Graduate
Management
Admission
Council®

Termination Criteria in Computerized Adaptive Tests: Variable-Length CATs Are Not Biased

Ben Babcock and David J. Weiss

University of Minnesota

Presented at the Realities of CAT Paper Session, June 2, 2009

Abstract

This simulation study examined a large number of computerized adaptive testing (CAT) termination rules using the item response theory framework. Results showed that longer CATs yielded more accurate trait estimation, but there were diminishing returns with a very large number of items. The authors suggest that a minimum number of items should always be used to ensure the stability of measurement when using CATs. Standard error termination performed quite well in terms of both a small number of items administered and high accuracy of trait estimation if the standard error used was low enough. Fixed-length CATs did not perform better than their variable-length termination counterparts; previous findings stating that variable-length CATs are biased were the result of a statistical artifact. The authors discuss the conditions that led to this artifact occurring.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Babcock, B. & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**Ben Babcock, The American Registry of Radiologic Technologists,
1255 Northland Drive, St. Paul, MN 55120, U.S.A. ben.babcock@arrt.org**

Termination Criteria in Computerized Adaptive Tests: Variable-Length CATs Are Not Biased

Computerized adaptive tests (CATs) are becoming increasingly popular in a variety of domains (Fliege et al., 2005; Simms & Clark, 2005; Triantafillou, Georgiadou, & Economides, 2007). Because of computer availability and advances in item response theory (IRT; Weiss & Yoes, 1991), adaptive testing is now used in a myriad of settings. CATs tailor the test to each individual examinee in order to obtain accurate measurement across the entire latent trait continuum. CATs also are advantageous over non-adaptive tests because CATs can administer fewer items to examinees while maintaining the same quality of measurement as non-adaptive tests (Weiss, 1982). A termination or stopping rule is what determines the length of the CAT. Although some studies have tested a few possibilities for termination (e.g., Dodd, Koch, & De Ayala, 1993; Gialluca & Weiss, 1979; Wang & Wang, 2001), no single study has thoroughly compared a large number of CAT termination criteria using multiple item banks. This study examined numerous termination rules with four different item banks to determine which termination rules led to the best CAT estimation of a latent trait.

Modern CAT uses IRT as a statistical framework. IRT models the probability of responses to a test item based on a person's underlying latent trait/ability, or θ. While there are a wide variety of IRT models, this study used the unidimensional 3-parameter logistic model (3PL) for dichotomously scored items. The mathematical form of this model is

$$P(x_{ip}=1\,|\,\theta_p,a_i,b_i,c_i)=c_i+(1-c_i)\frac{\exp[Da_i(\theta_p-b_i)]}{1+\exp[Da_i(\theta_p-b_i)]}, \tag{1}$$

where i is an item index, p is a person index, x is a person's response to an item (1 for a keyed response, 0 for a non-keyed response), D is the multiplicative constant 1.702, a is the item discrimination parameter, b is the item difficulty parameter, and c is the lower asymptote (pseudo-guessing) parameter (Birnbaum, 1968; Weiss & Yoes, 1991). A plot of the probability of a keyed response on θ has a familiar ogive shape and is called an item response function (IRF). The probability of responding in the keyed direction depends on the person taking the item and item characteristics. Increasing b will decrease the probability of a person responding in the keyed direction. Increasing a will cause the probability of responding in the keyed direction to change more quickly when θ values are near b, thus increasing the slope of the IRF for θ values near b. Increasing the lower asymptote c parameter increases the probability of a keyed response for low values of θ but only slightly increases the probability of response for high ranges of θ. IRT users can estimate the parameters of this model with marginal maximum likelihood using computer programs such as XCALIBRE (Assessment Systems Corporation, 1996) and BILOG (Mislevy & Bock, 1991). For a summary on the basics of IRT, see Embretson and Reise (2000) or De Ayala (2009).

CATs require six main components: (1) a response model, (2) an item bank of IRT-calibrated items, (3) an entry rule, (4) an item selection rule, (5) a method for scoring θ, and (6) a termination rule (Weiss & Kingsbury, 1984). IRT is a good statistical model for CAT, because there are a variety of options available in the IRT framework to fulfill these requirements. There has been substantial research on IRT item parameter estimation (e.g. Harwell, Stone, Hsu, & Kirisci, 1996), CAT entry rules (e.g. Gialluca & Weiss, 1979), item selection (e.g. Hau & Chang, 2001), and θ scoring methods (e.g. Wang & Vispoel, 1998). Much less work has been conducted in the area of CAT termination.

The two most popular termination rules thus far in the literature are fixed-length termination and standard error termination (Weiss, 1982; Gushta, 2003). Fixed-length CATs simply terminate when an examinee has taken a pre-specified number of items. The simplicity of this termination rule and its similarity to paper-and-pencil tests has made it popular in applied settings. Some researchers even argue that variable-length CATs are more biased than fixed-length CATs (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001). The present study tested whether the result that fixed-length CATs are less biased than variable-length CATs holds under conditions designed to control for alternative explanations.

Another popular termination rule for CATs is standard error (SE) termination (Weiss & Kingsbury, 1984). According to this termination rule, the examinee continues to take test items until the examinee's θ estimate reaches a specified level of precision, as indicated by its SE, resulting in equiprecise measurement across examinees. The SE of a θ estimate when using maximum likelihood scoring is calculated from the inverse of the square root of the second derivative of the negative log likelihood function at the θ estimate with respect to θ, or

$$SEM(\hat{\theta}) = \frac{1}{\sqrt{-\partial^2 \log L / \partial\theta^2}}, \tag{2}$$

where $\hat{\theta}$ is the current θ estimate and log L is the log likelihood function (Samejima, 1977). The log likelihood function is defined as

$$\log L(x_{1p}, x_{2p}, ..., x_{np}) = \sum_{i=1}^{n} x_{ip} \log[P_i(\theta)] + (1 - x_{ip}) \log[Q_i(\theta)], \tag{3}$$

where i is an item index, n is the number of items to which a person has responded, P is the probability of a person responding in the keyed direction (Equation 1), and Q is $1-P$ (Embretson & Reise, 2000, Ch. 7). When a person has a great deal of psychometric information in their responses (i.e., has answered questions with difficulties near true θ), the log likelihood function will be highly curved and steep at the θ estimate (the maximum of the likelihood). This curvature causes a decrease the SE around the θ estimate. Researchers have found that SE termination performs well in terms of accurate estimation of θ (Dodd, Koch, & De Ayala, 1993; Dodd, Koch, & De Ayala, 1989; Revuelta, & Ponsoda, 1998; Wang & Wang, 2001) if the information in the item bank allows a specific value to be attained at all levels of θ.

A third termination rule that has been studied in a limited fashion is the minimum information termination rule. This rule states that a CAT should end when there are no items remaining in the test bank that can provide more than a specified minimal amount of psychometric information at the current θ estimate (Gialluca & Weiss, 1979; Maurelli & Weiss, 1981). Fisher item information I is calculated by

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \tag{4}$$

where i is an item index and P' is the first derivative of the IRF (Samejima, 1977). Most item selection criteria in CAT involve choosing an item with high information at the current θ

estimate. The theory behind minimum information termination is that if there are no remaining items that will yield information about the examinee, the test should terminate for the sake of efficiency. While some research has demonstrated that this termination rule provides a great deal of efficiency over conventional tests (Brown & Weiss, 1977), other research has shown this method to be inferior to other termination methods. Dodd, Koch, and De Ayala, (1989, 1993) found that minimum information performed slightly worse than fixed SE in terms of correlation with true θ. The values of minimum information that they used, however, were relatively high (.45 to .5, versus .01 and .05 used by Brown & Weiss). This could have led to premature termination of CATs and inferior measurement.

A fourth termination rule that has received almost no research attention is the θ convergence, or change in θ, criterion. Because of the addition of new psychometric information, a person's θ estimate changes after answering each item in a CAT. Changes in θ are large at the beginning of a CAT and become smaller as the CAT tailors the test to the person and converges on a θ estimate (Weiss & Kingsbury, 1984). The convergence of a θ estimate could provide a good stopping rule for a CAT. Hart, Cook, Mioduski, Teal, and Crane (2006) and Hart, Mioduski, and Stratford (2005) investigated a hybrid CAT termination rule that combined SE and θ convergence. The researchers concluded that the convergence of θ yielded good θ estimates for a CAT.

Purpose

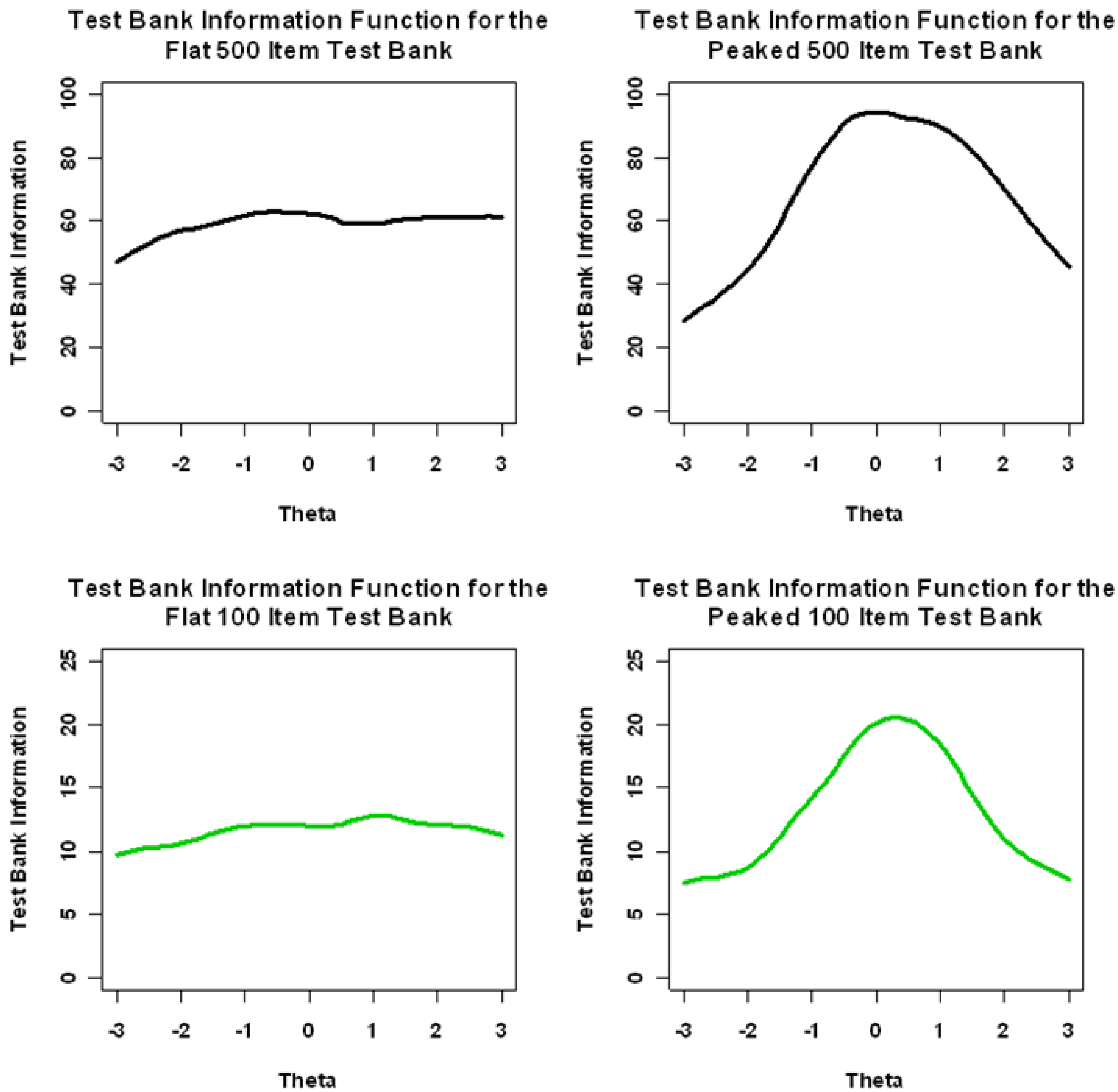
This study used simulation to investigate various termination rules for CATs. Studies in the past have investigated only a few termination rules, typically fixed-length and either SE or minimum information termination. This study used several conditions of four basic termination rules (SE, minimum information, change in θ, and fixed length) and two combinations of SE and minimum information termination. These combinations were used in order to take advantage of both high measurement precision where it was possible and terminating quickly when high precision was probably not possible. The four different item banks in this study were used to examine the conditions in which a given termination rule was superior or inferior in its ability to recover true θ compared to other termination conditions.

Method

Item Banks

There were four item banks in this study: (1) a flat information bank with 500 items, (2) a peaked information bank with 500 items, (3) a flat information bank with 100 items, and (4) a peaked information bank with 100 items. Figure 1 contains the information functions for these four banks. These item banks have several notable features. First, total information on the negative end of θ was somewhat lower than at the higher (positive) end of θ. This occurs because the *c* parameter (lower asymptote) in the 3PL decreases the slope of an IRF in the low ranges of the function; decreases in slope lead to lower item information. Second, the 500-item banks have a great deal of information across the entire range of θ. These banks should be able to satisfy all of the SE criteria used as stopping rules for most of the θ range. An information value of 20.67, for example, corresponds to a model-predicted SE of 0.22 [$1/(20.67)^{1/2} = 0.22$]. Both of the 500-item banks had model-predicted information values greater than 21 for the entire θ continuum. Finally, the smaller item banks had generally low levels of information. These item banks could not satisfy some of the lower SE termination criteria.

Figure 1. Bank Information Functions for the Four Item Banks



The *a* parameters for all four banks were generated to approximate a log normal distribution with a log mean of –0.1 and a log standard deviation of 0.3. This translated into mean discrimination values of about *a* = 0.94 with a standard deviation of about 0.28. The *b* parameters for the two flat banks were generated from a uniform distribution with a minimum of –4 and a maximum of 4. The *b* parameters for the peaked banks came from a mixed distribution; 300 and 50 items in the large and small banks, respectively, came from a uniform distribution with a minimum of –4 and a maximum of 4. The remainder of the *b* parameters for these two banks came from a standard normal distribution. The *c* parameters for all item banks came from a mixture of uniform distributions, one with a minimum of 0.1 and a maximum of 0.225, and the other with a minimum of 0.075 and a maximum of .35. These item parameter distributions were based on estimated parameter distributions in the achievement testing domain (Chen & Ankenman, 2004; Wang, Hanson, & Lau, 1999).

Simulees

In order to evaluate the performance of the stopping rules along the θ continuum, this study used 1,000 simulees at 13 evenly spaced points on θ from −3 to 3. Thus, 1,000 simulated people with true θ values of −3 took each CAT, 1,000 simulees with true θ values of −2.5 took the CAT, and so on. Each level of θ had 1,000 response vectors.

Data Generation and CAT Simulation

Based on the item parameters and θ, each θ level had a model-predicted probability of responding in the keyed direction to an item. The response simulation compared this probability of response with a random uniform deviate between 0 and 1 in order to determine how this simulee would respond to the item. If the probability of response in the keyed direction was greater than the random number, the person responded in the keyed direction. If the probability of response in the keyed direction was less than the random number, the person responded in the non-keyed direction. POSTSIM3, a computer program for post-hoc CAT simulation, simulated the CATs after the generation of the four full data sets (Assessment Systems Corporation, 2008). POSTSIM3 allows the user to conduct a simulated CAT based on specified item parameters and a person's full set of item responses. Estimated CAT θ values for each termination condition from POSTSIM3 were then compared to the true θs.

CAT Conditions

The following are the conditions that this research used:

1. *Starting rule*. All examinees started with an initial θ of 0. This made the start of the CAT equal for all conditions and eliminated chance starting variation from affecting the results of the study.

2. *Item selection rule*. This study used maximum Fisher information to select items for the CAT. POSTSIM3 administered the item with the highest information at the current estimate of θ for each simulee. This rule maximizes the efficiency of the CATs by reducing the SE of θ [SE(θ)] as quickly as possible.

3. *θ estimation*. All conditions used maximum likelihood (ML) estimation for θ. If the simulee did not have a mixed response vector (i.e., had all keyed or all non-keyed responses), ML scoring does not yield a finite maximum for the likelihood function. The CAT algorithm increased (all keyed responses) or decreased (all non-keyed responses) θ by a fixed step size of θ = 0.5 for the next θ estimate when the simulee had a non-mixed response vector. This rule was used to obtain a mixed response vector relatively quickly, and ML scoring was used to estimate θ after there was a mixed response vector. This research used ML estimation instead of Bayesian methods because previous research has demonstrated that Bayesian θ estimation methods produce biased θ estimates when true θ is extreme (Guyer, 2008; Stocking, 1987; Wang & Vispoel, 1998; Weiss & McBride, 1984).

4. *Termination rules*. This was the focus of the current study. For each item bank, the same simulated response data were run 14 times, with each run using a different termination rule. The following were the termination conditions:

1) SE(θ) was below 0.385 (analogous to a reliability of 0.85[1]) with a maximum of 100 items.

2) SE(θ) was below 0.315 (analogous to a reliability of 0.90) with a maximum of 100 items.

3) SE(θ) was below 0.220 (analogous to a reliability of 0.95) with a maximum of 100 items.

4) All items not yet administered at the current θ estimate had less than 0.2 information with a maximum of 100 items.

5) All items not yet administered at the current θ estimate had less than 0.1 information with a maximum of 100 items.

6) All items not yet administered at the current θ estimate had less than 0.01 information with a maximum of 100 items.

7) Either when the SE(θ) was below 0.315 or when all items not yet administered had less than 0.1 information at the current θ estimate with a maximum of 100 items, whichever occurred first.

8) Either when the SE(θ) was below 0.220 or when all items not yet administered had less than 0.01 information at the current θ estimate with a maximum of 100 items.

9) Absolute change in θ estimate was less than 0.05 with a minimum of 11 items and a maximum of 100 items. The minimum number of items was to ensure that the CAT did not terminate prematurely.

10) Absolute change in θ estimate was less than 0.02 with a minimum of 11 items and a maximum of 100 items.

11) Fixed-length CAT with the number of items equal to the mean number of items required in Condition 2.

12) Fixed-length CAT with the number of items equal to the mean number of items required in Condition 5.

13) Fixed-length CAT with the number of items equal to the mean number of items required in Condition 7.

14) Fixed-length CAT with the number of items equal to the mean number of items required in Condition 9.

The fixed-length conditions were used here in order to compare fixed-length CATs with variable-length CATs using a comparable number of items.

---

[1] In classical test theory, the standard error of measurement (*SEM*) is approximated with the equation SEM = $s_{\text{obs}}\,(1-\rho_{xx})^{½}$, where s_{obs} is the standard deviation of the observed scores and ρ_{xx} is the reliability. Assuming that the standard deviation of θ is 1, specifying a reliability of .85 for ρ_{xx} gives a standard error of .385.

Dependent Variables

Five dependent measures were used to evaluate the performance of the CATs.

1. *Length of the CAT*. This was simply the number of items the CAT required to terminate. Although this dependent variable was not particularly interesting for the fixed-length conditions, it was important for the variable-length conditions. This dependent variable was a measure of the efficiency of the CAT.

2. *Bias*. This statistic was the signed mean difference between the CAT estimated θ and true θ. Bias was calculated for each of the 13 θ values and combined across all values of θ. It was calculated by

$$\text{bias} = \frac{\sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)}{N} \tag{5}$$

   where N is the number of people at a θ point, j is a person index, θ_j is a person's true θ, an $\hat{\theta}_j$ is a person's CAT estimated θ.

3. *Root mean squared error* (*RMSE*). This statistic was a measure of absolute difference between the CAT estimated and true θ. It was calculated by

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)^2}{N}} \tag{6}$$

4. *Pearson correlation between estimated and true* θ. This is the familiar correlation between the CAT estimated θ values and the true θ values. The equation for the Pearson correlation r is

$$r = \frac{\text{cov}(\hat{\theta}, \theta)}{\text{SD}(\hat{\theta})\text{SD}(\theta)}. \tag{7}$$

5. *Kendall's Tau rank order correlation between estimated and true* θ. This measure indicated how similarly each of the conditions ranked the simulees. Kendall's Tau is sensitive to changes in the ordering of θ. The equation for Tau is

$$\tau = \frac{4P}{N(N-1)} - 1 \tag{8}$$

   where P is the sum of greater rankings in variable 2 after a given observation when the observations are ordered by the first variable. Kendall's Tau was used in addition to the familiar Pearson correlation because Tau is more sensitive to small changes in the ordering of θ than the Pearson correlation.

6. *Coverage*. This variable was the number of times that true θ fell within ±2 SEs of the CAT θ estimate. Coverage is a measure of the accuracy of the SEs. The expected coverage should be about 950 out of 1,000 (i.e. 95% confidence interval), so large deviations from this number indicate that the condition produced inaccurate SEs.

Results

Bank 1

Table 1 contains the major results from Bank 1. Several of the conditions produced identical results. Conditions 5, 6, and 12 all gave the maximum number of items to examinees. These conditions, thus, all administered the same tests. Conditions 11 and 13 were also equivalent, because these CAT conditions were the same fixed length with the same items administered. Conditions 2 and 7 and Conditions 3 and 8 were identical because Conditions 7 and 8 always reached SE termination before reaching minimum information termination with this item bank. These results were not surprising when the size and shape of the item bank is taken into consideration. Thus, Conditions 6, 7, 8, 12, and 13 were eliminated from this bank's analyses because of equivalence with other conditions.

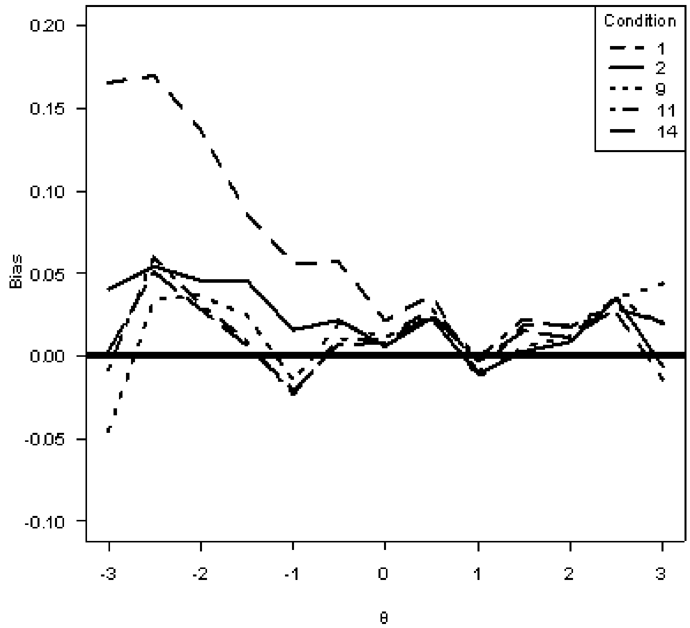
The SE below 0.385 (Condition 1), the SE below 0.315 (Condition 2), and the change in θ less than 0.05 (Condition 9) administered the fewest items among the variable-length termination criteria conditions. The minimum information criteria conditions administered close to or exactly the maximum allowed 100 items, because the item bank contained a great number of informative items across θ. The middle ranges in θ used slightly fewer items than the extreme θs for the variable-length conditions. This occurred because the starting value for each person was at $\theta = 0$. Because the starting value was close to the person's true θ in the middle ranges, the test terminated a few items more quickly. Overall, the numbers of items administered across different values of θ were relatively close to the mean value.

Table 1. Summary Results for Item Bank 1

	Condition								
Statistic	1	2	3	4	5	9	10	11	14
Mean Length	9.27	14.03	33.53	99.99	100	15.58	28.32	14.0	16.0
SD Length	2.84	3.59	5.44	0.15	0.00	3.37	6.59	NA	NA
Mean Bias	0.06	0.02	0.01	0.00	0.00	0.01	0.01	0.02	0.01
Mean RMSE	0.46	0.34	0.22	0.16	0.16	0.36	0.29	0.38	0.34
Kendall's τ	0.88	0.91	0.94	0.96	0.96	0.92	0.94	0.91	0.92
Pearson r	0.97	0.98	0.99	0.99	0.99	0.98	0.99	0.98	0.98
Mean Coverage	0.92	0.94	0.95	0.95	0.95	0.94	0.95	0.94	0.94

The mean bias for all of the conditions was very close to 0, with the exception of Condition 1. This positive bias occurred because this termination method did not estimate low values of θ very well. Figure 2 is a graph of the bias conditional on θ for Conditions 1, 2, 9, 11, and 14. It is clear that, except for Condition 1, the methods had very little conditional bias across θ. The results for Condition 1 were likely due to an interaction between administering a relatively low number of items and the c parameter in the low ranges of θ. Future research should investigate this interaction. It is worthy to note that Conditions 2 and 9 were variable-length termination conditions, and Conditions 11 and 14 were fixed-length conditions whose length corresponded to the mean length of Conditions 2 and 9, respectively. There were not any large differences in the bias between the variable- and fixed-length conditions.

Figure 2. Conditional Bias for Selected Conditions in Item Bank 1



The RMSEs yielded some interesting results concerning the length of a CAT and the accuracy of θ estimation. RMSE was strongly related to the number of items administered; the scatterplot in Figure 3 illustrates this relationship. Mean RMSE decreased quickly between 0 and 35 items. The point on the far right shows that the RMSE decreased more slowly once the test was over 40 items long. One interesting item of note was that Condition 2, one of the SE termination criteria, had a slightly lower RMSE than Condition 11, its fixed-length counterpart. The conditional RMSE values in Figure 4 indicate that the Condition 2 SE termination criterion had lower RMSE in the low regions of θ than the fixed-length Condition 11. Although Condition 1 was clearly the worst for RMSE, the fixed-length and variable-length termination conditions with comparable lengths all had similar RMSE values across most of the θ continuum.

Kendall's τ and the Pearson correlation between true θ and the CAT estimated θs followed generally the same pattern. Conditions that administered more items had higher correlations. The differences were more pronounced with Kendall's τ. These results are not surprising considering the results linking RMSE with test length.

The mean coverage from Table 1 for all conditions was relatively close to the nominal rate of 0.95. This indicates that the SEs of θ at the end of a CAT were relatively accurate, on average. Coverage was somewhat poor for Condition 1 in the low ranges of θ. Overall, however, the confidence intervals functioned near the nominal rate for Bank 1.

Figure 3. Scatterplot of Mean RMSE on Mean Test Length for Item Bank 1

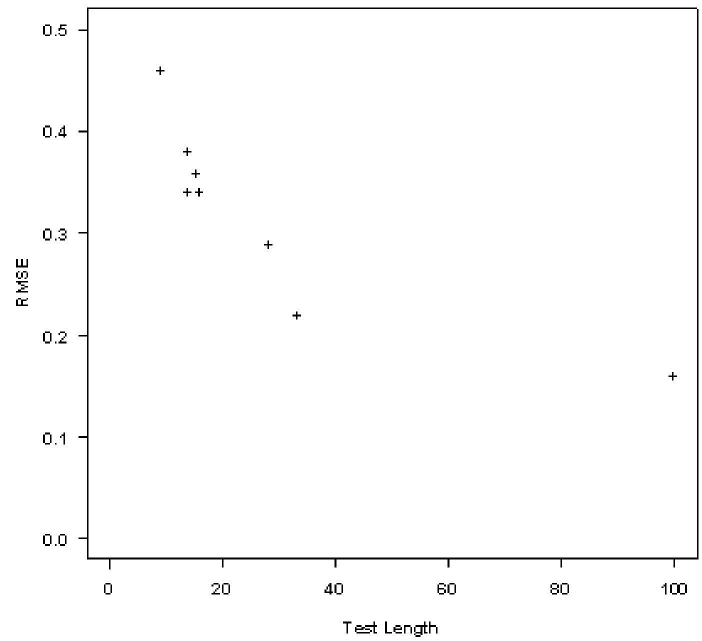
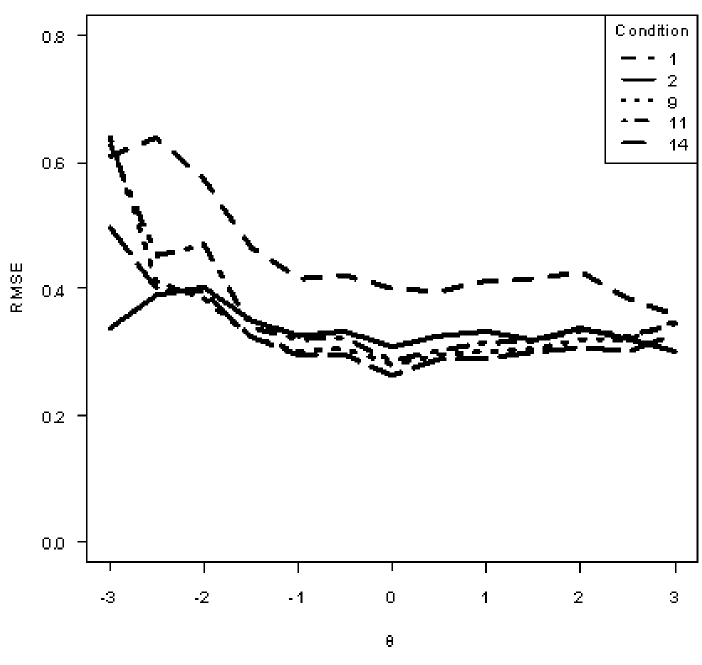


Figure 4. Conditional RMSE for Selected Conditions in Item Bank 1



It is of particular note that the variable-length CATs performed equivalently or slightly better than their fixed-length counterparts in terms of bias and RMSE. With a large flat information item bank, it appears that variable-length CATs are not biased as previous studies have claimed (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001).

Bank 2

Table 2 contains the results from Bank 2 combined across θ levels. Several of the conditions produced the same (or essentially the same) results. Conditions 11 and 13 were equivalent because these CAT conditions were fixed length with the same number of items administered. Conditions 2 and 7, and 3 and 8 were virtually identical because, with the exception of only a very few examinees, Conditions 7 and 8 reached SE termination before reaching minimum information termination. Conditions 7, 8, and 13 were eliminated from this analysis.

Table 2. Summary Results for Item Bank 2

	Condition										
Statistic	1	2	3	4	5	6	9	10	11	12	14
Mean Length	9.14	13.89	34.83	90.91	98.62	100	15.05	28.48	14.0	99.0	15.00
SD Length	2.81	3.90	10.17	14.92	4.62	0.00	3.20	6.34	NA	NA	NA
Mean Bias	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02
Mean RMSE	0.48	0.34	0.22	0.16	0.16	0.16	0.40	0.33	0.40	0.16	0.37
Kendall's τ	0.88	0.91	0.94	0.96	0.96	0.96	0.92	0.94	0.91	0.96	0.92
Pearson's r	0.97	0.98	0.99	1.00	1.00	1.00	0.98	0.98	0.98	1.00	0.98
Mean Coverage	0.92	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.94

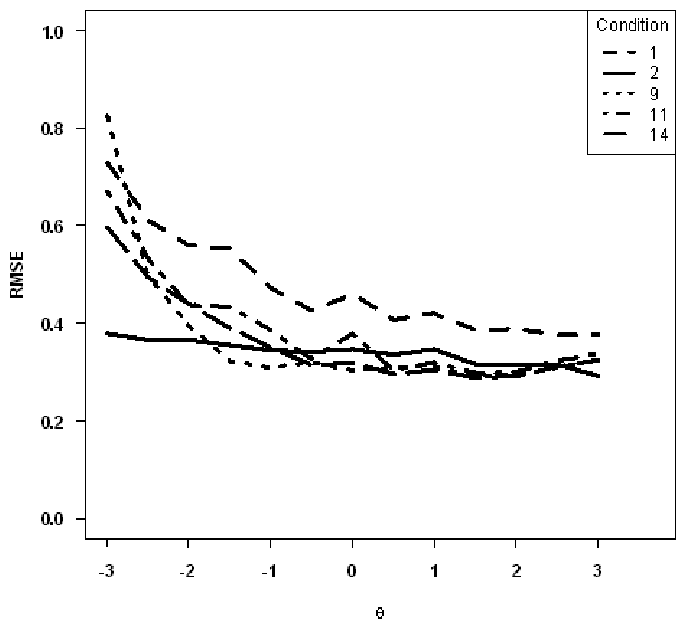
The Condition 1 (SE < 0.385), Condition 2 (SE < 0.35), and Condition 9 (change in $\theta < 0.05$) termination conditions once again administered the fewest items. The minimum information criterion conditions administered more than 90 items on average, because the item bank contained a great number of informative items across θ, even though the bank was somewhat peaked. The change in θ conditions (9 and 10), also gave relatively low numbers of items. The minimum information conditions used fewer items at the extremes of θ, because there were fewer items in these ranges that gave psychometric information. Condition 3, the lowest of the SE termination criteria, administered a larger number of items in the extreme ranges of θ, accounting for the increase in the number of items administered. This occurred because there were somewhat fewer items in Bank 2 with high information at the extremes of θ.

All of the conditions were relatively unbiased when combined across θ, except for Condition 1. This positive bias occurred because this termination method did not estimate low values of θ very well, a trend similar to the results from Bank 1. For every condition except Condition 1, the conditional bias across θ was always close to 0.

The mean RMSE from Table 2 yielded results that were similar to the RMSE results from Bank 1. The RMSE was quite strongly related to the number of items administered. The conditions that administered the largest number of items all had very low RMSE across the entire range of θ. The conditions that administered fewer items had larger RMSE values for low θs and relatively lower RMSE values for high θs. Figure 5 is a plot of the conditional RMSE values for Conditions 1, 2, 9, 11, and 14. Condition 2 had a much lower RMSE in the low ranges of θ than its fixed-length counterpart (Condition 11). The SE termination criterion administered a few more items to people in the low ranges of θ who were not measured well. Because this condition administered more items to the simulees who needed to take more items, the variable-length Condition 2 performed better than the fixed-length Condition 11. Condition 9, however, had a slightly higher RMSE than its fixed-length counterpart (Condition 14). It appears that the change in θ conditions (Conditions 9 and 10) performed slightly worse than CATs of comparable length that were terminated by SE or fixed length.

The results for Kendall's τ and Pearson's r closely matched the results from Bank 1. The conditions that used more items performed slightly better. Based on the correlations from Conditions 3 and 9, terminating with an average of about 30 items gave a high Kendall's τ, but adding more items did not seem to increase the correlation very much. The results for coverage matched closely the results from Bank 1 in that coverage was very close to the nominal rate for every condition except Condition 1. The coverage was below 0.90 for θs below −1 in Condition 1. When an insufficient number of items were administered, the final SE estimates for low values of θ were too small.

Figure 5. Conditional RMSE for Selected Conditions in Item Bank 2



Bank 3

Table 3 contains the results combined across θ levels for Bank 3. The SE termination criteria (1, 2, and 3) administered more items in this item bank, because there was not a large number of highly discriminating items at every point on θ. The minimum information criteria conditions (4, 5, and 6) administered fewer items than in Banks 1 and 2 for the same reason. The number of items administered for Conditions 9 and 10, the change in θ conditions, were not greatly affected by the change in the item bank.

All of the conditions yielded θ estimates that had a very small positive bias throughout the θ continuum. All conditions also had conditional RMSE values that were relatively close to the mean RMSE values. Conditions 9, 10 (change in θ) and 14 (fixed length) had slightly higher conditional RMSE at $\theta = -3$. Giving a very large number of items from this small item bank did not result in large decreases in RMSE in the same way as did giving more items from the large bank. This is because the large number of additional items that the CATs administered did not provide much psychometric information about the examinees. The correlational measures showed the same pattern as previous item banks: all correlations were high, and the highest correlations were for the conditions with the most items. Finally, the confidence interval coverage was very close to the nominal rate of 95%.

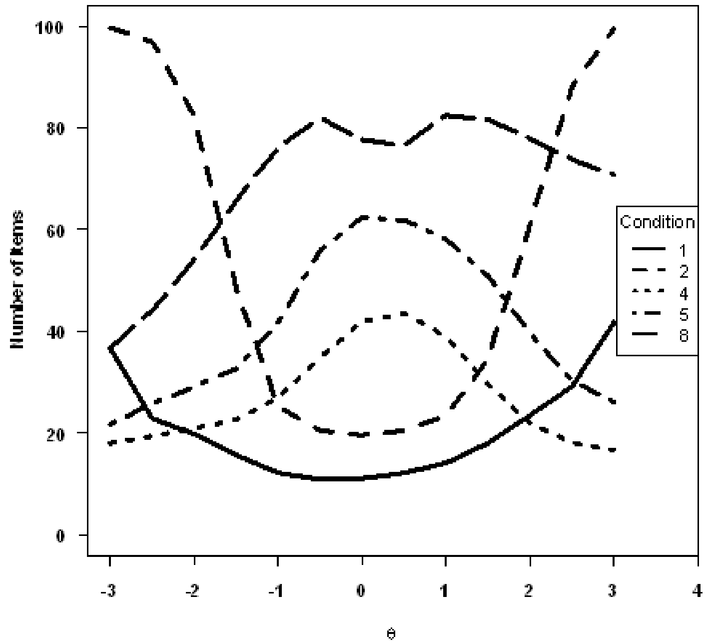
Variable-length CATs in this item bank performed roughly equivalently to their fixed-length counterparts. The mean bias, RMSE, and coverage rates were virtually identical between the variable-length CATs and the comparable fixed-length CATS. These results further demonstrate that variable-length CATs are not more biased than fixed-length CATs.

Bank 4

Table 4 contains the results combined across θ levels for Bank 4. The conditions that generally administered the fewest items for Bank 3 also administered the fewest items for Bank 4. The change in θ conditions administered the fewest items overall for this item bank. Figure 6 contains a plot of the conditional mean number of items administered across θ for selected conditions. Conditions with a great deal of overlap or relatively low conditional variability were not shown in this figure. As seen in Figure 6, the number of items administered varied widely within some conditions. These differences occurred because of the greater amount of information available in the middle of the θ distribution for this item bank. Variable-length conditions requiring a SE cutoff (Conditions 1 and 2) administered more items at the extremes of θ because the item bank often did not have enough items to fulfill the termination criterion. The minimum information criterion conditions (Conditions 4 and 5) administered fewer items in the extremes of θ because of a lack of informative items in this region.

There was a very slight positive bias across all conditions. The trends for conditional bias were quite similar across all conditions, varying slightly between just below 0 to 0.1. The conditional RMSE values were slightly lower in the middle of θ for all conditions, and the conditions that administered more items had slightly lower RMSE overall. The correlational measures showed the same pattern as previous item banks: all correlations were high, and the highest correlations were for the conditions with the most items. Confidence interval coverage across θ was fairly close to the nominal rate of 95% for all conditions. Similar to previous item banks, the fixed- and variable-length CATs that had comparable mean number of items performed similarly.

Figure 6. Conditional Number of Items Administered for Selected Conditions for Item Bank 4



Conclusions

This study examined a wide variety of CAT termination criteria. A few general trends were observed when examining the results of all of the item banks together. First, CATs that administered more items yielded better θ estimates. The gains in accuracy were largest when adding items to short exams. Adding items to exams that were already long (e.g., 50 items) did not produce sizable gains in the accuracy of estimating θ. Second, CATs that were too short (e.g., fewer than 15 items) did not give good θ estimates in low ranges of θ. Termination criteria that terminated very quickly were not stable. The large conditional biases and conditional RMSE values for the Condition 1 CATs demonstrated this. A CAT should always administer a minimum number of items, such as 15 to 20, before terminating, and CAT administrators using standard error termination should use a standard error that is equal to or smaller than 0.315 for accurate measurement of θ in terms of bias and RMSE. Third, the variable termination criteria that performed the best when taking test length and accuracy into consideration were the conditions that used a standard error below 0.315 as part of the termination rule. This termination rule also estimated low θ values more accurately than its fixed-length counterpart. Change in θ termination, a relatively new termination rule, performed slightly worse than the standard error conditions. Finally, using minimum information termination alone administered too many items for large item banks.

One clear conclusion was that, contrary to claims in the literature (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001), variable-length CATs were not biased nor did they perform worse than fixed-length CATs. Variable-length CATs either performed equally to or slightly better than their fixed-length counterparts when average test lengths were comparable. Previous results were a statistical artifact due to the number of items administered and the scoring methods used to

Table 3. Summary Results for Item Bank 3

Statistic	Condition													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Mean Length	16.02	42.09	100	24.77	36.76	61.10	28.93	61.10	15.44	23.87	42.0	37.0	29.0	15.0
SD Length	6.03	26.83	0.00	2.95	4.16	8.88	5.42	8.88	2.96	4.09	---	---	---	---
Mean Bias	0.03	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02
Mean RMSE	0.41	0.32	0.30	0.35	0.32	0.30	0.33	0.30	0.44	0.39	0.31	0.32	0.33	0.44
Kendall's τ	0.89	0.92	0.92	0.91	0.92	0.92	0.91	0.92	0.89	0.91	0.92	0.92	0.91	0.89
Pearson r	0.98	0.98	0.99	0.98	0.99	0.99	0.98	0.99	0.97	0.98	0.99	0.99	0.98	0.97
Mean Coverage	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.94

Table 4. Summary Results for Item Bank 4

Statistic	Condition													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Mean Length	20.66	55.38	97.81	27.21	41.33	70.21	25.82	69.22	14.88	24.20	55.0	41.0	26.0	15.0
SD Length	17.26	36.20	8.19	10.06	15.17	15.66	6.58	15.39	2.88	5.55	---	---	---	---
Mean Bias	0.03	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.03	0.02	0.01	0.01	0.02	0.04
Mean RMSE	0.42	0.34	0.31	0.37	0.33	0.31	0.36	0.31	0.45	0.39	0.32	0.33	0.36	0.44
Kendall's τ	0.89	0.91	0.93	0.91	0.92	0.93	0.91	0.93	0.89	0.91	0.92	0.92	0.91	0.89
Pearson's r	0.97	0.98	0.99	0.98	0.98	0.99	0.98	0.99	0.97	0.98	0.99	0.98	0.98	0.97
Mean Coverage	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.94

estimate θ. First, fixed-length CATs in previous simulation studies have generally been much longer than variable-length CATs. This means that fixed-length CAT conditions in these studies utilize more psychometric information (Revuelta & Ponsoda, 1998), thus giving fixed-length tests an unfair advantage in θ estimation. Second, many of these studies used Bayesian methods for θ estimation. Numerous studies have documented that Bayesian scoring methods produce biased results in the extremes of θ, particularly when tests are short (Guyer, 2008, Stocking, 1987; Wang & Vispoel, 1998). Previous results claiming that variable-length CATs are biased are due to Bayesian θ estimation techniques combined with variable–length CATs that terminated with too few items. The trend found in this study was that, no matter how a CAT is terminated, using two few items has a negative effect. This was especially true for the large item banks, where Condition 1 terminated in a mean of less than 10 items. The θ estimates from this condition were highly biased in the low ranges of θ, had high RMSEs, and had confidence intervals with the lowest coverage of any condition.

The best solution to the CAT termination issue might be to use one or more variable termination criteria in combination with a minimum number of items constraint. Based on this research, 15 to 20 items appears to be a reasonable *minimum* number of items for variable-length CAT termination, depending on the precision needs of the test user. A variable termination rule would supplement the minimum item termination rule by administering more items to people who are still not measured well. This would ensure stability of the CAT results and could fulfill some other desirable properties of θ, such as standard errors below a certain benchmark value and the efficiency that is a common desire for CAT users. Future research should explore the minimum number of items more thoroughly with various items banks.

The change in θ criterion, a relatively new termination criterion, performed just slightly worse than other termination criteria with a comparable number of items. This termination method may be a viable supplement to standard error termination when an item bank does not permit a given standard error to be reached in extreme ranges of θ, which can occur in peaked-information item banks. Minimum item information as a supplemental rule to standard error termination, however, is also a viable alternative to simply having one termination rule for a small item bank.

This research does have limitations. The most notable limitation is that this study did not control for item exposure or content balancing. It is probable that techniques controlling for item exposure and/or content balancing – both practical implementation issues in numerous CAT applications – would increase the number of items required for a CAT to give a desired level of measurement precision.

CAT is a good way to measure accurately and simultaneously increase test efficiency. This research demonstrated that a wide variety of termination criteria work well if a minimally sufficient number of items is used. CATs can dramatically reduce the number of items required for accurate measurement over non-adaptive methods, and more research will continue to demonstrate the effectiveness of this kind of measurement tool.

References

Assessment Systems Corporation (2008). POSTSIM3: Post-hoc simulation of computerized adaptive testing [Computer software]. Minneapolis, MN: Assessment Systems Corporation.

Assessment Systems Corporation. (1996). *User's manual for the XCALIBRE marginal maximum-likelihood estimation program.* St. Paul, MN: Assessment Systems Corporation.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*, pp. 397–479. Reading, MA: Addison-Wesley.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Chang, S.-W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40,* 71-103.

Chen, Y.-Y., & Ankenman, R. D. (2004) Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149-174.

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

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-143.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53*, 61-77.

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

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277–2291.

Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subset branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. Available from CAT Central (http://www.psych.umn.edu/psylabs/catcentral/).

Gushta, M. M. (2003). *Standard-setting issues in computerized-adaptive testing.* Paper Prepared for Presentation at the Annual Conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia, May 30th, 2003.

Guyer, R. D. (2008). *Effect of early misfit in computerized adaptive testing on the recovery of theta.* Ph.D. Dissertation, University of Minnesota.

Hart, D. L., Cook, K. F., Mioduski, J. E., Teal, C. R., Crane, P. K. (2006). Simulated computerized adaptive test for patients with shoulder impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology, 59*, 290–298.

Hart, D. L., Mioduski, J. E., & Stratford, P. W. (2005). Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. *Journal of Clinical Epidemiology, 58*, 629–638.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement, 20,* 101-125.

Hau, K.-T., & Chang, H.-H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266.

Maurelli, V., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries.* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory. Available from CAT Central (http://www.psych.umn.edu/psylabs/catcentral/).

Mislevy, R. J., & Bock, R. D. (1991). *BILOG user's guide*. Chicago: Scientific Software.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, *1*, 233-247.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17*, 28–43.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36*, 263-277.

Triantafillou, E., Georgiadou, E. & Economides, A.A. (2007). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education, 49*. Found online at http://www.conta.uom.gr/conta/publications/PDF/The%20design%20and%20 evaluation%20of%20a%20computerized%20adaptive%20test%20on%20mobile%20devices.pdf

Wang, T., Hanson, B. A., & Lau, C.-M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23,* 263-278.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25,* 317-331.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 109-135.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

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

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8,* 273-285.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R.K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 69-95). Boston: Kluwer Academic.

Yi, Q., Wang, T., & Ban, J.-C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38*, 267-292.