On Computing the Key Probability in the Stochastically Curtailed Sequential Probability Ratio Test

Applied Psychological Measurement
2016, Vol. 40(2) 142–156
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615611633
apm.sagepub.com

Alan R. Huebner¹ and Matthew D. Finkelman²

Abstract

The Stochastically Curtailed Sequential Probability Ratio Test (SCSPRT) is a termination criterion for computerized classification tests (CCTs) that has been shown to be more efficient than the well-known Sequential Probability Ratio Test (SPRT). The performance of the SCSPRT depends on computing the probability that at a given stage in the test, an examinee's current interim classification status will not change before the end of the test. Previous work discusses two methods of computing this probability, an exact method in which all potential responses to remaining items are considered and an approximation based on the central limit theorem (CLT) requiring less computation. Generally, the CLT method should be used early in the test when the number of remaining items is large, and the exact method is more appropriate at later stages of the test when few items remain. However, there is currently a dearth of information as to the performance of the SCSPRT when using the two methods. For the first time, the exact and CLT methods of computing the crucial probability are compared in a simulation study to explore whether there is any effect on the accuracy or efficiency of the CCT. The article is focused toward practitioners and researchers interested in using the SCSPRT as a termination criterion in an operational CCT.

Keywords

computerized testing, classification, item response theory

Introduction

Computerized classification tests (CCTs) are assessments that aim to categorize an examinee based on his or her responses to a set of items. CCTs are often used for professional certification exams that classify examinees as masters or non-masters (i.e., passes or fails). CCT methodology has been a consistently fertile area of psychometric research over the past several decades,

¹University of Notre Dame, IN, USA

²Tufts University School of Dental Medicine, Boston, MA, USA

Corresponding Author:

Alan R. Huebner, Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 153 Hurley Hall, Notre Dame, IN 46556, USA.

Email: Alan.Huebner.10@nd.edu

and there has been much inquiry into and many innovations proposed for CCT item selection (e.g., Eggen, 1999; Lin, 2011; Spray & Reckase, 1994; Thompson, 2009; Van Groen, Eggen, & Veldkamp, 2014), termination rules (e.g., Finkelman, 2008, 2010; Thompson, 2011; Weiss & Kingsbury, 1984), and test security via item exposure or overlap control (e.g., Chen, Lei, Chen, & Liu, 2014; Huebner, 2012). Although it is possible to construct a CCT using a variety of measurement models to classify examinees into one of two or more categories, much research has been directed toward CCTs based on item response theory (IRT) models for dichotomously scored items that classify examinees into one of exactly two categories. This type of CCT will be the focus of the present study.

An essential consideration for the design of a CCT is its efficiency or ability to classify examinees correctly using a minimal number of items. Appropriate item selection rules may bolster the efficiency of CCTs, and termination rules can increase efficiency by stopping CCTs when an accurate classification can be made; the Sequential Probability Ratio Test (SPRT; Wald, 1947) and ability confidence intervals (ACI; Weiss & Kingsbury, 1984) are two well-known examples. This article will focus on a variation of the SPRT proposed by Finkelman (2008) for cases in which a maximum number of items is imposed on the CCT, a common practice in operational settings. (Some authors use the term *truncated SPRT* to describe this case, but the present study will use SPRT for simplicity.) Finkelman (2008) showed that the SPRT is inefficient in this situation; specifically, the SPRT may allow items to be administered even if the responses will not affect the ultimate classification decision. Thus, the exposure of these items is unnecessary, and he proposed the more efficient Stochastically Curtailed SPRT (SCSPRT). In a simulation study using item parameters from an operational CCT using item exposure and content balancing, the SCSPRT reduced average test length (ATL) by approximately 7 to 10 items compared with the SPRT for examinees with ability levels near the pass/fail threshold with virtually no loss of accuracy.

To facilitate further discussion of the SCSPRT and describe the purpose of this article, some notation must be introduced. Suppose a given CCT allows an examinee to respond to a maximum of J items, and let D_J denote the classification decision after the administration of all J items. This ultimate decision may be $D_J = m$ or n , indicating mastery or non-mastery, respectively. Also, let D_k^* represent the tentative, or interim, classification decision at Stage k of the test, after the administration of the k th item. Similar to D_J , D_k^* may be m or n . As will be described in detail in the next section, at each stage, the SCSPRT computes $P(D_k^* = D_J)$, the probability that the current classification decision will not change before the test is terminated after item J . If $P(D_k^* = D_J)$ is greater than a pre-specified probability threshold γ , the test is stopped at Stage k even if the more conservative SPRT criterion does not indicate termination.

Clearly, the computation of $P(D_k^* = D_J)$ is a crucial element of the SCSPRT; Finkelman (2008) proposed two methods of computing it. The authors of the present study give a conceptual overview of the methods here and save the detailed explanation for the next section. At Stage k of the test, $P(D_k^* = D_J)$ may be computed exactly by considering all possible response patterns to the $J - k$ remaining items. However, if $(J - k)$ is large, the exact method may become computationally burdensome, and a central limit theorem (CLT) approximation may be used for simplicity. As will be described in the next section, the CLT method does not explicitly take into account each possible remaining response pattern. Rather, it uses a normal approximation to compute $P(D_k^* = D_J)$ conditional on the examinee's interim classification status. Hereafter, $P(D_k^* = D_J)$ calculated under the exact and CLT methods are referred to as P_{Exact} and P_{CLT} , respectively. This article will examine the relation between P_{Exact} and P_{CLT} from a practical perspective, geared toward researchers and psychometricians interested in using the SCSPRT method to increase the efficiency of an operational CCT. Although Finkelman (2008)

demonstrated the general efficacy of the SCSPT in reducing ATLs, many issues need to be investigated if the SCSPT were to be used in an operational setting.

For example, consider a CCT having a maximum of $J = 50$ dichotomously scored items. Computing P_{Exact} at Stage 30 of the test, when $(J - k) = 20$, would require evaluating the probabilities of 2^{20} potential response patterns. This may be too computationally burdensome for an operational CCT, as it is undesirable for examinees to experience a noticeable delay between the administrations of items. Thus, in this situation, the use of P_{CLT} would be preferable. However, when $(J - k)$ is small, say 5, the P_{CLT} may not be accurate, as the CLT generally performs better with larger sample sizes. Moreover, P_{Exact} could be easily computed in this case, as there are only 2^5 response patterns for which to account. Thus, it is reasonable that P_{CLT} should be used early in the test whereas P_{Exact} should be used at later stages, but there are many yet unanswered concerns that would be raised by practitioners. This research addresses the following questions:

Research Question 1: Is there an optimal test stage for switching from using P_{CLT} to P_{Exact} ?

Research Question 2: Does one method consistently under- or overestimate the true probability that the classification decision will be changed?

Research Question 3: How are the overall ATLs and classification accuracy affected by P_{CLT} versus P_{Exact} ?

Research Question 4: Do the answers to these questions depend on the quality of the item bank or other elements of the CCT design?

Moreover, this is the first study that compares P_{CLT} and P_{Exact} with the quantity P_{True} . As will be explained in the following sections, P_{True} is the true probability of keeping the same classification decision, based on the true value of the examinee's latent ability. Although P_{True} would not be known in practice, it is treated in this simulation study as a gold standard with which P_{CLT} and P_{Exact} are compared.

Method

IRT

The three-parameter logistic (3PL) IRT model quantifies the interaction between an item and an examinee with a given ability level. Specifically, let X_j denote the response of an examinee to item j , where $X_j = 1$ or 0 indicates a correct or incorrect response, respectively, and θ is the latent ability level of the examinee. Then, the item response function (IRF) of the 3PL is given by

$$P(X_j = 1|\theta) = c_j + \frac{1 - c_j}{1 + \exp[-1.7a_j(\theta - b_j)]}, \quad (1)$$

where a_j , b_j , and c_j are the discrimination, difficulty, and pseudo-chance parameters of item j , respectively (Birnbaum, 1968). It is usually assumed that the responses to a series of k items for a particular examinee are independent conditional on the examinee's θ level (Lord, 1980). Then, the likelihood function for the responses to items 1, $\dots$, k from a particular examinee is given by

$$L(\mathbf{X}_k; \theta) = \prod_{j=1}^k L(X_j; \theta) = \prod_{j=1}^k P(X_j = 1 | \theta)^{X_j} [1 - P(X_j = 1 | \theta)]^{1-X_j}. \quad (2)$$

Note that the boldface notation is used to denote a vector of responses, $\mathbf{X}_k = (X_1, \dots, X_k)'$, while X_j represents the response to a single item.

SPRT

To classify examinees into two categories, a cut point on the θ scale is required. The authors denote this value θ_0 , and examinees with $\theta \geq \theta_0$ are deemed masters, and non-masters otherwise. Reflecting the fact that classifications cannot be perfect because θ is a latent variable, a small quantity, δ , is added to and subtracted from θ_0 to create an “indifference” region. The authors define $\theta_+ = \theta_0 + \delta$ and $\theta_- = \theta_0 - \delta$ as the upper and lower bounds of the indifference region, and misclassifications of examinees with true θ in this range are not considered severe errors. The SPRT formulates the classification decision as accepting a null (H_0) or alternative (H_a) hypothesis, $H_0 : \theta = \theta_-$ versus $H_a : \theta = \theta_+$. Under this formulation, the null and alternative serve as surrogates for hypotheses $\theta \leq \theta_-$ and $\theta \geq \theta_+$, respectively (Spray & Reckase, 1996). If H_a is accepted, the examinee is classified as a master (pass); if H_0 is accepted, the examinee is judged a non-master (fail). The decision is based on the log of the likelihood ratio at Stage k of the test defined as

$$\log \lambda_k = \log \frac{L(\mathbf{X}_k; \theta_+)}{L(\mathbf{X}_k; \theta_-)} = \sum_{j=1}^k \log \frac{L(X_j; \theta_+)}{L(X_j; \theta_-)}. \quad (3)$$

Also, the desired Type I (judging a non-master as a master) and Type II (judging a master as a non-master) error rates are specified and denoted as α and β , respectively. In addition, define constants $A = \alpha / (1 - \beta)$ and $B = (1 - \alpha) / \beta$ such that $A < 1 < B$. Then, the SPRT termination criterion proceeds as follows:

1. At Stage k , calculate $\log \lambda_k$ as shown in Equation 3.
2. If $\log \lambda_k \geq \log B$ or $\log \lambda_k \leq \log A$, the test terminates with an m or n decision, respectively.
3. If the test is not terminated by the criterion in Step 2, proceed to Stage $k + 1$ by administering the next item and repeat Steps 1 to 2.
4. If the maximum number of items J is administered without a classification being made, a “forced” decision results. If $\log \lambda_J \geq C$ or $\log \lambda_J < C$, the test terminates with an m or n decision, respectively, where $C = [\log A + \log B] / 2$. (Note that if α and β are equal, C simplifies to zero.)

The individual elements of the SPRT discussed above, α , β , θ_0 , and δ , are quantities that would be determined in practice by collaboration between psychometricians, subject-matter experts, and the testing agency. In recent CCT literature, δ values between .1 and .3 have been common. For example, Finkelman (2008, 2010) used $\delta = .2$, Huebner and Fina (2015) used $\delta = .1$ and .2, Lin (2011) examined $\delta = .1$, .2, and .3, and Eggen (1999) conducted simulations in which δ was varied from .10 to .23 by increments of .01. Also, all simulations in those articles set $\alpha = \beta = .05$. Values of θ_0 used in the above references range from -1.32 to 1.50 . Huebner and Fina (2015) presented a simulation study in which θ_0 was set to -1 , 0 , and 1 in various

testing conditions. On average, conditions with $\theta_0 = 0$ tended to have larger ATLs and lower classification accuracy than conditions with $\theta_0 = -1$ or 1 .

SCSPRT

As mentioned in the previous section, the SCSPRT termination criterion relies on the calculation of $P(D_k^* = D_J)$ to increase test efficiency compared with the SPRT. The authors first give a general description of the method and then provide details on computing the exact and CLT approximation versions. The SCSPRT is implemented according to the following steps:

1. At Stage k , calculate $\log\lambda_k$ as shown in Equation 3.
2. If $\log\lambda_k \geq \log B$ or $\log\lambda_k \leq \log A$, the test terminates with an m or n decision, respectively.
3. If the test is not terminated at Stage k by the criterion in Step 2, then calculate $P(D_k^* = D_J)$. If $P(D_k^* = D_J) \geq \gamma$, the test terminates with an m decision if $\log\lambda_k \geq C$ or an n decision if $\log\lambda_k < C$.
4. If the test is not terminated by the criteria in Steps 2 or 3, proceed to Stage $k + 1$ by administering the next item, and repeat Steps 1 to 3.
5. If the maximum number of items J is administered without a classification being made, a forced decision results as described in the previous subsection for the SPRT.

In summary, the key difference between the SCSPRT and the traditional SPRT is that the SCSPRT aims to shorten the exam by taking into account potential future items, whereas the SPRT does not. The SCSPRT uses the SPRT within it, but it is allowed to terminate the exam earlier by calculating $P(D_k^* = D_J)$. The authors note that the SCSPRT terminates the exam whenever the SPRT does, and thus, the SCSPRT never administers more items to a given examinee than the SPRT. Also, note that as γ increases, the SCSPRT is less likely to terminate the exam before the SPRT. In many practical settings, it would be desirable to maintain classification accuracy, and hence, a conservative value (close to 1) should be chosen. Past studies on the SCSPRT (Finkelman, 2008, 2010) have used $\gamma = .95$.

Computation of P_{Exact} and P_{CLT}

The authors now discuss the computation of $P(D_k^* = D_J)$, starting with the exact version P_{Exact} . First, at Stage k , the current interim classification is obtained by setting D_k^* to m or n if $\log\lambda_k \geq C$ or $\log\lambda_k < C$, respectively. Then, for $(J - k)$ remaining items, there are $2^{(J-k)}$ possible response patterns. The probability of each response pattern may be computed using Equations 1 and 2, given the parameters of the remaining items and a θ value. It is assumed that an operational CCT has a well-calibrated item bank, so the only issue is choosing a θ value to input. Finkelman (2008) used a conservative approach originating in the field of clinical trials (Lan, Simon, & Halperin, 1982). This approach, termed “conditional power” by Jennison and Turnbull (2000), sets $\theta = \theta_+$ if $D_k^* = n$, and $\theta = \theta_-$ if $D_k^* = m$. In addition to the probability of each potential response pattern, the $2^{(J-k)}$ corresponding potential λ_J s are computed. Those λ_J s are used to indicate potential classification decisions, and the probabilities of those decisions that match D_k^* are summed to obtain P_{Exact} .

Note that if the true value of θ was known, it could be input into Equation 1 rather than θ_- or θ_+ and used in the subsequent calculations to obtain the “true” value of $P(D_k^* = D_J)$. This is, of course, not possible in practice, but it can easily be done in a simulated exam when the true θ s are known and used to generate item responses. This true calculation of $P(D_k^* = D_J)$ is

denoted as P_{True} ; it is interpreted as the probability of keeping the same classification decision calculated using the examinee's true θ value. As will be further explained in the next section, P_{True} will be regarded as a gold standard by which to judge the performance of P_{Exact} and P_{CLT} . Specifically, in "Simulation Study" section, the authors will describe several measures to examine the agreement of P_{Exact} and P_{CLT} to P_{True} .

Finkelman (2008) provides the formulas for a CLT approximation of $P(D_k^* = D_J)$. In contrast to P_{Exact} , the calculation of P_{CLT} does not explicitly take into account each potential remaining response pattern. Instead, P_{CLT} is a function of $\log \lambda_k$, the θ value obtained by the conditional power approach described above, and the parameters of the remaining $(J - k)$ items. Thus, computing P_{CLT} is much more efficient than P_{Exact} , especially when $(J - k)$ is large. A review of the CLT formulas, as well as a worked example illustrating P_{Exact} , P_{CLT} , and P_{True} , are included in the supplementary online appendix.

Simulation Study

A simulation study was designed to compare the performance of the SCSPT using P_{CLT} and P_{Exact} at different stages of a CCT and to investigate whether there are any sizable differences in the two methods in terms of ATL or classification accuracy. For clarity of discussion, the study is divided into two parts, Part 1 and Part 2. All analyses described in both parts of the study were performed using the 3PL IRT model described in "Item Response Theory" section for four different simulated testing conditions. For all conditions, examinees were administered CCTs with a maximum of $J = 50$ items, cut point $\theta_0 = 1.0$, desired error rates $\alpha = \beta = .05$, and, for the SCSPT, probability threshold $\gamma = .95$. One CCT factor that was varied was the half length of the indifference region δ . In two conditions, $\delta = .1$, and in two conditions, $\delta = .2$. These CCT parameter settings were chosen to be representative of the settings chosen in other recent studies as summarized at the end of "SPRT" section. Two different item banks were constructed. For both banks, the discrimination and pseudo-chance parameters were generated from the distributions $a_j \sim \text{Normal}(0.7, 0.2)$ and $c_j \sim \text{Normal}(0.25, 0.03)$, respectively. The difficulty parameters were generated in such a way that the item information for one bank was "peaked" about θ_0 and "broad" about θ_0 for the other. For the peaked bank, $b_j \sim \text{Normal}(\theta_0 - 0.25, 0.2)$, and $b_j \sim \text{Normal}(\theta_0 - 0.25, 1)$ for the broad bank. The reasoning behind this setup is that the information for both banks should be centered near the cut point, but the spread around the cut point may differ. The authors note that this method of generating peaked and broad item banks as well as the specific item parameter settings are very similar to those of Thompson (2009). The items were assumed to be calibrated without error, a simplifying assumption commonly used in the CCT and computer adaptive testing literature, for example, Thompson (2009), Wang and Chang (2011), Huebner and Fina (2015), and Fan, Wang, Chang, and Douglas (2012).

Both levels of δ were applied to both item banks, resulting in the four testing conditions. Furthermore, for each condition, the CCTs were administered to $N = 5,000$ simulated examinees with θ s generated from the $\text{Normal}(0, 1)$ distribution. Items were selected by the well-known method of maximizing Fisher information (FI) at θ_0 . The authors note that although Kullback–Leibler information has also been studied as an alternative to FI, the two item selection criteria perform very similarly in terms of efficiency and accuracy in practical testing situations (Lin, 2011). Moreover, recent studies on SCSPT methodology (Finkelman, 2008, 2010; Huebner & Fina, 2015) use the FI index. Also, to mimic a practical testing situation, the Symptom–Hetter method of item exposure control (Symptom & Hetter, 1985) was used with a desired maximum exposure rate of 0.20.

Simulation Study—Part 1

Part 1 of the simulation study investigated the performance of the SCSPT termination rule when P_{CLT} and P_{Exact} were used at different stages of test for the computation of P_{CLT} . As stated in the introduction, computing P_{CLT} is not burdensome and can be done at any stage of the test. However, calculating P_{Exact} very early in the test would not be feasible in practice or in a simulation study because of the large number of potential response patterns for which to account. Thus, this study begins comparing P_{CLT} and P_{Exact} at Test Stage 35, when there are 15 items remaining and 2^{15} possible remaining response patterns. Specifically, let T_{Exact_35} be the SCSPT termination rule that uses P_{CLT} until Stage 34 of the test and then switches to using P_{Exact} at Stage 35 and for the remainder of the test. Then, let T_{Exact_36} be the SCSPT termination rule that switches from using P_{CLT} to P_{Exact} at Stage 36 of the test, and so on. Also, let T_{CLT} be the SCSPT termination rule that uses P_{CLT} throughout the entire exam; note that for a 50-item test, $T_{CLT} \equiv T_{Exact_50}$.

All the above termination rules, along with the original SPRT to serve as a baseline, were implemented simultaneously on the test of each simulated examinee. If one rule stopped the test, the others were allowed to keep administering items until they also terminated. Thus, a “matched” comparison of the methods was produced, as was done in the simulation study reported by Finkelman (2008). The termination rules were evaluated by their ATLs and proportion of examinees classified correctly (PCC).

Simulation Study—Part 2

Part 2 of the simulation study examined the relationships between P_{CLT} , P_{Exact} , and P_{True} for Test Stages 35 through 49. For each of those 15 test stages, the mean bias (MB) and correlation (COR) between (P_{CLT}, P_{True}) , (P_{CLT}, P_{Exact}) , and (P_{Exact}, P_{True}) were computed for those examinees whose test had not yet been terminated by the T_{CLT} (the reasoning behind this will be discussed momentarily). Specifically, $MB(P_{CLT}, P_{True}) = \frac{1}{N_j} \sum_{N_j} (P_{CLT} - P_{True})$ was calculated at each stage $j = 35, \dots, 49$, where N_j denotes the number of examinees with exams still in progress at Stage j . Also, $COR(P_{CLT}, P_{True})$ is the Pearson product-moment correlation between P_{CLT} and P_{True} . The authors note that these quantities were computed also for (P_{CLT}, P_{Exact}) and (P_{Exact}, P_{True}) , but for the sake of brevity, the formulas are displayed only for (P_{CLT}, P_{True}) . The MB and COR describe the relationships between the three probabilities as the test progresses. As mentioned in the previous section, P_{True} is regarded as a gold standard because it is based on the true, or generating, examinee θ value. Thus, small MB and high COR with P_{True} would indicate good performance for P_{CLT} and P_{Exact} .

The study also examined quantities at Test Stages 35 through 49 that describe the relationships between P_{CLT} , P_{Exact} , P_{True} , and γ . The false positive and false negative rates (FPR and FNR) for P_{CLT} and P_{Exact} at Stage j of the test are defined as

$$FPR(P_{CLT}) = P(P_{CLT} \geq \gamma | P_{True} < \gamma) = \frac{(\# \text{examinees with } P_{CLT} \geq \gamma \text{ and } P_{True} < \gamma)}{(\# \text{examinees with } P_{True} < \gamma)}$$

and

$$FNR(P_{CLT}) = P(P_{CLT} < \gamma | P_{True} \geq \gamma) = \frac{(\# \text{examinees with } P_{CLT} < \gamma \text{ and } P_{True} \geq \gamma)}{(\# \text{examinees with } P_{True} \geq \gamma)},$$

and similar for P_{Exact} . As above, the FPR and FNR were based on only those examinees whose exams had not been terminated yet by T_{CLT} at that stage. For a given Test Stage j , the FPR and

FNR have the following interpretations. $FPR(P_{CLT})$ is the proportion of times the SCSPT terminates the exam based on P_{CLT} , among cases where testing would not have stopped based on P_{True} . $FNR(P_{CLT})$ is the proportion of times the SCSPT does not terminate the exam based on P_{CLT} , among cases where testing would have stopped based on P_{True} . That is, $FNR(P_{CLT})$ is the observed conditional probability that the SCSPT fails to stop the test using P_{CLT} , given that the SCSPT would stop the test using P_{True} . $FPR(P_{CLT})$ indicates whether P_{CLT} is causing the SCSPT to terminate the exam too aggressively compared with P_{True} , possibly resulting in decreased classification accuracy. However, the $FNR(P_{CLT})$ indicates whether the P_{CLT} is making the SCSPT too conservative, possibly resulting in reduced efficiency. Of course, these interpretations also hold for $FPR(P_{Exact})$ and $FNR(P_{Exact})$.

The choice to compute the quantities described above is now discussed based on examinees whose exams had not yet been terminated by the T_{CLT} criterion. First, consider the alternate plan of using all examinees at a given test stage regardless of whether their tests had been terminated or not. This plan would not be representative of the actual usage of the SCSPT, because in practice, P_{CLT} , P_{Exact} , and P_{True} are only computed for examinees whose tests have not yet been terminated. Moreover, because the examinees whose tests have already terminated are those who exhibit stable classifications, they typically exhibit values of P_{CLT} , P_{Exact} , and P_{True} that are all close to 1 (and therefore are close to one another). Including such examinees in the analysis would falsely inflate the agreement between these probability estimates compared with their agreement values when the SCSPT is used in practice. Thus, to avoid an overly optimistic view of the agreement between P_{CLT} , P_{Exact} , and P_{True} , the computations should be based only on examinees whose tests have not yet been terminated at the given test stage. Of the termination criteria described in the previous subsection, the T_{CLT} criterion is used because it had been the primary SCSPT stopping rule in previous work (Finkelman, 2008, 2010).

Results

Part I Results

The PCC and ATL computed over all values of examinee θ for termination rules in each testing condition are displayed in Table 1 (for brevity, results were omitted for termination rules with an even number of items remaining). It is seen that the PCCs for all versions of the SCSPT are never more than 0.001, or 0.1%, away from the PCC for the SPRT. Also, all versions of the SCSPT always yield lower ATLs than the SPRT, with the largest reductions yielded with $\delta = .10$ in Conditions 1 and 2. The difference in ATL between the SCSPT and SPRT was smallest in Condition 3. For the various versions of the SCSPT, a gradual increase in ATL going from $T_{Exact_{35}}$ to T_{CLT} can be seen. This implies that the earlier in the test the SCSPT is computed using P_{Exact} , the shorter the test, on average. However, this effect is small; the largest difference in ATL among $T_{Exact_{35}}$ and T_{CLT} , which occurred in Condition 2, was only approximately 0.50.

A somewhat surprising result is revealed when the effect of changing δ for the SCSPT is examined while holding the bank constant, that is, comparing Conditions 1 with 3 and 2 with 4. It is seen that for all SCSPT termination rules, regardless of whether they are based on P_{CLT} or P_{Exact} , the ATLs are slightly larger for the higher value of δ . In contrast, Table 1 shows that increasing δ greatly reduces ATL for the SPRT, an effect that is well known. This difference in the interplay between δ and the SPRT and SCSPT can be explained by once again considering the conditional power approach used in the calculation of P_{CLT} and P_{Exact} . For example, when $\delta = .2$ and the current decision is m , the conditional power approach assumes a smaller θ value of 0.80 while estimating the probability of changing decisions. In this case, making an early

Table 1. Simulation Results: PCC and ATL Computed Over All Values of Examinee θ for All Studied Termination Rules.

Rule	Condition 1:		Condition 2:		Condition 3:		Condition 4:	
	$\delta = .1$, Peaked		$\delta = .1$, Broad		$\delta = .2$, Peaked		$\delta = .2$, Broad	
	PCC	ATL	PCC	ATL	PCC	ATL	PCC	ATL
T_{Exact_35}	0.955	21.87	0.942	22.86	0.953	22.12	0.935	24.37
T_{Exact_37}	0.955	21.96	0.942	22.99	0.953	22.19	0.935	24.45
T_{Exact_39}	0.955	22.02	0.942	23.08	0.953	22.25	0.935	24.54
T_{Exact_41}	0.955	22.10	0.942	23.20	0.953	22.33	0.935	24.62
T_{Exact_43}	0.955	22.13	0.942	23.24	0.953	22.36	0.935	24.65
T_{Exact_45}	0.955	22.17	0.942	23.27	0.953	22.39	0.935	24.68
T_{Exact_47}	0.955	22.20	0.942	23.32	0.953	22.42	0.935	24.72
T_{Exact_49}	0.954	22.23	0.942	23.37	0.953	22.44	0.936	24.75
T_{CLT}	0.954	22.25	0.942	23.39	0.953	22.45	0.936	24.76
SPRT	0.954	38.64	0.942	42.62	0.953	24.46	0.936	29.08

Note. PCC = proportion classified correctly; ATL = average test length; T_{Exact_35} = SCSPT termination rule that switches to using P_{Exact} at Stage 35, and so on; SPRT = Sequential Probability Ratio Test; SCSPT = Stochastically Curtailed Sequential Probability Ratio Test.

Table 2. Percent of Examinees With Forced Decisions and the Proportion of Those Classified Correctly for All Conditions.

Condition	T_{Exact_35}		T_{CLT}		SPRT	
	% FD	PCC of FD	% FD	PCC of FD	% FD	PCC of FD
1	1.1	0.717	2.5	0.622	42.6	0.892
2	1.9	0.547	4.2	0.588	58.4	0.901
3	1.8	0.584	2.9	0.615	19.2	0.778
4	2.3	0.504	4	0.572	28.6	0.781

Note. T_{Exact_35} = SCSPT termination rule that switches to using P_{Exact} at Stage 35; T_{CLT} = SCSPT termination rule that uses CLT approximation throughout; SPRT = Sequential Probability Ratio Test; FD = forced decisions; PCC = proportion classified correctly; SCSPT = Stochastically Curtailed Sequential Probability Ratio Test; CLT = central limit theorem.

termination is less likely than if $\delta = .1$, when a θ value of 0.90 would be used. However, this is attenuated by the fact that SCSPT also stops when the SPRT does, and the SPRT tends to stop earlier when δ is higher. It seems that the two effects nearly cancel each other out, especially for the conditions with peaked banks.

Table 2 displays information on forced decisions that were described in "SPRT" section. Specifically, for each condition, it has shown the percent of examinees that received forced decisions and the proportion of those that were classified correctly for the T_{Exact_35} , T_{CLT} , SPRT stopping rules. The SPRT had, by far, the largest percentage of forced decisions in every condition; this is not surprising, given the relatively large ATLs shown in Table 1. Looking at the SCSPT rules, in every condition, T_{CLT} had larger percent of forced decisions than T_{Exact_35} . Again, this is expected given the trends seen in Table 1. The PCC of the forced decisions for T_{Exact_35} and T_{CLT} was markedly lower than that of the SPRT. However, because forced

Figure 1. Correlations among P_{CLT} , P_{Exact} , and P_{True} for Test Stages 35 to 49.

decisions occurred at a much lower rate for these rules than for the SPRT, their overall PCC was not adversely affected.

Part 2 Results

Part 1 presented results indicating that the use of stochastic curtailment termination rules based on P_{CLT} yielded slightly longer tests than P_{Exact} . Part 2 delves more deeply into this phenomenon by examining the relationship between P_{CLT} , P_{Exact} , and P_{True} at Test Stages 35 to 49; that is, when the number of remaining items is 15, 14, and so on. Figure 1 displays the correlations between the three probabilities as the number of remaining items decreases. For all conditions, it can be seen that $COR(P_{Exact}, P_{True})$ increases as the tests progress, while $COR(P_{CLT}, P_{True})$ either levels off or drops when several items remain. This is due to the fact that the CLT does not generally perform well for very small sample sizes. Also, $COR(P_{CLT}, P_{Exact})$ decreases or levels off as the tests end; the pattern is more erratic for the broad item banks than for the peaked.

Figure 2. Bias among P_{CLT} , P_{Exact} , and P_{True} for Test Stages 35 to 49.

Figure 2 displays the MB among P_{CLT} , P_{Exact} , and P_{True} . The values of $MB(P_{CLT}, P_{True})$ and $MB(P_{Exact}, P_{True})$ are negative for all test stages studied in all four conditions, though they both approach zero as the test progresses. This indicates that both P_{CLT} and P_{Exact} consistently underestimate $P(D_k^* = D_J)$ compared with P_{True} ; thus, the use of P_{CLT} and P_{Exact} terminate the tests less aggressively than would P_{True} . Furthermore, P_{CLT} generally underestimates $P(D_k^* = D_J)$ more than P_{Exact} , especially in Conditions 1 and 3 with peaked item banks. This explains the results in Table 1 showing increasing ATLs the longer P_{CLT} is used for the computation of $P(D_k^* = D_J)$. In addition to the MB and Pearson correlation as described above, the mean absolute bias and Spearman's correlation were examined in a similar manner. Both measures yielded plots showing trends very similar to the Pearson correlations depicted in Figure 1; thus, for the sake of brevity, these plots were omitted.

Figures 3 and 4 display FPR and FNR , respectively, for P_{CLT} and P_{Exact} . In Figure 3, it is shown that the FPR are generally smaller in Conditions 1 and 3 when the item bank is peaked. However, in Condition 3, $FPR(P_{CLT})$ may be as high as 10% when there are only a couple items remaining. The pattern revealed by Figure 4 for FNR is quite clear; $FNR(P_{CLT})$ is greater

Figure 3. False positive rates for P_{CLT} and P_{Exact} .
Note. FPR = False positive rates.

than $FNR(P_{Exact})$ in all conditions. In fact, for most of the studied test stages, $FNR(P_{CLT})$ is far above 50%. This indicates that P_{CLT} does not terminate the exam in the majority of cases that P_{True} would do so. This is consistent with the MB results depicted in Figure 2 and with Table 1. Also, it is noticed that in Figure 4, there are obvious increases, or jumps, in the FNR occurring when there are 9 or 10 items remaining. To investigate these, further simulations were run with different cut points (0.25 and -1 ; the full results are not presented due to space considerations). Jumps occurred for these conditions with 7 or 8 items remaining in these additional testing conditions. The authors conclude that these jumps are randomly occurring phenomena that depend on the quality of the item bank and the CCT settings. Overall, the results suggest that these are quirks that can occur with the SCSPT, but they do not negatively affect the general performance of the method.

Discussion

The SCSPT termination criterion was proposed by Finkelman (2008) to increase the efficiency of CCTs. His results suggested that the SCSPT was capable of significantly reducing

Figure 4. False negative rates for P_{CLT} and P_{Exact} .
 Note. FNR = False negative rates.

ATLs compared with the SPRT while maintaining virtually identical classification accuracy. The crucial element of the SCSPT is the computation of $P(D_k^* = D_J)$, the probability that the interim classification decision at Stage k of the test will not change before the maximum number of items J is reached. There are two methods of calculating $P(D_k^* = D_J)$, an exact method and a CLT approximation, resulting in P_{Exact} and P_{CLT} , respectively. This article presented the results of a simulation study that aimed to provide detailed analyses concerning the relationship of P_{Exact} to P_{CLT} as the test progresses and the performance of SCSPT termination rules that switch from using P_{CLT} to P_{Exact} at various stages of the test. Although past studies have shown the SCSPT and its variations to be generally superior in efficiency to the SPRT, no previous study (to the best of the authors' knowledge) has addressed such specific comparisons between P_{CLT} and P_{Exact} .

At the outset, the authors stated that this article is intended to be of use to practitioners interested in implementing the SCSPT in an operational CCT. The results of the simulation study suggest the following concrete recommendations and guidelines as the authors refer back to the original research questions stated in the introduction: (a) Based on their analyses, the authors

cautiously suggest the switch from using P_{CLT} to P_{Exact} be made when there are between five and eight items remaining in the exam. This range should assure adequate performance from P_{CLT} and should not be too computationally burdensome for an operational setting. (b) P_{CLT} tends to overestimate the gold standard P_{True} , and P_{Exact} is less biased overall. (c) The use of P_{CLT} generally causes the SCSPT to terminate the test less aggressively than P_{Exact} . If test efficiency is of the utmost importance, then the switch from using P_{CLT} to P_{Exact} in the SCSPT should be made as early as possible. Although it is not theoretically appealing, using P_{CLT} throughout the exam in the conditions studied did not significantly decrease PCC and only very slightly increased ATL. However, unsurprisingly, P_{CLT} tended to behave somewhat erratically at later stages of the tests. And (d) the above conclusions are consistent for the four different simulation conditions examined. The reduction in ATL yielded by the SCSPT is generally more significant for smaller values of δ than for larger values. Even for conditions in which the SCSPT produces a small increase in efficiency, its use is worthwhile because there is virtually no reduction in classification accuracy.

Of course, the conditions studied in this article represent only a small subset of a vast array of potential CCT designs. One study cannot coherently address all possible variations of CCT parameters such as δ , θ_0 , test length, and so on, and item bank characteristics. Researchers and practitioners should perform similar simulations on their particular CCTs and item banks when considering the SCSPT method. Also, this article assumed a calibrated item bank with no estimation error. It is currently unknown how the performance of the SCSPT is affected by a small amount of mismatch between the true values of the item parameters and their estimated values.

Finally, there are several ways in which this research may be extended. Finkelman (2010) proposed variations on the SCSPT that terminate CCTs even more aggressively than the original SCSPT. Future research could explore whether P_{CLT} and P_{Exact} perform differently for these variations than they did in this study. Also, it has been proposed that Bayesian methodology be used in the design of a CCT (Lewis & Sheehan, 1990; Vos, 1999, 2000). In addition, a reviewer pointed out that the pass/fail decision in a CCT may be posed as two separate models. Then, test stoppage could be based on the Bayes factor of the two models exceeding a pre-determined threshold, where the Bayes factor may be calculated using a method similar to that proposed by Klugkist and Hoijtink (2007). An examination of the relative merits of SCSPT-based methods versus Bayesian frameworks would be appealing from both a theoretical and practical viewpoint.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.
- Chen, S. Y., Lei, P. W., Chen, J.-H., & Liu, T. C. (2014). General test overlap control: Improved algorithm for CAT and CCT. *Applied Psychological Measurement*, 38, 229-244.

- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261.
- Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*, 655-670.
- Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*, 442-463.
- Finkelman, M. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*, 27-45.
- Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research, & Evaluation, 17*(12). Retrieved from <http://pareonline.net/getvn.asp?v=17&n=12>
- Huebner, A., & Fina, A. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods, 47*, 549-561. doi:10.3758/s13428-014-0490-y
- Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall/CRC.
- Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis, 51*, 6367-6379.
- Lan, K. K. G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long term clinical trials. *Communications in Statistics: Sequential Analysis, 1*, 207-219.
- Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.
- Lin, C.-J. (2011). Item selection criteria with practical constraints for computerized classification testing. *Educational and Psychological Measurement, 71*, 20-36.
- Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Spray, J. A., & Reckase, M. D. (1994, April 5-7). *The selection of test items for decision making with a computerized adaptive test*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, New Orleans, LA.
- Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized adaptive test. *Journal of Educational & Behavioral Statistics, 21*, 405-414.
- Sympon, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center.
- Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793.
- Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, & Evaluation, 16*(4). Retrieved from <http://pareonline.net/pdf/v16n4.pdf>
- Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classifying respondents into multiple levels. *Applied Psychological Measurement, 38*, 187-200.
- Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24*, 271-292.
- Vos, H. J. (2000). A Bayesian procedure in the context of sequential mastery testing. *Psicológica, 21*, 191-211.
- Wald, A. (1947). *Sequential analysis*. New York, NY: John Wiley.
- Wang, C., & Chang, H. (2011). Item selection in multidimensional computerized adaptive testing—Gaining information from different angles. *Psychometrika, 76*, 363-384.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.