Graduate Management Admission Council®

Computerized Classification Testing with Composite Hypotheses

Nathan A. Thompson and Shungwon Ro
Prometric

Presented at the Item Calibration and Special Applications Paper Session, June 7, 2007

Abstract

Reckase (1983) proposed a widely used method of applying the sequential probability ratio test (SPRT; Wald, 1947) to computerized classification testing with item response theory. This method formulates the classification problem as a point hypothesis that an examinee's ability, θ, is equal to a point, θ_1, below the cutscore or a point, θ_2, above the cutscore. The current paper argues that the actual goal of classification testing is a composite hypothesis (Weitzman, 1982) that an examinee's ability θ is in the *region* of θ either above or below the cutscore, rather than equal to an arbitrarily defined point. A formulation of the SPRT to reflect this testing paradigm is proposed.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Thompson, N. A. & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Nathan A. Thompson, Prometric, 1260 Energy Lane, Saint Paul, MN 55108.
nathan.thompson@thomson.com

Computerized Classification Testing With Composite Hypotheses

The proliferation of testing in today's society is primarily manifested in industry and education, and within these realms many testing programs exist with the goal of accurately classifying examinees into categories along a continuum of achievement, ability, or trait. Efficient and accurate classification is critical for these applications because of the important decisions resulting from the classification; a classification of "pass" can lead to hiring, certification, licensure to practice a profession, or the completion of high school. A testing approach often used for the design of these high-stakes examinations is variable-length computerized classification testing (CCT; Lin & Spray, 2000). A CCT is similar in design to a computerized adaptive test, but the purpose of classification rather than point estimation of ability introduces different issues into the design.

The mathematical algorithm used to classify an examinee is referred to as the *termination criterion*. Three termination criteria have been applied to CCT: Bayesian decision theory (Vos, 1999; Rudner, 2002), ability confidence intervals (Kingsbury & Weiss, 1983; Eggen & Straetmans, 2000), and the sequential probability ratio test (SPRT: Wald, 1947; Eggen, 1999). Each has been shown to work efficiently, accurately classifying examinees with fewer items that would be required for a conventional fixed-form test.

The most commonly used and most efficient (Spray and Reckase, 1996; Eggen & Straetmans, 2000) termination criterion, the SPRT, was first applied to item response theory (IRT: Hambleton & Swaminathan, 1985) by Reckase (1983). Reckase formulated the classification problem as a point hypothesis test on the ability (θ) metric, in which an examinee's θ is equal to a point, $\theta_{1,}$ below the cutscore or a point, $\theta_{2,}$ above the cutscore, θ_c. These two points are arbitrarily defined by the test user. One drawback to the SPRT is the introduction of a certain amount of arbitrariness into the procedure by this definition, which is important because the distance between the two points directly affects the performance of the CCT. Therefore, a formulation of the classification problem with the SPRT that reduces this arbitrariness would be beneficial.

Weitzman (1982) suggested that the classification problem could also be formulated as a *composite* hypothesis, namely that $\theta \in \Theta_1$ or $\theta \in \Theta_2$ where Θ_1 represents all values of θ below the cutscore and Θ_2 all values above the cutscore. This conceptually matches the goal of CCT more closely. Weitzman proposed a method of specifying parameters for the SPRT with a composite hypothesis, but used classical test theory. Some of the issues encountered by Weitzman can be addressed by the application of item response theory (IRT) to the termination criterion.

The goal of the current study was to implement a combination of the approaches of Reckase (1983) and Weitzman (1982). Weitzman's composite hypothesis paradigm is more appropriate for CCT, but his approach is not fully realized without the advantages of IRT, as applied by Reckase. The ability confidence interval and Bayesian decision theory termination criteria can assume a composite hypothesis, but since the SPRT is the uniformly most powerful test of two competing hypotheses (Spray & Reckase, 1996), a composite formulation of an IRT-based SPRT is advantageous.

The SPRT

The SPRT compares the ratio of the likelihoods of two competing hypotheses. In CCT, the likelihoods are calculated using the probability *P* of an examinee's response if each of the hypotheses were true, that is, if the examinee were truly a "pass" (P_2) or "fail" (P_1) classification. This is expressed in general form after *n* items as, where *X* is the observed response to item *i*:

$$LR = \frac{\prod_{i=1}^{n} P_{2i}^{\ X} (1 - P_{2i})^{1-X}}{\prod_{i=1}^{n} P_{1i}^{\ X} (1 - P_{1i})^{1-X}} . \qquad (1)$$

This ratio is then compared to two decision points *A* and *B*. The full computations of *A* and *B* are complex, and Wald (1947) stated as valid approximations

$$\text{Lower decision point} = B = \beta / (1 - \alpha) \qquad (2)$$
$$\text{Upper decision point} = A = (1 - \beta)/\alpha . \qquad (3)$$

If the ratio is above the upper decision point after *n* items, the examinee is classified as above the cutscore. If the ratio is below the lower decision point, the examinee is classified as below the cutscore. If the ratio is between the decision points, another item is administered.

Formulations of the SPRT for CCT differ in the calculation of the probabilities by composing the structure of the hypotheses differently. The calculation of the ratio and the decision points remain the same.

Reckase's Method

Reckase's (1983) method first requires the cutscore to be set on the θ metric. This can be done in one of two ways. A point can be specified directly on θ, such as a cutscore of 0.0 to identify the top half of the population. The cutscore can also be translated from a cutscore previously set on the proportion-correct metric by applying a test response function and solving for the value of θ linked to the proportion-correct cutscore.

Next, two points, θ_1 and $\theta_{,2}$ must be specified on either side of the cutscore. Conceptually, this is done by defining the highest θ level that the test designer is willing to fail (θ_2) and the lowest θ level that the test designer is willing to pass (θ_1). Based on these interpretations, the area between the two is called the *indifference region*, as the test designer is indifferent to whether examinees with $\theta_1 < \theta < \theta_2$ are classified as a "pass" or "fail." In practice, however, these points are often determined by specifying an arbitrary small constant δ, then adding and subtracting it from the cutscore (e.g., Eggen, 1999; Eggen & Straetmans, 2000).

Therefore, the hypothesis test is structured as

$$H_0: \theta = \theta_1 \qquad (4)$$
$$H_1: \theta = \theta_2. \qquad (5)$$

The likelihood ratio for point hypotheses is defined as

$$LR_p = \frac{L(\theta = \theta_2)}{L(\theta = \theta_1)} = \frac{\prod_{i=1}^{n} P_i(X = 1 | \theta = \theta_2)^X \, P_i(X = 0 | \theta = \theta_2)^{1-X}}{\prod_{i=1}^{n} P_i(X = 1 | \theta = \theta_1)^X \, P_i(X = 0 | \theta = \theta_1)^{1-X}}. \qquad (6)$$

The probability of an examinee's response X to item i is calculated with an IRT item response function. An IRT model commonly applied to multiple-choice data for achievement or ability tests when examinee guessing is likely is the three-parameter logistic model (3PL). With the 3PL, the probability of an examinee with a given θ correctly responding to an item is (Hambleton & Swaminathan, 1985, Eq. 3.3):

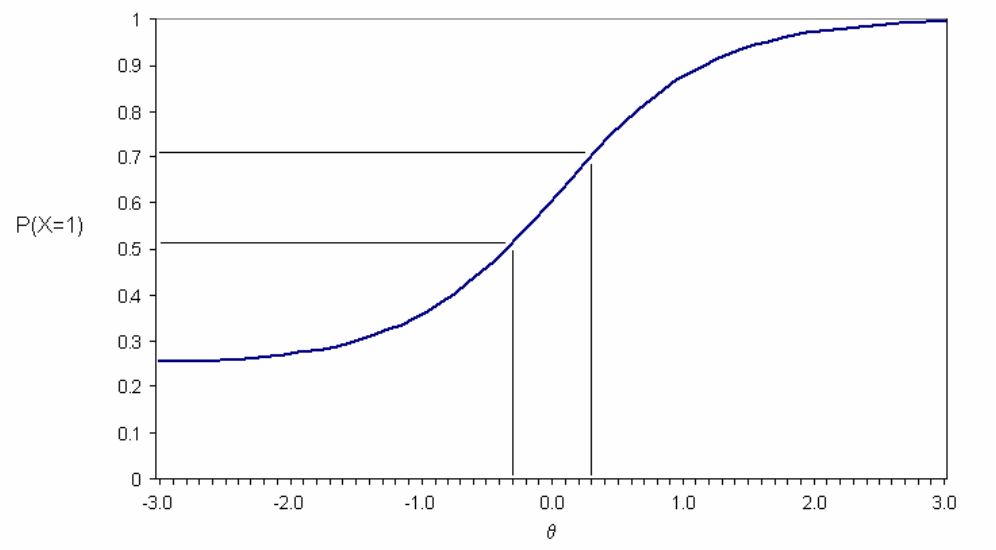
$$P_i(X_i = 1 | \theta_j) = c_i + (1 - c_i) \frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]} \qquad (7)$$

where

a_i is the item discrimination parameter,
b_i is the item difficulty or location parameter,
c_i is the lower asymptote, or pseudoguessing parameter, and
D is a scaling constant equal to 1.702 or 1.0.

A graphic representation of Reckase's method is shown in Figure 1. Let a = 1.0, b = 0.0, and c = 0.25, with θ_c = 0.0 and δ = 0.3. Then θ_1 = -0.3 and θ_2 = 0.3, with p_1 = 0.53 and p_2 = 0.72. If this was the only item presented to an examinee, the likelihood ratio would be 0.72/0.53 = 1.36 for a correct answer and 0.53/0.72 = 0.74 for an incorrect answer.

Figure 1. Reckase's SPRT Formulation



The relatively small value of δ that is illustrated produces a relatively small $P_1 - P_2$ difference. A larger value could be chosen that increases the difference, but an increase in δ entails an increase in classification error (Spray and Reckase, 1996). Moreover, a smaller value

of δ better fits the concept of an indifference region, and recent research actually employs smaller values (Eggen, 1999; Eggen & Straetmans, 2000).

Weitzman's Method

Weitzman (1982) proposed that the population of examinees be divided into *K* quantiles, with the border between two of the quantiles representing the cutscore. For example, suppose that examinees can be divided into the groups *Remedial*, *Insufficient*, *Sufficient*, and *Advanced* in terms of their knowledge of the subject matter. The purpose of the test might then be to classify examinees as sufficient (or higher) or insufficient (or lower). The group above the cutscore, in this example *Sufficient*, is denoted as *K**. The probabilities used to calculate the likelihood ratio are classical difficulty statistics for the item *i* in each group *k*, averaged across the groups above and below the cutscore. Mathematically, after *n* items this is

$$LR = \frac{\left\{(K - K^* + 1)^{-1} \sum_{k=K^*}^{K} \prod_{i=1}^{n} P_{ik}{}^{X} (1 - P_{ik})^{1-X}\right\}}{\left\{(K^* - 1)^{-1} \sum_{k=1}^{K^*-1} \prod_{i=1}^{n} P_{ik}{}^{X} (1 - P_{ik})^{1-X}\right\}} \qquad (8)$$

Because the procedure averages P_{ik} for *all* levels of ability above and below the cutscore, it is evaluating composite hypotheses, which can be expressed as:

H_0: $\theta \in$ one of groups below cutscore (9)

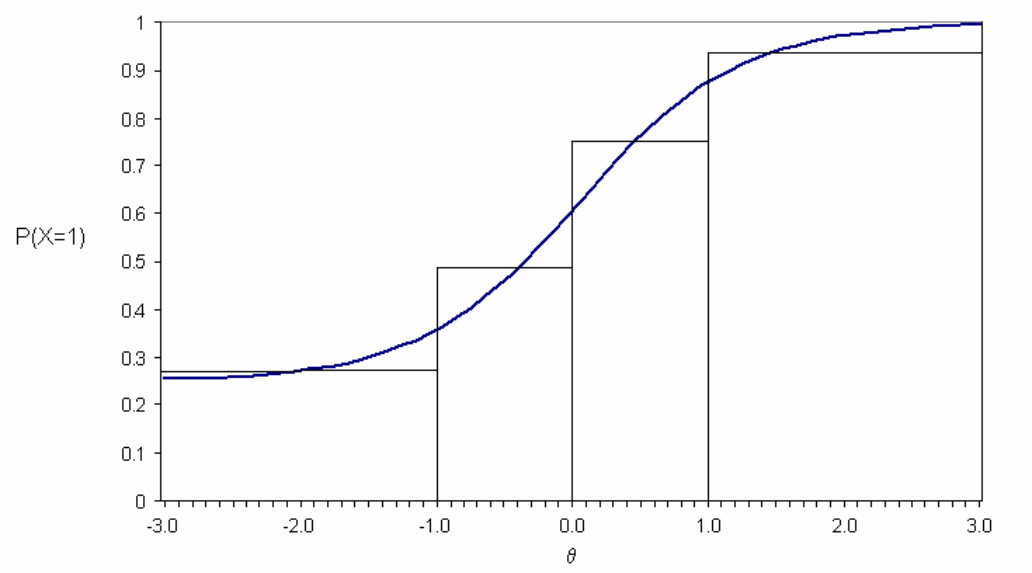
H_1: $\theta \in$ one of groups above cutscore. (10)

Note that although there are *K* quantiles, the procedure still only classifies examinees into two groups.

Figure 2 depicts an example item with Weitzman's method. This item resembles the previous example item with Reckase's method, but it is important to note that Weitzman did not employ IRT. The overlaid IRF is only for a frame of reference.

In this example, suppose that examinees have previously been determined to fall into four groups, with the groups delineated on ability by the values -1.0, 0.0, and 1.0. Weitzman's method would then sample a number of examinees in each group and determine the proportion of each group in the sample correctly responding to the item. Suppose these are the plausible values of 0.27, 0.49, 0.76, and 0.92; then the item would be graphically represented as in Figure 2. The probability of a correct response from an examinee that was *Sufficient* or above is (0.76 + 0.92)/2 = 0.84, and the probability of a correct response from an examinee that was *Insufficient* or below is (0.27 + 0.49)/2 = 0.38. The likelihood ratio after one item is then 0.84/0.38 = 2.21 for a correct response and 0.38/0.84 = 0.45 for an incorrect response.

Figure 2. Weitzman's SPRT Formulation



Composite Hypothesis with IRT

Weitzman (1982) noted that the procedure would theoretically become more accurate as the number of quantiles increased, as this would provide more detail regarding the performance of different levels of examinees on the item. However, he noted that, given a fixed calibration sample size, as the number quantiles increased the number of examinees in each quantile for the sample would decrease. This in turn would lead to greater error in the estimation of the classical difficulty statistics for each quantile, offsetting the increase in information provided by the use of more quantiles.

IRT offers a solution to this issue. Weitzman's (1982) procedure can be considered an empirical item response function because it analyzes the probability of a correct response for successively higher levels of ability. Rather than increase the number of quantiles to obtain a more detailed estimate of the probability of a correct response across different levels of ability, a more detailed estimate can be obtained by estimating the theoretical item response function, as in Equation 7. The same goal can be achieved by calculating the average probability of a correct response for examinees above and below the cutscore using an average Riemann sum procedure (Ostebee & Zorn, 2001). In fact, Weitzman's approach can be interpreted as a very approximate unweighted Riemann procedure.

With IRT, this formulation can be expressed in terms of a hypothesis test:

$$H_0: \theta \in \Theta_1 \quad (11)$$

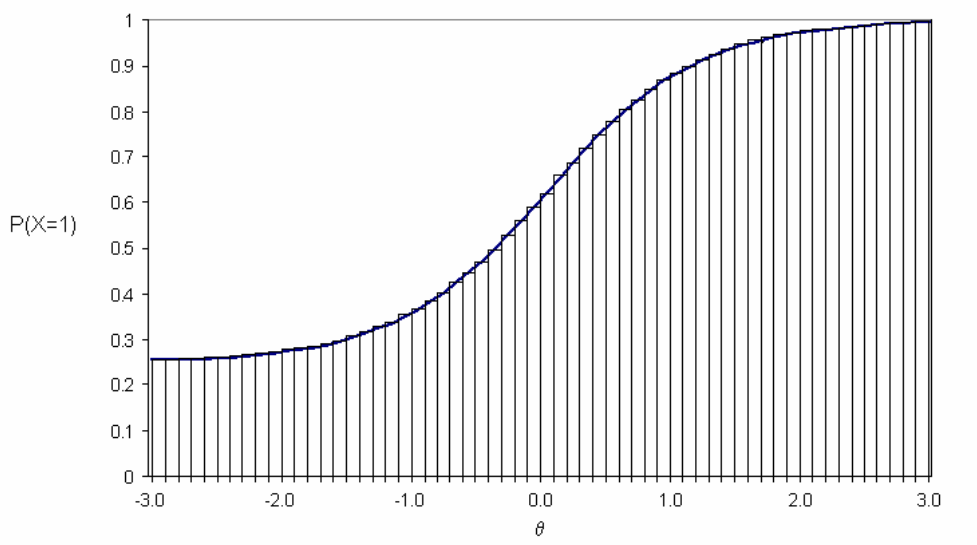
$$H_1: \theta \in \Theta_2. \quad (12)$$

where Θ_1 represents the range of θ below the cutscore, and Θ_2 represents the range of θ above the cutscore.

Consider the example item with $a = 1.0$, $b = 0.0$, and $c = 0.25$, and with $\theta_c = 0.0$. The mean probability of a correct response for examinees below the cutscore is approximately 0.35, evaluated from -3.0 to -0.1 in intervals of 0.1. The mean probability of a correct response for

examinees above the cutscore is approximately 0.90, calculated similarly. This is represented graphically in Figure 3. The likelihood ratios in this case would be 0.90/0.35 = 2.57 and 0.35/0.90 = 0.39. This produces a greater difference in the ratio of the two values than the previously discussed methods, although a large value of δ could be selected to increase the ratio for the point SPRT.

Figure 3. Composite IRT SPRT Formulation



When considered across two or more items, this approach conveniently translates to the commonly used IRT likelihood function for a response vector for *n* items,

$$L(X_{1j}, X_{2j}, ..., X_{nj} \mid \theta_j) = \prod_{i=1}^{n} P_i(\theta_j)^{X_{ij}} (P_i(\theta_j) - 1)^{1-X_{ij}} \qquad (13)$$

Rather than calculate a Riemann integration for each item, the likelihood function is integrated on either side of the cutscore. The ratio of the values, the likelihood ratio for composite hypotheses,

$$LR_c = \frac{L(\theta \in \Theta_2)}{L(\theta \in \Theta_1)}, \qquad (14)$$

is evaluated. Because the values of the likelihood function in the extreme regions of θ are very small, the integration can be done for a ±*k* interval symmetrical about the cutscore for some constant *k*, such as ±3. Note that this must be symmetrical about the cutscore, as an interval symmetrical about another point would be biased in the same direction as that point is in relation to the cutscore.

Theoretically, this ratio is a continuous integration,

$$LR_c = \frac{L(\theta \in \Theta_2)}{L(\theta \in \Theta_1)} = \frac{\int_{\theta_c}^{\theta_c+k} L(u \mid \theta)}{\int_{\theta_c-k}^{\theta_c} L(u \mid \theta)} \tag{15}$$

but can be calculated with a Riemann integral. With the midpoint method, this is expressed as

$$LR_c = \frac{L(\theta \in \Theta_2)}{L(\theta \in \Theta_1)} = \frac{\sum_{\theta_c}^{\theta_c+k} L(u \mid \theta + 0.5\Delta\theta)\Delta\theta}{\sum_{\theta_c-k}^{\theta_c} L(u \mid \theta + 0.5\Delta\theta)\Delta\theta} \tag{16}$$

for a Riemann interval width of $\Delta\theta$.

After each item or testlet in a test, the composite likelihood ratio LR_c can then be compared to bounds A and B, as is done for LR_p. If LR_c exceeds the bounds, a decision is made. If not, another item or testlet is administered.

Thus, the composite likelihood ratio (CLR) offers a conceptually attractive and appropriate alternative to the point-hypothesis SPRT. However, while the CLR led to the greatest $P_1 - P_2$ difference in the single-item example above, which would lead to an examinee classification with fewer items, this is not necessarily generalizable. Because δ for the point-hypothesis SPRT is a parameter chosen by the test designer, it could theoretically be specified at a relatively large value, leading to relatively large $P_1 - P_2$ differences. Moreover, specification of δ involves tradeoffs, where large values lead to decreased test length but increased classification error, and vice versa. To better compare the point and composite versions of the SPRT, a monte carlo simulation was used.

Method

The item bank parameters used for this study assumed that items were dichotomously scored and could be calibrated with the 3PL. The bank contained 750 items. The scaling constant *D* was specified as 1.702. The *c* parameters were randomly generated from a N(0.25,0.03) distribution, the *b* parameters were randomly generated from a N(0,2) distribution with the constraint $-3 < b < 3$, and the *a* parameters were randomly generated from a N(0.7,0.2) distribution.. A sample of 10,000 examinees was generated from a N(0,1) distribution. The cutscore was fixed at 0.5, to represent a testing situation where the purpose is to identify a set of highly performing or knowledgeable examinees. Maximum test length was specified as 200 items.

A classification was determined for each examinee using monte carlo simulation. A response was randomly generated for an item, after which the termination criterion was evaluated. If the termination criterion was not satisfied, another item was administered based on specific item selection criteria. CCT conditions were evaluated and compared on two dependent variables commonly used in CCT research: the average test length (ATL) across the sample, and the percentage of correct classifications (PCC) made across the sample.

The primary independent variable was the termination criterion. Three criteria were investigated: the point SPRT, the CLR, and ability confidence intervals (ACI: Kingsbury & Weiss, 1983; Thompson, 2007). ACI is an alternative method of using the likelihood function to make a classification decision. However, rather than considering the entire likelihood function, it constructs a confidence interval around the maximum likelihood (or Bayesian) estimate of ability, θ. This can be expressed as (Thompson, 2007):

$$\hat{\theta}_j - z_\varepsilon(CSEM) \le \theta_j \le \hat{\theta}_j + z_\varepsilon(CSEM) \tag{17}$$

where z_ε is the normal deviate corresponding to a 1-ε confidence interval, given $\varepsilon = \alpha + \beta$ for nominal error rates α and β.

Because the value of δ affects the results of the SPRT, it was varied to provide a better opportunity for comparison. Five values were used: 0.30, 0.35, 0.40, 0.45, and 0.50. This range was selected because it produced error rates near the nominal levels, which were specified as $\alpha = \beta = 0.025$ for all termination criteria.

The CLR parameter *k* was specified as 6.0, so that the likelihood function was integrated from -5.5 to 0.5, and 0.5 to 6.5. This implicitly assumes that values of the likelihood function outside the interval are low enough or symmetrical enough that they would not affect the ratio. The Riemann interval width Δθ was specified as 0.01.

During the course of the simulations, it was noticed that the CLR had a difficult time making classifications after 10 or 20 items had been administered. At this point, the likelihood function had extremely low values and subsequently administered items had little effect. ATL was greater than 50 items. Therefore, the bounds *A* and *B* were modified so that after a large number of items, the CLR did need to attain a very large or very small value to make a decision. Because the modification needed was greater after a larger number of items *n*, the bounds were multiplied by the inverse of the square root of *n* and a constant γ:

$$\text{Lower decision point} = B = \frac{\beta}{1-\alpha} \times \frac{1}{\gamma\sqrt{n}} \tag{18}$$

$$\text{Upper decision point} = A = \frac{1-\beta}{\alpha} \times \frac{1}{\gamma\sqrt{n}}. \tag{19}$$

The constant γ is analogous to the constant δ for the SPRT, where larger values will cause the criterion to be satisfied after fewer items, all other things equal. Two levels of γ were investigated: 0.8 and 1.0.

Two item selection methods were crossed with the CLR termination criteria. Both utilized Fisher information, which with the 3PL is (Embretson & Reise, 2000, Eq. 7 A.2)

$$I_i(\theta) = \left[a_i^{\,2} \frac{1-P_i(\theta)}{P_i(\theta)} \right] \left[\frac{(P_i(\theta)-c_i)^2}{(1-c_i)^2} \right]. \tag{20}$$

Cutscore-based (CB) item selection selects the next item to maximize information at the cutscore. Estimate-based (EB) item selection selects the next item to maximize information at the current θ estimate. Thompson (2007b) showed that CB selection is more appropriate for the SPRT, while EB selection is more appropriate for ACI, so these termination criteria only utilized one selection method..

Results

Simulation results are presented in Table 1. The CLR performed comparably to the point SPRT, with the SPRT having slightly higher PCC. For example with $\gamma = 1.0$ and CB item selection, the CLR had an ATL of 20.92 and PCC of 93.48. The point SPRT with $\delta = 0.5$ had an ATL of 21.69 and the PCC ATL was 93.91.

Table 1. ATL and PCC by Item Selection Method and Termination Criterion

Method and Termination Criterion	ATL	PCC
Point SPRT		
$\delta = 0.2$	85.75	95.60
$\delta = 0.3$	42.81	95.50
$\delta = 0.4$	29.93	94.61
$\delta = 0.5$	21.69	93.91
CLR: CB Selection		
$\gamma = 0.8$	24.56	93.73
$\gamma = 1.0$	20.92	93.48
CLR: EB Selection		
$\gamma = 0.8$	28.09	94.01
$\gamma = 1.0$	23.55	93.44
ACI Confidence Interval		
95%	38.88	93.63
99%	59.59	94.84

Decreasing γ had the same effect as decreasing δ: more items were needed to make a decision, while the accuracy of the simulation also increased. ACI also exhibited this effect; ATL was higher than both the CLR and point SPRT, but it was also more accurate. However, this level of accuracy was still below nominal levels. The maximum test length caused many examinees with θ values near the cutscore to have tests concluded before the termination criterion could be met.

Discussion and Conclusions

The purpose of this paper was to argue that, at a paradigm level, composite hypotheses are more appropriate for CCT. As noted earlier, a composite hypothesis is assumed with the other

two termination criteria available for CCT. Ability confidence intervals implicitly make this assumption, as they are designed to evaluate whether a confidence interval of an examinee's θ level is completely within the region above or below the cutscore. Bayesian decision theory evaluates loss or utility structures with regard to whether the examinee is truly a "pass" or "fail." The point hypothesis formulation currently in use with the SPRT termination criterion is not incorrect, but can be viewed as an approximation to the composite hypothesis formulation that reflects the purpose of classification testing and is currently in use with other termination criteria.

The requirement of an additional test parameter for the point SPRT and arbitrariness introduced by it can be seen as a drawback when compared to the ability confidence interval termination criterion. One reason for the development of a composite likelihood ratio approach is that the use of a composite paradigm theoretically eliminates the need for an arbitrary specification of δ. However, this was not achieved in the current study because of the application of γ to the CLR. Future research should explore an analytical solution to the issue addressed by γ, namely, that it becomes difficult for the CLR to make a decision after a substantial number of items has been administered.

The tradeoff between test length and classification accuracy introduced by the specification of δ has been an issue in SPRT research, causing researchers to investigate various values and then being faced with a decision of which set of results to interpret (Eggen, 1999; Eggen & Straetmans, 2000). Nevertheless, the use of δ and γ can alternatively be seen as an additional layer of flexibility in the procedure. They can be adjusted in small increments to obtain observed PCC very close to the nominal PCC. ACI does not have such a feature.

The CLR approach also has a relationship to ACI: conceptually, they are equivalent if the likelihood function is perfectly symmetrical. A 95% confidence interval will make a decision when the θ estimate is 1.96 standard errors above or below the cutscore, which is comparable to when approximately 97.5% of the likelihood function falls on either side of the cutscore. At that juncture, the CLR would also be approximately 0.975/0.25 = 39.00 which is also the calculation of *A*, so the CLR would make a decision at that same point in time. ACI and CLR should differ with respect to the asymmetry of the likelihood function.

It is likely due to this conceptual similarity that CLR initially had a high ATL and PCC, similar to ACI, before *A* and *B* were adjusted with γ. The development of an *n*-related adjustment is also plausible for ACI, which would introduce some flexibility with respect to PCC..

In conclusion, the composite hypothesis model that is acknowledged in other CCT termination criteria can also be applied to the SPRT without any reduction in efficiency. However, this approach needs to be explored further, as well its relationships with the point SPRT and ACI.

References

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 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,* 713-734.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications.* Norwell, MA: Kluwer Academic Publishers.

Kingsbury, G.G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Lin, C.-J. & Spray, J.A. (2000). Effects of item-selection criteria on classification testing with the sequential probability ratio test. (Research Report 2000-8). Iowa City, IA: ACT, Inc.

Ostebee, A., & Zorn, P. (2001). *Calculus from graphical, numerical, and symbolic points of view* (2nd Ed.). Boston: Houghton Mifflin.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), New horizons *in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Rudner, L. M. (2002). An examination of decision-theory adaptive testing procedures. Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, 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 test. *Journal of Educational & Behavioral Statistics, 21,* 405-414.

Vos, Hans J. (1999). Applications of Bayesian Decision Theory to Sequential Mastery Testing. *Journal of Educational and Behavioral Statistics, 24*(3), 271-92.

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

Weitzman, R. A. (1982). Sequential testing for selection. *Applied Psychological Measurement, 6,* 337-351.