changes are needed, which may include reconsidering the use of the item response model, reducing test length, and/or increasing testing time. Moreover, CAT designers might consider using a model that incorporates speed and/or additional items for examinees more prone (i.e., \(-1 > \theta > 1\)) to biased and unstable estimates.

Although CATs that rely on common IRT models are not meant to be utilized under conditions of speededness, these tests are certainly in use. This study provides valuable insight into how the commonly employed 3-PL model performs when a CAT becomes speeded with different item pools and ability level groups. Again, these results showed how test speededness does not have a constant effect across the ability spectrum, but instead the magnitude of bias is test- and examinee-dependent. It is our hope that practitioners, test developers, and examinees will consider this information when developing and interpreting the results of CATs. As mathematician Yuri Burago stated about his former student and famed mathematician Grigory Perelman who proved the Poincaré conjecture, which was considered by many mathematicians the “holy grail” of mathematical problems:

‘There are a lot of students of high ability who speak before thinking,’ Burago said. ‘Grisha was different. He thought deeply. His answers were always correct. He always checked very, very carefully.’ Burago added, ‘He was not fast. Speed means nothing. Math doesn’t depend on speed. It is about deep.’ (Nasar & Gruber, 2006, p. 3)

NOTES

1. Ability is defined as an unobservable trait and is used interchangeably throughout with latent trait, proficiency, and aptitude.
2. It should be noted that in CATs item selection is also commonly based on content balance (see Swanson & Stocking, 1993; van der Linden, 1998) and item exposure constraints (see Stocking & Lewis, 1998; Symson & Hetter, 1985). Content balance and item exposure constraints were not included in the current study because they are not part of the IRT model, the methods vary greatly (Wang & Vispoel, 1998), and when slightly less than optimal items
are being administered there will generally be an adverse effect on the accuracy of ability estimation within a CAT environment (van der Linden, 1998). Thus, the primary purpose of this study was to evaluate the model to provide a standard for further simulation research and to demonstrate the effect of speededness on ability estimation. Scoring penalties were also not implemented as they also play no formal role in the IRT model and are implemented in various ways.

REFERENCES


