COMPARISON OF FOUR STOPPING RULES IN COMPUTERIZED ADAPTIVE TESTING AND EXAMINATION OF THEIR APPLICATION TO ON-THE-FLY MULTISTAGE TESTING

BY

CHEN TIAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Master's Committee:

Professor Hua-Hua Chang, Chair
Professor Carolyn J. Anderson
Associate Professor Jinming Zhang

ABSTRACT

Computerized adaptive testing (CAT) is a powerful and efficient approach in educational testing for both estimating ability and classifying examinees into groups. When the purpose is to classify students as either proficient or not proficient in ability, an accurate estimate is not necessary, and the test can stop whenever a satisfactory decision can be made. Therefore, the stopping rule is a critical element in variable-length adaptive testing. In this study, the efficiency of four stopping rules was compared in variable-length CAT designs (vl-CAT): ability confidence interval (ACI), sequential probability ratio test (SPRT), generalized likelihood ratio (GLR), and the truncation rule. In addition, their application to the newly-developed adaptive testing design, on-the-fly multistage testing (OMST), was also examined and compared with vl-CAT.

Two simulation studies were conducted. In study 1, since the fourth stopping rule cannot be executed independently, ACI, SPRT, and GLR were combined with the truncation rule, which resulted in 6 CAT and 6 OMST designs in total. With the classification accuracy (CA) controlled at the same level, the test length of 12 variable-length designs was examined. All test designs in study 1 have a length between 10 and 30. In study 2, the lower and upper bound of the test length was extended to 30 and 100, and only ACI- and GLR- CAT designs were conducted to provide a more general comparison of these two stopping rules. In both studies, the ability was estimated by maximum likelihood estimation or expected a posterior. The next item(s) is (are) selected with the maximum priority index at the current ability estimate. 1000 theta values were simulated from a standard normal distribution, and 30 replications were conducted under each design.

The results show that OMST produced similar results to CAT. Regarding the efficiency of four stopping rules, the truncated versions of ACI, SPRT, and GLR produced shorter test lengths than their corresponding counterparts. Among ACI, SPRT, and GLR, SPRT yielded the longest test length with the highest estimation accuracy. The results of GLR and ACI designs are similar, but ACI is more efficient for examinees whose ability is far from the cutoff point, and GLR is more efficient for examinees whose ability is near the cutoff point.

It can be concluded that the stopping rules designed for CAT also function for OMST in a similar way. When the item selection method is estimate-based rather than cutscore-based, SPRT performed less efficiently than ACI and GLR. The efficiency of GLR and ACI is comparable, and each has its own strengths. The truncation rule is useful because it prevents examinees from taking unnecessary items. These results imply the good statistical properties of variable-length OMST, which facilitates its future application. The research also provides a direct comparison between different stopping rules, giving practitioners more information about the above-mentioned applicable situations of different rules. The studies also develop a new simple truncation rule and indicate its feasibility in a future adaptive testing context.

ACKNOWLEDGEMENTS

I want to thank my advisor, Dr. Hua-Hua Chang. Your valuable advice, rich experience and support provided me much insight and encouraged me to overcome difficulties. I also want to thank my committee members, Dr. Jinming Zhang and Dr. Carolyn Anderson, who offered me generous help and guidance through my master's study.

Lastly, I want to thank my family and friends for their long-time supports and unconditional love.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Computer-based testing is gaining its popularity in educational contexts. Nowadays, there are three major types of computer-administered tests: computer-based tests (CBT), computerized adaptive tests (CAT), and multistage tests (MST) (Zheng and Chang, 2015). Compared with CBT, which has no adaptation algorithm, adaptive testing, like CAT and MST, is more extensively researched and widely applied in measurement tasks. Adaptive testing is an assessment method that sequentially selects items from the item pool to match the performance of the examinee during the administration of a test (Wang, Lin, Chang, and Douglas, 2016). Therefore, a more tailored test will be given, and this individualized test improves the accuracy and efficiency of tests dramatically with the same test length as a fixed-form test.

While some tests are designed to obtain accurate ability estimates, the purpose of mastery testing is to classify examinees into mutually exclusive groups. Mastery testing is used in educational or certification contexts to decide whether or not an individual has achieved a sufficient command of a given subject matter (Lewis, Sheehan, DeVore, & Swanson, 1991). Mastery testing is useful when deciding the degree of a student's proficiency and identifying individuals who need further training in a specific subject (Nitko & Hsu, 1974). Given the usefulness of both mastery and adaptive testing technologies, a fusion of these two occurred to reduce test length while maintaining the accurate decisions necessary for correct diagnoses (Kingsbury and Weiss, 1983). There are many stopping rules or termination criteria designed for adaptive testing with the purpose of reducing the classification error while using as few items as necessary (Thompson &Prometric, 2007).

In this project, the purposes are to compare four stopping rules designed for adaptive classification testing and to explore if these rules can also function well in a newly-developed

adaptive testing design: on-the-fly multistage testing (OMST). Chapter 2 is a literature review, which introduces two kinds of adaptive testing and important elements in adaptive mastery testing, including the item selection methods and stopping rules. This chapter also contains a summary and analysis of previous studies that investigated similar research questions. Chapter 3 and 4 describe the methods and results of two studies in detail. Chapter 5 presents the conclusions, implications, and limitations of the studies.

CHAPTER 2: LITERATURE REVIEW

Adaptive Classification Testing

Computerized Adaptive Testing (CAT) was first proposed by Lord (1971) and developed further by Owen (1975) and Weiss (1976). CAT was designed to sequentially tailor the item difficulty to an examinee's ability so that the examinee is always challenged. Specifically, during the test, every item is selected based on the examinee's responses to previous items and a set of constraints, such as content coverage, answer-key distribution, and item exposure (e.g., Zheng & Chang, 2015; Wang et al., 2016). The major advantage of CAT is that it can estimate the latent trait with fewer items than traditional linear tests (e.g., Weiss, 1976; Chang, 2015). Schnipke and Reese (1999) found that the CAT design had the lower mean square error and bias compared to the 25-item paper-pencil-based design. Multistage testing (MST) is another kind of adaptive testing, but instead of choosing individual items on-the-fly, it uses a bundle of items as the building blocks for a test and adapts between blocks (e.g., Zheng and Chang, 2014; Chang, 2015), and tests are pre-assembled before the actual administration.

The framework of on-the-fly assembled multistage tests (OMST) which is a compromise between CAT and the traditional MST, was proposed by Zheng and Chang (2015). The principal idea is to use the well-developed item selection algorithms of CAT to assemble a bundle of items on-the-fly to match each examinee's ability. Like MST, OMST is administered in stages and only adapts between stages (Figure 2.1). However, compared with MST where the modules in every stage are all pre-assembled and panel designs are fixed, the stages in OMST are assembled during the process of testing, so the difficulty levels of modules in OMST in later stages can be any point in the whole ability scale with more individualized adaptation. After the implementation of the first moderately difficult stage, the examinee's ability is provisionally

estimated, and the second stage will be assembled based on the estimate. After the completion of the second stage, the ability estimate is updated, and a new stage is assembled on-the-fly to match the new ability estimate. OMST can have as many stages as desired, and the process continues until the test is terminated.

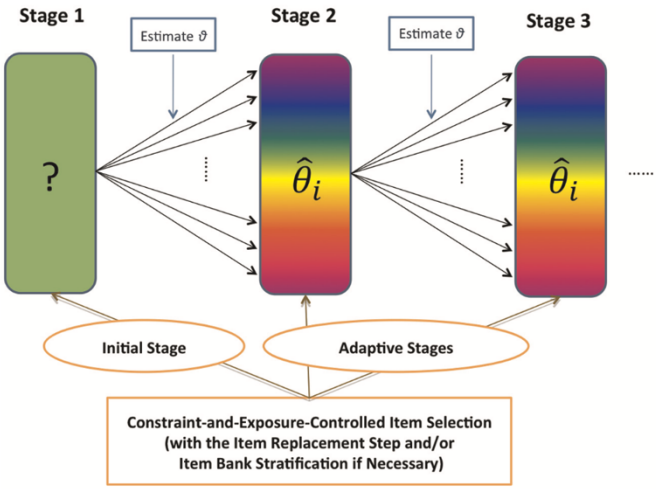


Figure 2.1. The structure of OMST. Reprinted from "On-the-fly assembled multistage adaptive testing," by Y. Zheng, and H. H. Chang, 2015, *Applied Psychological Measurement, 39*(2), 104-118. Copyright 2015 by SAGE Publications.

OMST combines the advantages of both CAT and MST and offsets their limitations. Compared with CAT, OMST can avoid the over-/underestimation problem of CAT, due to its robust initial estimation based on bundles of items. Also, examinees are allowed to skip and review items freely within a stage, which relieves the stress of test takers. Those two advantages are quite important from test takers' perspective because the robust initial estimation facilitates the test fairness, and a lower level of anxiety can improve the measurement accuracy. Compared with MST, simulation studies have shown that OMST has a higher measurement efficiency, especially at the two polar ends of the ability scale (Zheng & Chang, 2015). Besides the statistical advantages, OMST also eliminates the difficulty of assembling parallel panels by

implementing computer algorithms automatically and avoiding considering complex constraints (e.g., Zheng & Chang, 2014; Wang et al., 2016). Also, due to the decreased probability that two examinees share the same panel, OMST improves test security (Zheng & Chang, 2014; 2015).

Item Selection Methods

An important element in CAT is the item selection method that is used to select items during test administration. As Wang and her colleagues (2016) summarized, the most commonly used item selection method is the Maximum Fisher Information approach. However, this algorithm, which solely relies on Fisher's information, requires a large sample (number of items) to be reliable. Specifically, at the beginning of a test, when the sample size is small, if capable examinees happen to answer the first few items incorrectly, or less capable examinees happen to guess the first few items correctly, the estimation will be heavily biased and cannot easily adjust back within a short test (Chang & Ying, 2002; 2008).

With the consideration of practical purposes and non-statistical constraints, there are many other modifications or substitutes to the original method (e.g., Chang & Ying, 1999; Cheng & Chang, 2009; Luecht, 1998; Swanson & Stocking, 1993; van der Linden & Reese, 1998). According to Cheng and Chang (2009), there are two categories of the item selection methods proposed to manage non-statistical constraints: mathematical programming approaches and heuristic approaches. The methods in the first category are effective in constraint management but are computationally intensive when the number of constraints is large, such as the network-flow programming method (Armstrong, Jones, & Kunce, 1998) and the shadow-test method (van der Linden, 2000; van der Linden & Reese, 1998). Methods in the second category, such as the weighted deviation modeling method (Swanson & Stocking, 1993), are simpler and more

computationally efficient, despite the fact that they treat constraints as desirable properties and do not guarantee that all will be strictly met.

The maximum priority index (MPI) method is a heuristic method that can accommodate various non-statistical constraints simultaneously (Cheng and Chang, 2009). MPI can be considered as a variant of the maximum information method, and the next item is selected if this item maximizes the priority index (PI): $PI_j = I_j \prod_{k=1}^{K} (w_k f_k)^{c_{jk}}$, where I_j is the Fisher's information of item *j* at the relevant θ, w_k is the weight associated with *k*th constraint, and f_k is the scaled quota left of constraint *k*. c_{jk} is the element in the *j*th row and *k*th column of the constraint relevancy matrix $\boldsymbol{C}$. $c_{jk} = 1$ means constraint *k* is relevant to item *j*, and $c_{jk} = 0$ means constraint *k* is irrelevant to item *j*. For the scaled quota left f_k, if the *k*th constraint is a content constraint, then $f_k = (X_k - x_k)/X_k$, where X_k is the total number of items from a certain content area, and x_k indicates so far, x_k such items have been selected. If the *k*th constraint is an exposure rate constraint, then $f_k = (r - (n/N))/r$, where r is the pre-set maximum value of an acceptable exposure rate, n is the number of the examinees who have seen item *j*, and N is the number of the examinees who have taken the test.

Stopping Rules

Another important element of variable-length adaptive classification tests is the stopping rule (also called termination criterion), which is an algorithm determining whether the examinee can be classified. If a satisfactory classification decision can be made, the test is terminated; otherwise, one or more items will be administered. One stopping rule is the IRT-based ability confidence interval (ACI) approach that formulates the classification purpose as a statistical estimation problem; one is the sequential probability ratio test (SPRT) approach that treats the classification purpose as a hypothesis testing problem (Eggen & Straetmans, 2000); and one is

the generalized likelihood ratio (GLR) approach, which is the same as SPRT but allows the likelihoods of two competing hypotheses to vary. Also, a truncation component can be designed into the variable-length adaptive testing to recognize a situation when the completion of the remaining questions would be unlikely to alter the current decision.

ACI is a method that makes a confidence interval around the estimate of ability using the conditional standard error of measurement (CSEM), and once the current confidence interval (CI) does not hold the cut score, the test will stop. Thompson (2007) quantified this CI as the following expression: $\widehat{\theta_j} - z_\alpha(CSEM) \leq \theta_j \leq \widehat{\theta_j} + z_\alpha(CSEM)$. The CSEM can be either theoretical or observed. The theoretical SEM is model predicted and calculated using the test information function at the relevant θ regardless of the response pattern: $SEM = 1/\sqrt{TI(\theta)}$ (Embretson & Reise, 2000). The observed SEM is calculated based on the second derivative of the likelihood function: $SEM = 1/\sqrt{-E(\partial^2 L/\partial\theta_j^2)}$ (Baker & Kim, 2004). Figure 2.2 is a graphic depiction of the ACI method: when the 95% CI around the ability estimate does not include the cutoff point, 0, the test will stop even though the test has not reached its maximum length, 25. In real life test situations, the National Council Licensure Examination for Registered/Practical Nurse employs the ACI approach: this institution calculates the 95% CI after administering 60 items and sets a maximum test length for candidates with abilities very close to the passing standard (National Council of State Boards of Nursing, 2012).

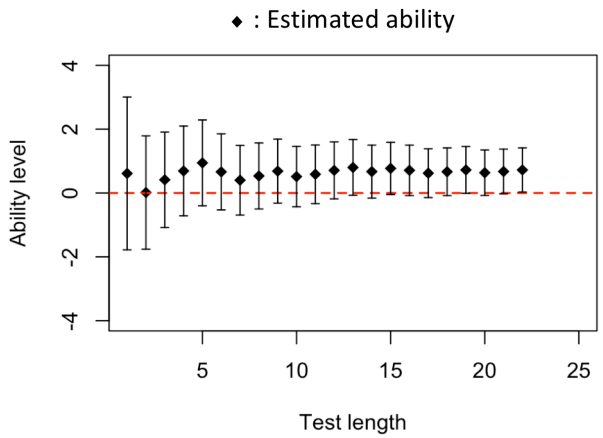


Figure 2.2. An ACI design stops when the 95% CI does not include the cutoff point

The SPRT compares the ratio of the likelihoods of two competing hypotheses (Reckase, 1983; Wald, 1947). Once the likelihood ratio (LR) is greater than the upper decision point A or smaller than the lower decision point B, the test will stop. To calculate LR, test developers should define an indifference region first. The indifference region is a region of ability that the test designer think is acceptable with going either way (pass or fail), and it is defined by two endpoints, θ_1 and θ_2, on either side of the cutoff score. In practice, an arbitrary small constant δ usually is specified to determine this region by adding and subtracting this constant from the cut score (e.g., Eggen, 1999; Eggen & Straetmans, 2000; Thompson & Prometric, 2007). θ_1 is the lowest acceptable θ level that the test designer allows an examinee to pass, and θ_2 is the highest acceptable θ level that the test designer allows an examinee to pass. Given a response pattern, the LR is expressed as the ratio of the likelihoods at θ_1 and θ_2: $LR = L(\theta = \theta_2)/L(\theta = \theta_1) = \prod_{i=1}^{n} P_i(X = 1|\theta = \theta_2)^X P_i(X = 0|\theta = \theta_2)^{1-X} / \prod_{i=1}^{n} P_i(X = 1|\theta = \theta_1)^X P_i(X = 0|\theta = \theta_1)^{1-X}$, where n is the number of administered items, P_i is the possibility that an examinee answer item i correctly, and X is the response, 0 or 1, for item i. To calculate the decision points A and B, test developers should specify the nominal error rates, α and β. According to Wald (1947), $A = (1 - \beta)/\alpha$, and $B = \beta/(1 - \alpha)$.

GLR is specified and calculated with the same methods as the fixed-point SPRT, but instead of using two fixed endpoints to calculate LR, GLR uses the highest points beyond the endpoints (Thompson, 2011). If the maximum of the likelihood function lies out of the indifference region, this maximum will be used in the likelihood ratio for that side. Figure 2.3 shows the different endpoints used in SPRT and GLR when the maximum likelihood estimation (MLE) method is used. For Figure 2.3, the likelihood of a given response at $\hat{\theta}$ rather than at θ_2 will be used to calculate the generalized likelihood ratio. In other words, if the ability estimate is not in between of θ_1 and θ_2, compared with SPRT, GLR is more likely to be larger than A or smaller than B. When MLE is used to estimate ability, GLR can be expressed as such:

$$GLR = \begin{cases} \dfrac{L(\theta = \theta_2)}{L(\theta = \theta_1)}, if\ \theta_1 \leq \hat{\theta} \leq \theta_2 \\ \dfrac{L(\theta = \hat{\theta})}{L(\theta = \theta_1)}, if\ \hat{\theta} \geq \theta_2 \\ \dfrac{L(\theta = \theta_2)}{L(\theta = \hat{\theta})}, if\ \hat{\theta} \leq \theta_1 \end{cases}$$

Another stopping rule is to include a truncation component. A test with a truncation rule will stop when a satisfactory decision can be made or administering more items will not alter the current classification decision. This idea was summarized by Thompson and Prometric (2007): "A component may be designed into the variable-length computerized classification tests to recognize this situation and determine when the remaining items in the bank do not have enough information to make a decision either way, even if the examinee answered all of the remaining items correctly (or incorrectly)" (p. 9).

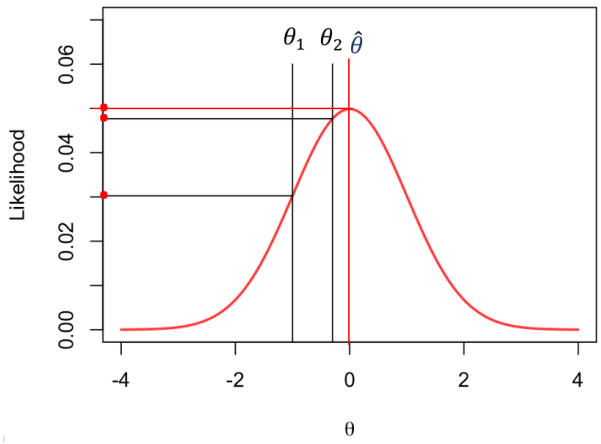


Figure 2.3. Different endpoints used in SPRT and GLR when MLE is used

Previous Studies

As for which stopping rule is the most efficient, a consensus has not been reached. Regarding ACI and SPRT, the appropriateness of each rule and the maximization of its benefits depend on several other factors, such as the psychometric model and item selection algorithm (Thompson & Prometric, 2007). While SPRT works most effectively with the cutscore-based item selection method because it increases the P_2-P_1 difference (Lin & Spray, 2000), ACI works most effectively with estimate-based item selection method because it decreases the standard error of measurement (Thompson & Prometric, 2007). As for GLR and SPRT, Thompson (2011) found that GLR yielded shorter test lengths especially when δ is small because a wider indifference region forces the GLR and SPRT to utilize the same calculations. As for the test designs with a truncation rule, Finkelman (2003) proposed truncated forms of SPRT. In Finkelman's study, a test is stopped early if completion of the remaining questions would be unlikely to alter the classification decision of the examinee. The results showed that the truncation rule offered substantially shorter test lengths with only a slight decrease in classification accuracy. However, this method is designed for the fixed-form test, which means all items are selected by maximizing the information at the cut score. It assumes the administer

knows which question is going to be presented, and in what order, so the method developed in that study is not applicable to the adaptive testing which uses an estimate-based item selection method.

The statistical properties of OMST, MST, and CAT were compared in previous studies, but few research was found investigating the variable-length OMST. Studies have shown that OMST and CAT designs have comparable results. In Zheng and Chang's simulation study, both CAT and OMST designs used the same item selection, exposure rate controlling, and item bank stratification method. The results show that the RMSEs and biases of OMST and CAT designs were comparable across all the experimental conditions, and they were smaller than that of the traditional MST designs. This study shows the similarities between CAT and OMST and the promising statistical properties of OMST over MST. Wang et al. (2016) proposed a hybrid design which has a flexible transition point connecting OMST and CAT. In this hybrid design, the test would move from an OMST step to a CAT step when the length of confidence interval for the current ability estimate is smaller or equal to a predetermined length. If the predetermined criterion is not satisfied even when the maximum test length is reached, the test will end up as an OMST design. The results show that hybrid designs provided comparable or better estimation accuracy and efficiency than standard CAT designs. However, though the hybrid design has a flexible transition point connecting OMST and CAT, the properties of variable-length OMST were not directly examined and compared with variable-length CAT.

In my study, the efficiency of different stopping rules, ACI, SPRT, GLR, and their combination with the truncation component, are examined and directly compared in variable-length CAT designs and OMST designs. The first purpose is to compare the efficiency of these four stopping rules, and the second purpose is to explore whether these rules designed for CAT

can also function well and yield similar results in OMST designs. Two simulation studies were conducted: the first study provided answers to the two research questions when test length was set in between 10 and 30, and the second study provides a further comparison of the ACI and GLR methods when the test length was extended.

CHAPTER 3: STUDY 1

General Specifications

The cutoff point classifying examinees into the pass and fail groups, θ_{cut}, was set to 0. Although four stopping rules would be compared, the fourth stopping rule which introduces a truncation component is only executable when combined with one of the other three. Therefore, two versions of the ACI, SPRT, and GLR were produced. The standard version was administered without a truncation rule, and the truncated version has the truncation rule. In total, 12 designs were produced (6 under CAT and 6 under OMST), and 30 replications were conducted under each design to reduce the random sampling error.

The total test length for all 12 designs was set between 10 to 30. For OMST designs, each stage contains five items; thus examinees take at least 2 stages and at most 5 stages. Every examinee is required to take at least 10 items before being classified. When the test length reaches 30, the test will stop regardless of whether a satisfactory classification decision can be made according to the current stopping rule; the final decision will be made based on the final point estimation. Both CAT and OMST designs have a medium-difficult initial stage containing five items from each content category. Those five items were also randomly selected from 5 difficulty levels. The cut points for the five difficulty levels are -0.84, -0.254, 0.254, and 0.84, which resulted in a similar number of items falling into each level. The adaptive testing begins with the 6th item, adapting between each item in CAT and between every 5 items (1 stage) in OMST.

The item pool contains five content categories and has 360 items, with each item belonging to only one category and an equal number of items in each category. All items are dichotomous and are simulated according to the three-parameter logistic IRT model. The a

parameter was drawn from $N(1, 0.2)$, the b parameter was drawn from the standard normal distribution, and the c parameter came from a uniform distribution with the minimum value of 0 and the maximum value of 0.2.

The Maximum Priority Index method (MPI) proposed by Cheng and Chang (2009) was used in both CAT designs and OMST designs to select the next item(s). In this study, the item information used in the MPI method was calculated based on an examinee's current ability estimate rather than the cutoff point. Content constraints and exposure rate were also controlled: a maximum number of 6 items was set for each of the five content categories, and the maximum exposure rate was set to 0.2. All w_k were set to 1 for five content constraints and one exposure rate constraint. Examinees' abilities, θs, were estimated by Maximum Likelihood Estimation (MLE) when MLE is available and by Expected A Posteriori (EAP) when a constant response pattern was produced.

Stopping Rule Specifications

For the standard versions of ACI, SPRT, and GLR, the test will stop either when the test length reaches its maximum or when a satisfactory classification decision can be made. For the truncated versions of ACI, SPRT, and GLR, if a satisfactory classification decision was still not available when the examinee had already answered 25 items, the test would temporarily stop and examine the necessity of administering the remaining 5 items. The necessity of continuing the test was determined by supposing two extreme situations: answering all the remaining 5 items correctly or incorrectly. Under these two circumstances, two extreme final point estimates could be gained. If those two extreme estimates are both greater/smaller than the cutoff point, which means answering the remaining 5 items correctly/incorrectly will not change the current classification decision, the test will stop at the 25th item. Otherwise, the test will continue until a

satisfactory decision can be made or the test reaches its maximum length. The reason why we use two extreme final estimates is that it is reasonable to assume that the point estimate at the 25th item falls in between those two extreme estimates

Some parameters are needed to define what constitutes a satisfactory decision. In ACI designs, if the 95% confidence interval (CI) does not include θ_{cut}, and the test length is in between 10 and 30, the test will stop. The results of the pilot study show that a standard ACI design with a 95% CI yielded a mean classification accuracy of 88.40% ($sd = 0.0107$) over 30 CAT replications and 88.55% ($sd = 0.0090$) over 30 OMST replications. To match this CA level, two δ values, 0.2 and 0.3, were applied to standard SPRT and GLR designs. Results show that 0.3 can produce CAs at the same level and is more appropriate, which is consistent with Thompson's previous study (2011). The value $\delta = 0.3$ was used in both GLR and SPRT designs.

Evaluation Criteria

To compare four stopping rules, the mean test lengths of 1000 examinees and test length distribution would be examined, with the classification accuracy of all 12 designs controlled at the same level. To explore whether these stopping rules designed for CAT can also be applied to OMST and produce similar results, the test length distribution of 6 CAT designs and 6 OMST designs will be compared. Other indexes such as the correlation coefficient between estimated and true abilities will also be calculated to provide more information.

Data Source

1000 ability values were generated from a standard normal distribution. The same set of true θ values was used across all test designs and replications. Response data was generated using Monte Carlo simulation. For examinee *i*, the probability of answering item *j* correctly was calculated from the formula:

$$P_j(\theta_i) = c_j + \frac{1-c_j}{1+\exp\left[-a_j(\theta_i - b_j)\right]}.$$

After calculation, the probability was compared with a random number r ($0.0 < r < 1.0$). If the probability is larger than r, examinee i's simulated response on item j was set as incorrect (denoted by 0). Otherwise, the response was set as correct (denoted by 1). Responses were generated as each item was administered to each examinee in the simulation.

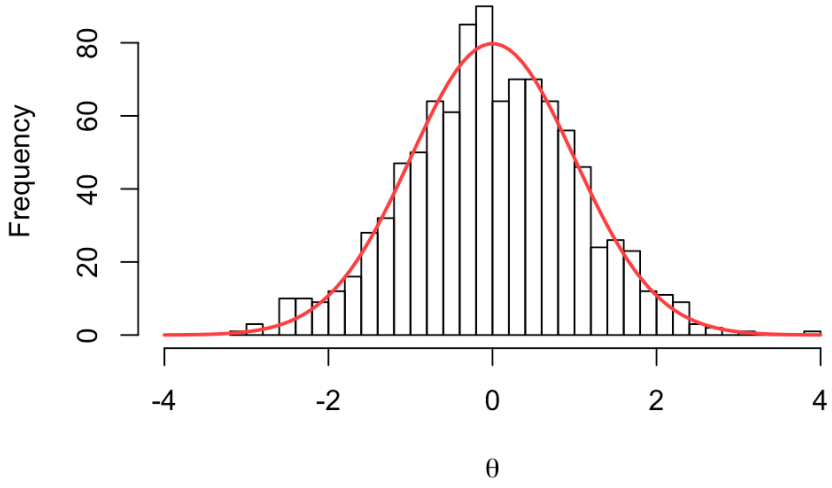


Figure 3.1. The histogram of 1000 simulated ability parameters from $N(0,1)$

Results

In this study, the classification accuracy was controlled at the same level across all 12 designs. Under each design, the CAs from 30 replications were averaged and summarized in Table 3.1 and Figure 3.2. The results show that for all designs, the averaged CA ranges from 88.40% to 88.87%. A one-way ANOVA was conducted to examine whether the 12 averaged CAs are different, and there was no evidence showing that they are different ($F(11,348) = 0.64, p = 0.79$).

Table 3.1

The mean and standard deviation of CAs from 30 replications under each of the 12 designs

	CAT Mean (SD)		OMST Mean (SD)	
	without truncation	with truncation	without truncation	with truncation
ACI	88.40% (0.0107)	88.47% (0.0080)	88.55% (0.0090)	88.76% (0.0087)
GLR	88.50% (0.0094)	88.53% (0.0075)	88.48% (0.0081)	88.51% (0.0104)
SPRT	88.66% (0.0086)	88.87% (0.0094)	88.67% (0.0097)	88.62% (0.0110)

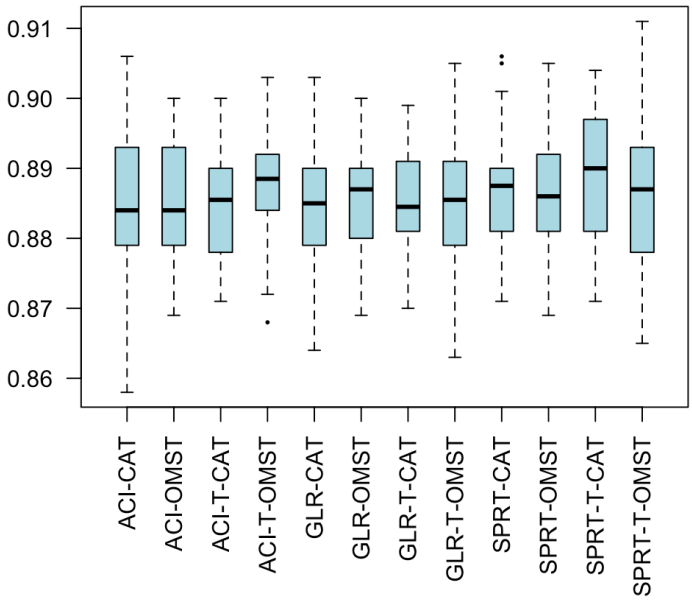


Figure 3.2 The boxplot of CAs from 30 replications under each of the 12 test designs

Note. ACI-T, SPRT-T, and GLR-T refer to the truncated versions of ACI, SPRT, and GLR, respectively.

In terms of test efficiency, the mean test length of 1000 examinees was calculated for 30 × 12 replications. Under each design, the 30 values were averaged, and the averaged means were reported in Table 3.2 and Figure 3.3. As Figure 3.3 shows, the variation trends in OMST and CAT are similar. In both OMST and CAT designs, the test length increased gradually as the

stopping rule switches from ACI, GLR, to SPRT; the test lengths produced by truncated designs are shorter than their corresponding counterparts.

Table 3.2

The mean and standard deviation of the test lengths from 30 replications under each of the 12 designs

	CAT Mean (SD)		OMST Mean (SD)	
	without truncation	with truncation	without truncation	with truncation
ACI	20.46 (0.174)	19.55 (0.205)	22.24 (0.262)	21.03 (0.150)
GLR	21.80 (0.207)	20.74 (0.197)	23.50 (0.176)	21.90 (0.175)
SPRT	27.14 (0.101)	25.05 (0.067)	28.01 (0.069)	25.38 (0.089)

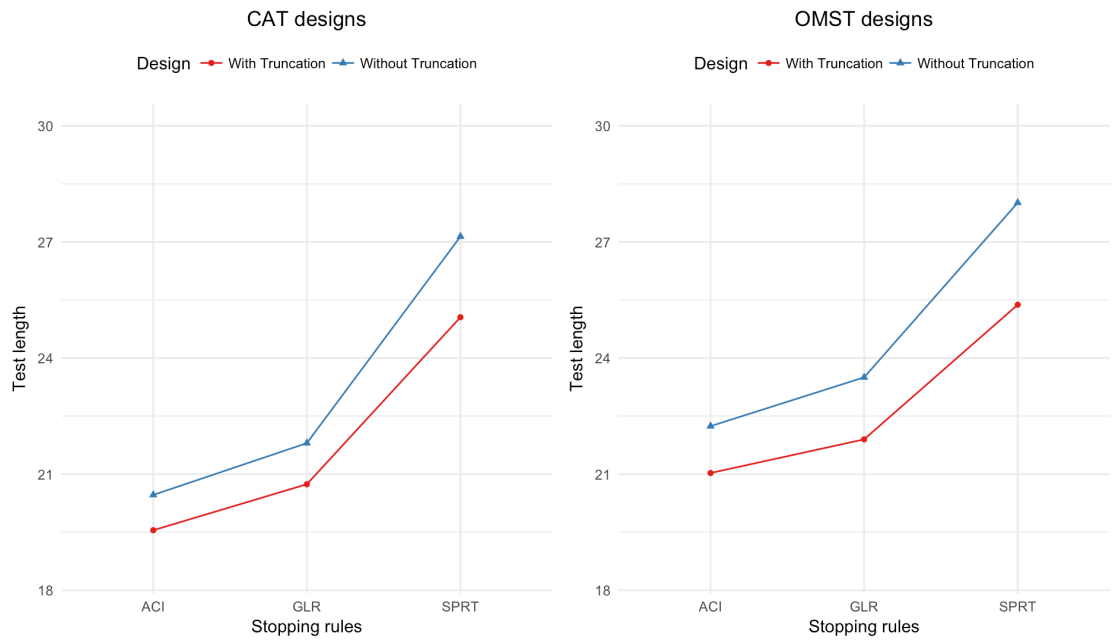


Figure 3.3. The averaged mean of test lengths produced by four different stopping rules.

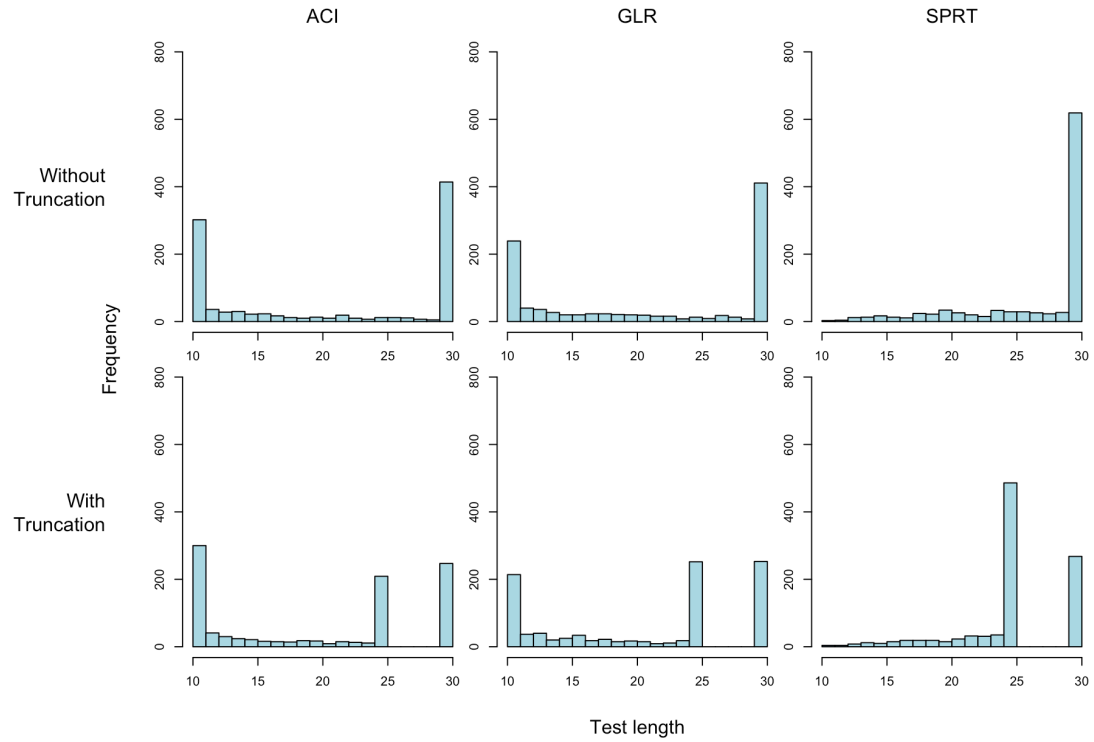
To have a better understanding of test efficiency, besides considering the mean test length over 1000 examinees, the distribution of test length was also examined. Since all the 30

replications of one design have a similar distribution, Figure 3.4 shows the histograms of test length in one randomly-selected replication under 12 test designs. Figures 3.4a and 3.4b present the test length distribution of CAT designs and OMST designs, respectively. The first and second row of these two graphs show the test length distribution of standard versions and truncated versions.

Although in OMST designs, there are only five possible values of test length, it is still apparent that the test length distribution in OMST designs is similar to that of CAT designs under each stopping rule. Across these stopping rules, the proportion of examinees who took a full 30-item test or a shortest 10-item test varies.

Overall, GLR designs produced similar results as ACI designs, and examinees took longer tests under SPRT designs. Though ACI and GLR produced comparable results, GLR performed slightly less efficiently for examinees whose ability is far from θ_{cut}, which is indicated by the smaller proportion of examinees who took a 10-item test. As for GLR's performance for examinees whose ability is near the θ_{cut}, no information can be gained through Figure 3.4 since the test would be stopped at the 30th item anyway. While in both ACI and GLR designs, about 200 hundred examinees out of 1000 took the shortest 10-item test, few examinees can finish the test after 10 items in SPRT designs. Unsurprisingly, SPRT designs also have the most examinees who took a full-length test. The truncation rule also plays a role in determining the proportion of examinees who took a full 30-item test. The difference between standard and truncated versions of ACI, SPRT, and GLR designs is apparent: some of the examinees who took more than 25 items in the standard test designs took a 25-item test in truncated designs instead. In truncated CAT designs, few examinees took tests of length larger than 25 but smaller than 30.

a. 6 CAT designs



b. 6 OMST designs

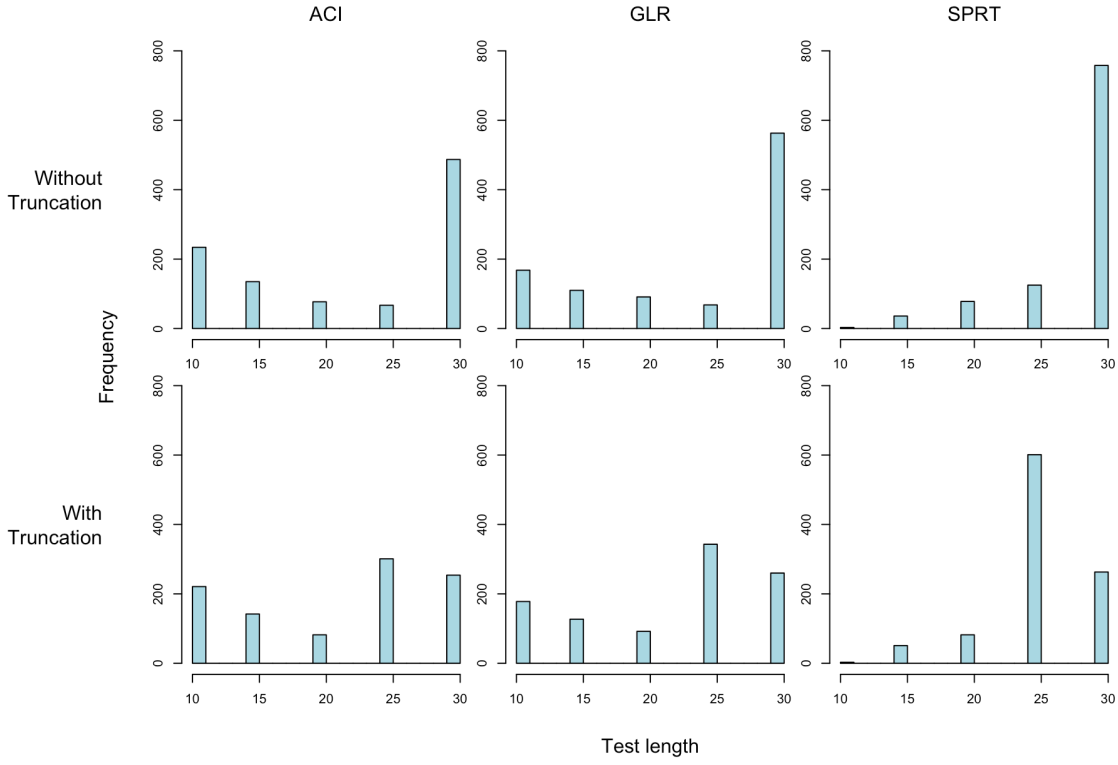
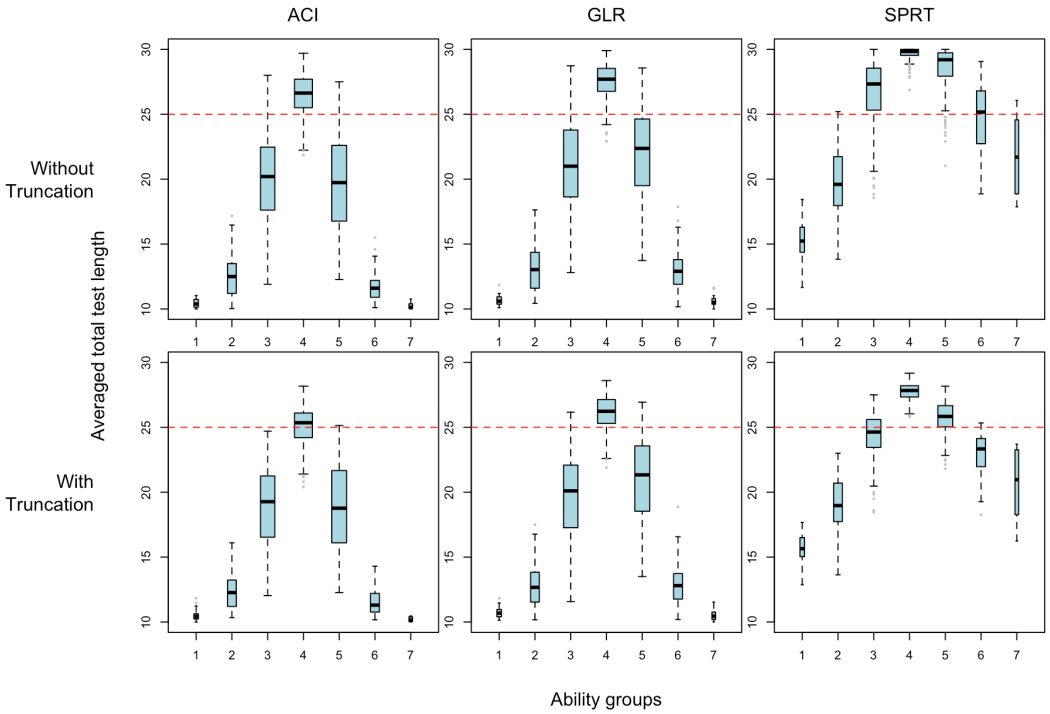


Figure 3.4. The histogram of test length in one replication under 12 test designs

Different ability groups may take tests of different length. Under each design, the 30 test lengths coming from 30 replications for every examinee were averaged. Then 1000 examinees were grouped into seven ability groups, and the test length distribution of each group was summarized in Figure 3.5. Group 1 contains examinees with ability parameters ranging from -4 to -2.22; group 2 is from -2.22 to -1.33; group 3 is from -1.33 to -0.44; group 4 is from -0.44 to 0.44; group 5 is from 0.44 to 1.33; group 6 is from 1.33 to 2.22; group 7 is from 2.22 to 4. The width of boxes is proportional to the square-roots of the number of observations in the groups, and outliers are denoted by grey points.

The test length distribution across theta levels in 6 OMST designs is similar to that of CAT designs; there are some common characteristics and some differences among 12 designs. The results of all designs show that for an examinee whose ability is far from θ_{cut}, 0, a shorter test was given to him/her; for an examinee with ability near to 0, a longer test was given. The closer one examinee's ability is to θ_{cut}, the longer test was given. The results of SPRT designs indicate a longer overall test length across all groups than ACI and GLR designs. Compared with the dashed line representing a test length of 25 items, the three graphs in the second row are "lower," especially for middle-ability groups (group 3, 4, and 5), which shows the test lengths of truncated designs are shorter across all ability groups.

a. 6 CAT designs



b. 6 OMST designs

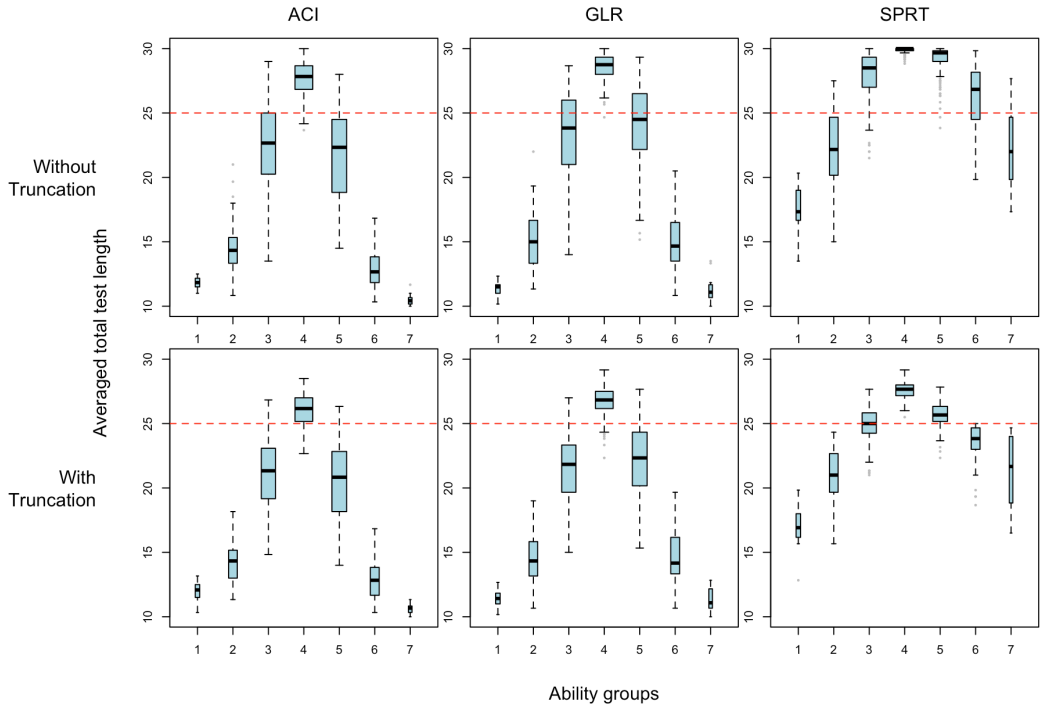


Figure 3.5. The boxplot of averaged test lengths for seven ability groups

The characteristics of the examinees whose tests were truncated to 25 items from longer ones were also examined. It can be seen from Figure 3.6 that, those examinees have true abilities ranging from -1.5 to 1.5, but their estimated abilities kept a certain distance from θ_{cut}.

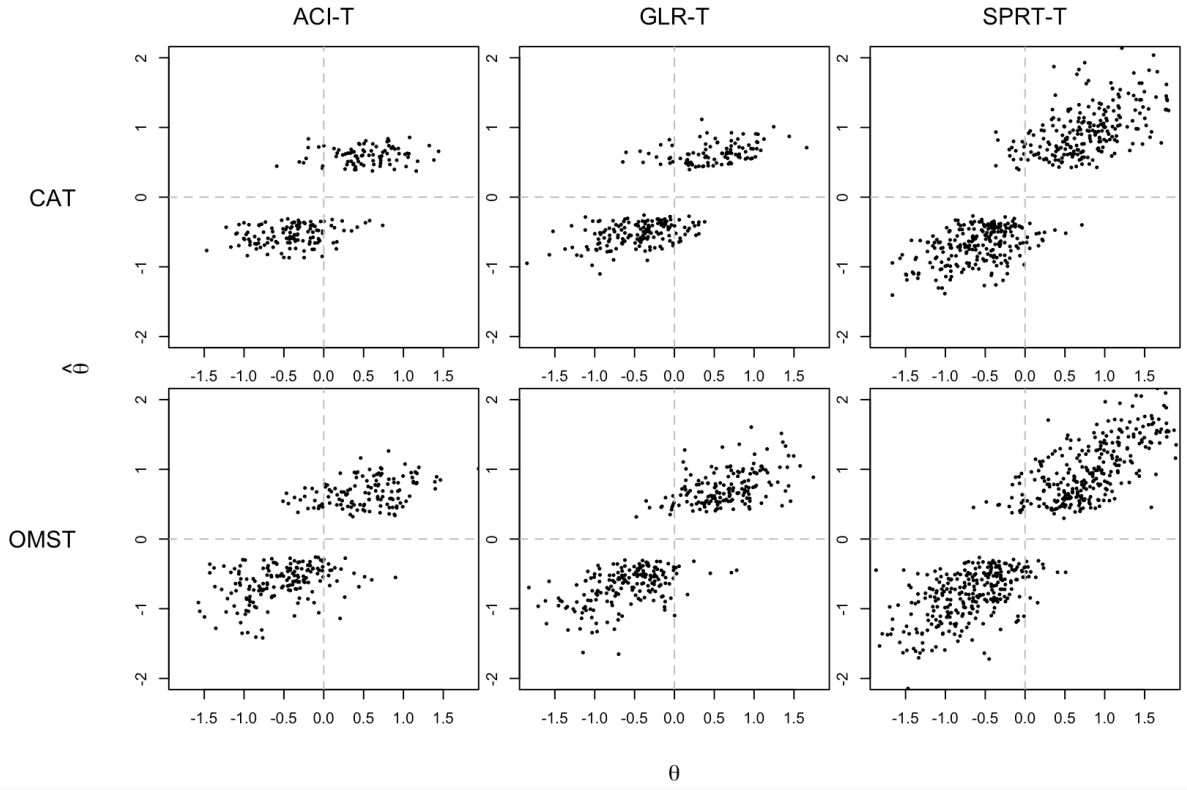


Figure 3.6. True and estimated abilities of examinees whose tests were truncated at the 25th item

Although achieving a high estimation accuracy is not the purpose of mastery tests, correlation coefficients between the true and estimated ability were still calculated and compared. Results are summarized in Table 3.3 and Figure 3.7. SPRT designs produced the highest averaged correlation coefficients, and ACI designs have the lowest ones. This is consistent with the results of mean test length. A test design with a longer test length also has a higher correlation coefficient.

Table 3.3

The mean and standard deviation of correlation coefficients from 30 replications under each of the 12 designs

	CAT Mean (SD)		OMST Mean (SD)	
	without truncation	with truncation	without truncation	with truncation
ACI	0.873 (0.0068)	0.871 (0.0056)	0.882 (0.0080)	0.879 (0.0059)
GLR	0.882 (0.0053)	0.881 (0.0067)	0.887 (0.0057)	0.883 (0.0068)
SPRT	0.918 (0.0040)	0.914 (0.0049)	0.921 (0.0035)	0.915 (0.0043)

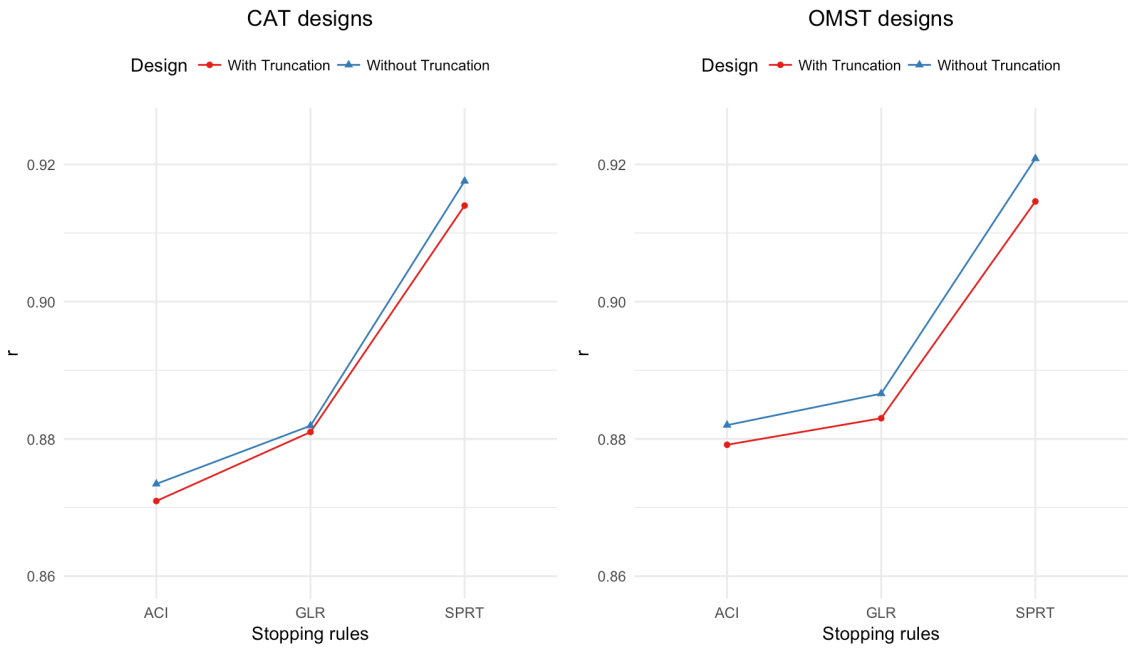


Figure 3.7. The averaged correlation coefficients between the true and estimated abilities under 12 test designs.

Non-statistical constraints were all well-controlled due to the MPI method. In 12 × 30 times of testing, none of the 360 items had an exposure rate greater than the pre-set value, 0.2. In terms of content constraints, no violation was observed. Thanks to the random selection of items in the initial stage, all items in the pool were used in one repetition.

CHAPTER 4: STUDY 2

The purpose of study 2 is to compare the performance of ACI and GLR designs when the test lengths are allowed to be longer. The methods are the same as in Study 1, except that the item pool consists of 1200 items and the total test length was set in between of 30 and 100. Given the similarity between CAT and OMST, only non-truncated CAT designs were conducted. 30 replications were conducted to reduce the random sampling error.

Results

The mean and standard deviation of 30 CAs and test lengths under both design were reported in Table 4.1. CA was controlled at the same level (0.939 for ACI and 0.938 for GLR). The test length of the ACI design is 4 items longer than that of the GLR design, which is opposite to Study 1's results. Figure 4.1 shows the test length distribution in one randomly selected repetition for these two designs. Compared with the GLR design, the ACI design had more examinees taking the shortest 30-item or full-length 100-item test than GLR, and fewer examinees taking the test of length in between.

Table 4.1

The mean and standard deviation of the classification accuracy and mean test length from 30 replications under 2 CAT designs

	ACI	GLR
	Mean (*SD*)	Mean (*SD*)
CA	93.86% (0.0063)	93.82% (0.0057)
Test length	50.70 (0.8050)	46.60 (0.7177)

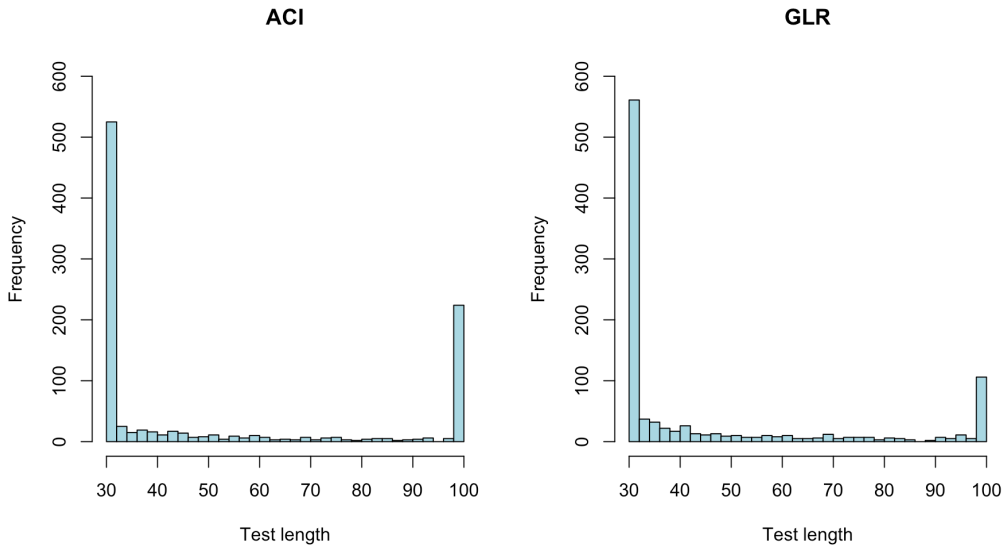


Figure 4.1. The histogram of test length in one replication for 2 CAT designs (30 ≤ test length ≤ 100)

CHAPTER 5: CONCLUSION AND DISCUSSION

The purpose of these studies is to compare the test lengths that resulted from four different stopping rules with classification accuracy controlled and to compare the results of CAT and OMST designs across different stopping rules. In study 1, Table 3.1 (page 17) shows that CA was successfully controlled: the averaged CAs under all designs are at the same level. An overall F-test from an ANOVA indicated that there was no significant difference among those 12 groups of CAs, so it is reasonable to assume the variation was caused by the random error. This control was achieved by varying the δ parameter in GLR and SPRT designs to match the CAs produced by ACI designs (Thompson, 2011).

OMST designs yielded similar results as CAT designs, which means the stopping rules designed for CAT also work for OMST in a similar way. In terms of the averaged mean of test length, OMST designs have around 1 or 2 items more than the corresponding CAT designs (Table 3.2, page 18). Although the values of test length in OMST can only be multiples of 5, it is apparent that the distribution of test length in OMST designs is similar to that of CAT designs (Figure 3.4, page 20). Both CAT and OMST designs with the SPRT rule have largest proportions of examinees who took a full-length test; the truncated versions of both CAT and OMST have many examinees whose tests were terminated at the 25th item rather than at the 30th item. The test length distribution across different ability groups of CAT and OMST designs are nearly the same (Figure 3.5, page 22). By comparing OMST with CAT, we can conclude that the four stopping rules established for CAT also function for OMST in a similar way. Despite less time of adaptation in OMST, OMST designs performed only slightly worse than CAT in terms of mean test length, which shows the statistical advantage of OMST. The high similarity between CAT and OMST stems from the same item selection method.

With CA controlled at the same level, comparing the averaged mean of test length and test length distribution can give us information about the efficiency of different stopping rules. ACI and GLR designs yielded comparable results, and SPRT is the least efficient rule. In Study 1, regardless of whether a design has a truncation rule or not, for both CAT and OMST, the averaged test lengths of 4 ACI designs are only one item less than that of 4 GLR designs (Table 3.2, page 18). Furthermore, the test length distribution of ACI and GLR is also similar: roughly 20%-30% of examinees finished the test after answering 10 items; 40%-60% of examinees took a test consisting of 25 items or more (Figure 3.4, page 20). As a result of the shorter averaged test length, the estimation accuracy of ACI designs is slightly lower than that of GLR designs (Table 3.3, page 24). However, since the purpose is to make a pass-or-fail decision in a mastery test, rather than to get an accurate ability estimate, and given the similar CAs of both ACI and GLR, the slightly lower correlation coefficient is not a problem. In study 2, though GLR outperforms ACI, the results are still similar. Compared with ACI and GLR, SPRT is the least efficient design. The test lengths of SPRT designs are always the largest (Table 3.2, page 17), with few examinees taking the shortest 10-item test and a big proportion of examinees taking the longest 30-item test or 25-item test (Figure 3.4, page 20).

It was predicted that ACI and GLR outperform SPRT in this study because the item selection method here is estimate-based instead of cutscore-based. As previous researchers analyzed, ACI works most effectively with estimate-based item selection because it decreases the standard error of measurement (Thompson & Prometric, 2007), while SPRT works most effectively with the cutscore-based item selection because it increases the P2-P1 difference (Lin & Spray, 2000). Unlike the ACI method, which takes advantage of the increasingly accurate ability estimate, the SPRT method does not fit well with the estimate-based item selection

method. The GLR method is an improved approach which solves this problem by introducing the likelihood of getting a specific response pattern given the current ability estimate. It allows $\theta_1$and θ_2 to vary, so the generalized likelihood ratio is more likely to be larger than A or smaller than B. This inclusion of a current ability estimate decreases the test length of examinees whose ability is far from the cutoff score, thus improving the test efficiency.

Though the overall results of ACI and GLR designs are comparable, Studies 1 and 2 provide different angles and show each method's advantages. The results of ACI designs in Study 1 are slightly better than that of GLR designs. However, when the test length was altered in Study 2, the GLR design performed more efficiently. In Study 1, it can be seen that for examinees whose ability is far from the cutoff point, GLR performed slightly worse than ACI: fewer examinees in GLR designs took a 10-item test (Figure 3.4, page 20). In other words, for low/high ability examines, GLR needs more items to give a satisfactory decision than ACI. However, since the maximum test length is only 30, it is hard to know whether GLR performs better for examinees whose ability is near the cutoff point. This question was explored in Study 2, where the maximum test length was set to 100. Results show that, compared with the GLR design, the ACI design had more examinees take a full-length test (Figure 4.1, page 26), which means that GLR does perform better for examinees whose ability is near the cutoff point. Therefore, to choose an appropriate stopping rule for a test, the composition of test takers and the desired test length also influence the efficiency of a test. Rather than assuming a normal distribution of examinees' abilities, if most test takers' abilities are near the cutoff point, GLR may be more appropriate.

When comparing the truncated versions of ACI, SPRT, and GLR designs with their non-truncated standard versions in Study 1, we can conclude that the truncation rule can shorten the

mean test length by preventing a proportion of examinees from taking the unnecessarily long 30-item test. Since the classification decision will not change even if these examinees answered the remaining items all correctly or all incorrectly, this truncation rule will not cause a sacrifice in CA. By comparing the test length distribution of truncated CAT designs with standard CAT designs, it can be seen that some of the examinees with test length 25 in the truncated version would take a 30-item test in a standard design (group 1); some would take a test of length more than 25 and less than 30 (group 2); and some would still take a 25-item test (group 3). Intuitively, before truncating the tests of group 1, the tests of group 2 are more likely to be truncated, so it is reasonable to see few examinees in truncated CAT designs took a test of length larger than 25 and smaller than 30.

As for examinees who took the 25-item truncated test, a closer look is given in Figure 3.6 (page 23), and it shows the truncation rule can discriminate examinees who do not really need a full-length test. It is expected that only the examinee with his/her true ability near the cutoff point can proceed to the upper bound of the test length limit. However, given that a 30- or 25-item test is still short, it is reasonable to observe that the true ability of those examinees has a relatively wide range, -1.5 to 1.5. As for their estimated ability values, if the estimated ability is near θ_{cut}, two extreme final estimates are more likely to fall on to both sides (one greater than θ_{cut} and one smaller than θ_{cut}), and this examinee would take a full-test instead. Therefore, the estimated abilities of examinees who took the 25-item truncated test are not close to the cutoff point.

This project has some implications for future research. First, it provides evidence supporting the good statistical properties of variable-length OMST. In this study, the only difference between OMST and CAT is that after the initial stage, CAT designs would estimate

the current ability and choose the next one item after administering each item, while OMST designs would estimate the ability and choose the next stage which consists of 5 items after administering each stage. Though the times of adaptation were decreased in the OMST design, its statistical property was still comparable with CAT designs. This is because unlike the traditional multistage testing that has pre-defined difficulty levels or statistical properties for each module, the properties of the next stages in OMST designs are allowed to be decided based on the current ability estimate and previous responses. It enables OMST to achieve a high level of flexibility and to compensate for the disadvantages of less adaptation. On the other hand, because the items in OMST are administered in a bundle, OMST allows test takers to review and skip items in one stage, which is a great relief for test takers and may lead to a more accurate measurement due to the lower level of anxiety. OMST is a good candidate for test developers who want to cause less stress for examinees while sustaining a similar level of estimation or classification accuracy. Beyond the previous study, this study provides extra evidence showing the advantages of OMST even when the test is designed to be variable-length, and it indicates the stopping rules designed for CAT can also be applied to OMST.

Second, this research provides a comparison between different stopping rules and applied a new truncation rule in on-the-fly adaptive testing designs. The results of SPRT and ACI are consistent with previous studies, which showed the efficiency of the ACI method combined with estimate-based item selection method. As for ACI and GLR, a thorough comparison was made under 2 situations. The results indicate that since both ACI and GLR take the current ability estimate into consideration, ACI and GLR yielded overall similar results. ACI is more efficient for examinees far from the cutoff point, while GLR can reduce the test length of examinees whose ability is near the cutoff point. This finding is useful for practitioners when choosing

stopping rules for a specific population. The truncated test design is straightforward and shows the promising future of applying the truncation rule in on-the-fly adaptive testing rather than fixed-form testing.

There are some limitations that can be explored in future studies. First, this study examined the results of a mastery test which classifies students into only two categories. Intuitively, due to the similarities between OMST and CAT in item selection, the variation trend across different stopping rules in OMST is expected to be similar to that in CAT. Since more categories are included, and a more accurate estimation is needed for each examinee to be classified, OMST may produce even longer tests than CAT. For tests designed for multi-classification, different stopping rules may perform differently. While the test lengths of SPRT designs may remain at the same level, the test length of the others may increase. This is because even though the estimation precision for abilities far from the cutoff point is not necessary for this study, SPRT still achieved higher accuracy than others for examinees at the two ends. In a mastery test, this improvement of the overall accuracy is redundant, but for tests designed for multi-classification, it is needed. As for GLR and ACI, since GLR performs better for examinees near the cutoff point, GLR may produce a shorter test length. Second, a more sophisticated truncation rule can be developed. In this study, when to temporarily stop and check the necessity of administering more items was arbitrarily set. Though this method is effective and not computationally demanding, future studies may develop a method which can estimate the probability of altering the decision if administering more items after the administration of each item.

REFERENCES

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*(3), 237-247.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques.* CRC Press.

Chang, H. H., & Ying, Z. (1999). *A-stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23*, 211–222.

Chang, H-H., & Ying, Z. (2002, April). *To weight or not to weight? Balancing influence of initial items in adaptive testing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Chang, H-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*(3), 441–450.

Chang, H. H. (2015). *Psychometrics behind computerized adaptive testing. Psychometrika, 80*(1), 1-20.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369-383.

Eggen, T. J. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249-261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, *60*(5), 713-734.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Maheah.

Finkelman, M. (2003). *An Adaptation of Stochastic Curtailment to Truncate Wald's SPRT in Computerized Adaptive Testing*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In *New horizons in testing* (pp. 257-283).

Lewis, C., Sheehan, K. M., DeVore, R. N., & Swanson, L. C. (1991). *U.S. Patent No. 5,059,127*. Washington, DC: U.S. Patent and Trademark Office.

Lin, C. J., & Spray, J. (2000). Effects of Item-Selection Criteria on Classification Testing with the Sequential Probability Ratio Test. ACT Research Report Series.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, *31*, 3–31.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224–236.

National Council of State Boards of Nursing. (2012). 2013 Nurse Licensee Volume and NCLEX® Examination Statistics.

Nitko, A., & Hsu, T. C. Using domain referenced tests for student placement, diagnosis, and attainment in a system of adaptive individualized instruction. *Educational Technology,* 1974, 14, 48-53.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing *Journal of the American Statistical Association*, *70*(350), 351–356.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In *New horizons in testing* (pp. 237-255).

Schnipke, D. L., & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing* (Law School Admissions Council Computerized Testing Report 97-01). Newtown, PA: Law School Admission Council.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151–166.

Thompson, N. A., & Prometric, T. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, *12*(1), 1-13.

Thompson, N. A. (2011). Termination Criteria for Computerized Classification Testing. *Practical Assessment, Research & Evaluation*, *16*.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259–270.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp. 27-52). Norwell, MA: Kluwer Academic Publishers. DOI: 10.1007/0-306-47531-6

Wald, A. (1947). Sequential Analysis Wiley. *New York*.

Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: from group sequential design to fully sequential design. *Journal of Educational Measurement*, *53*(1), 45-62.

Weiss, D. J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In *Proceedings of the first conference on computerized adaptive testing* (pp. 24–35). Washington, DC: United States Civil Service Commission.

Zheng, Y., & Chang, H. H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. *Advancing methodologies to support both summative and formative assessments*, 21-39.

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118.