Likelihood Ratio Based Computerized Classification Testing

Nathan A. Thompson

Assessment Systems Corporation & University of Cincinnati

Shungwon Ro

Kenexa

Abstract

An efficient method for making decisions is to evaluate the ratio of likelihoods for competing hypotheses, such as "Pass" or "Fail." It is typically used in the context of only two categories and item response theory, but this approach is easily extended to multiple categories (Eggen, 1999) and classical test theory (Rudner, 2002). This paper will provide a review of this approach as well as a comparison of the most current methods, most notably the comparison of a point hypothesis structure to a composite hypothesis structure. It will also present current issues facing the method and recommendations for further research.

Termination Criteria for Computerized Classification Testing

In educational assessment, a common purpose of a test is to classify examinees into mutually exclusive groups rather than obtain accurate estimates of individual scores. This is often termed *mastery testing* when the test is designed to determine if a student has mastered material by classifying them as "pass" or "fail." There are several methods of calculating this decision, the most obvious of which is utilizing the observed number-correct score on a traditional fixed-form test. However, more sophisticated methods have been suggested in which the computerized test delivery mechanism is designed to be intelligent and adapt both the number and nature of the items in the test to each examinee as they proceed sequentially through the test. The variable-length statistical mechanism that decides when to stop the test and classify the examinee is known as the termination criterion or stopping rule (Kingsbury & Weiss, 1983).

These computerized classification tests (CCTs; Parshall, Spray, Kalohn, & Davey, 2002) can be based on item response theory (IRT; Hambleton & Swaminathan, 1985) or classical test theory (Rudner, 2002). While the classical approach can be quite efficient (Frick, 1992), this paper will focus on the utilization of IRT. When based on IRT, the classification decision is made with two paradigms: likelihood ratios (e.g., Reckase, 1983) and confidence intervals (e.g., Kingsbury & Weiss, 1983). Both utilize the likelihood function of examinee ability, and have been termed *statistical classification* and *statistical estimation* by Eggen (1999).

The likelihood ratio was originally formulated as a point hypothesis sequential probability ratio test (SPRT) by Reckase (1983). The SPRT operates by testing that a given examinee's ability value θ is equal to a fixed value below (θ_1) or above (θ_2) the classification cutscore. The space between these two points is referred to as the indifference region, as the test developer is nominally indifferent to the classification assigned. The SPRT has been shown to be more efficient than confidence intervals around ability estimates as a method for CCT delivery (Spray & Reckase, 1996; Eggen & Straetmans, 2000).

However, Weitzman (1982) suggested that the classification problem could also be formulated as a *composite* hypothesis, namely that a given examinee's θ is *below* θ_1 or *above* θ_2. This conceptually matches the goal of CCT more closely, which is to test whether θ is above or below the cutscore. Weitzman proposed a method of specifying parameters for the likelihood ratio with a composite hypothesis, but used classical test theory as an approximation of IRT. Some of the issues encountered by Weitzman can be addressed by the application of item response theory directly to the termination criterion as a composite hypothesis. However, this line of research was not continued until recently.

Bartroff, Finkelman, and Lai (2008) and Thompson (2009a) independently suggested using a generalized likelihood ratio (GLR: Huang, 2004) based on the IRT likelihood function. The purpose of this paper is to explore the application of the GLR to CCT with two monte carlo simulation studies. The first study provides a comparison of the GLR to other methods of CCT. The second study explores the difference between the GLR and the SPRT, and the role of the indifference region.

Termination criteria

The likelihood ratio 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 to item *i* if each of the hypotheses were true, that is, if the examinee were truly a "pass" (P_2) or "fail" (P_1) classification. With IRT, the probability of an examinee's response *X* to item *i* is calculated with an 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):

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

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.

The likelihood ratio is expressed as the ratio of the likelihood of a response at two points on θ, θ_1 and θ_2,

$$LR = \frac{L(\theta = \theta_2)}{L(\theta = \theta_1)} = \frac{\prod_{i=1}^{n} P_i(X = 1 | \theta = \theta_2)^X P_i(X = 0 | \theta = \theta_2)^{1-X}}{\prod_{i=1}^{n} P_i(X = 1 | \theta = \theta_1)^X P_i(X = 0 | \theta = \theta_1)^{1-X}}. \qquad (2)$$

Note that, since the probabilities are multiplied, this is equivalent to the ratio of the value of the IRT likelihood function at two points. A value greater than 1.0 indicates a higher likelihood of the examinee being a "pass" classification. The ratio is then compared to two decision points *A* and *B*, (Wald, 1947):

$$\text{Lower decision point: } B \geq \frac{\beta}{1-\alpha} \qquad (3)$$

$$\text{Upper decision point: } A \leq \frac{1-\beta}{\alpha}. \qquad (4)$$

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. Note that the decision points do not need to be specified directly. Instead, the nominal error levels α and β are specified to reflect the need of the testing program, and used to calculate *A* and *B*. As a practical example, setting both to 0.025 would indicate 95% accuracy, and translate to $A = 39.0$ and $B = 0.026$.

Formulations of the likelihood ratio 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. The point hypothesis SPRT calculates P_1 and P_2 at fixed points selected by the test developer, while the composite hypothesis method calculates at variable points, wherever the likelihood function is the highest.

Because IRT is utilized, this 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 characteristic curve and solving for the value of θ linked to the proportion-correct cutscore (Parshall, Spray, Kalohn, & Davey, 2002).

Point hypothesis formulation

The point hypothesis method, known as the SPRT, suggested by Reckase (1983) specifies two *fixed* points θ_1 and θ_2 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), hence the term *indifference region* for this range. 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).

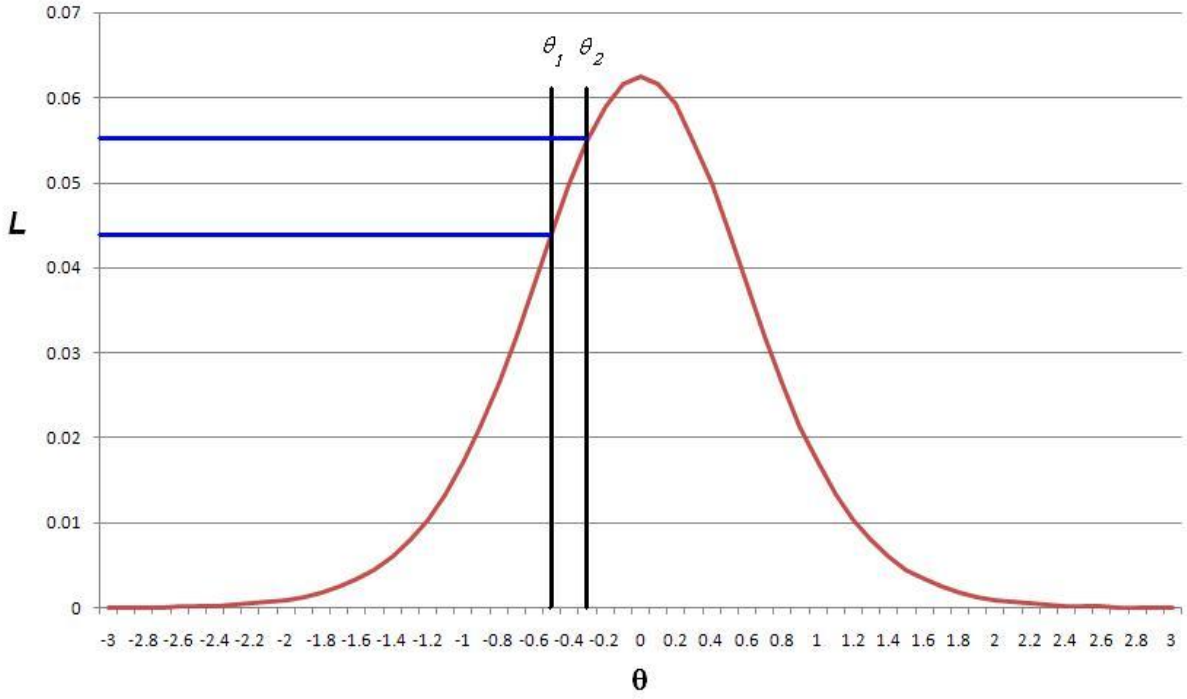
Therefore, the hypothesis test is structured as

H_0: $\theta = \theta_1$ (5)

H_1: $\theta = \theta_2$. (6)

A graphic representation of this method is shown in Figure 1. In this example, the cutscore is -0.4 and δ= 0.1, such that θ_1 = -0.3 and θ_2= -0.5. The likelihood function is evaluated at these two points, producing a ratio of approximately 0.055/0.044 = 1.25. The likelihood that the examinee is a "pass" is greater than the likelihood they are a "fail," but the classification cannot be made with much confidence at this point in the test.

Figure 1: Example likelihood function and indifference region



This is partially due to the relatively small value of δ that is illustrated, which produces a relatively small $P_2 - P_1$ difference. It is evident from Figure 1 that increasing the space between θ_1 and θ_2 would increase this difference and therefore the likelihood ratio. The generalized likelihood ratio (GLR) is designed to take advantage of this.

Eggen (1999) and Eggen and Straetmans (2000) utilize the point hypothesis approach for classification into more than two categories. The likelihood ratio is evaluated at each cutscore to make decisions. For example, if there are two cutscores defining three categories, failing at both cutscores would translate to a classification in the lowest group. Passing both cutscores translates into classification as the highest group, while passing the lower and failing the upper is classified as the middle group.

Composite hypothesis: the generalized likelihood ratio

The GLR is specified and calculated with the same methods as the fixed-point SPRT, with the exception that θ_1 and θ_2 are allowed to vary. Rather than evaluate the likelihood function at each endpoint of the indifference region, instead it is evaluated at the highest points beyond the endpoints. If the maximum of the likelihood function is outside the indifference region, that maximum will be utilized in the likelihood ratio for that side. For example, in Figure 1 the maximum is to the right of the

indifference region, at 0.0, and will be utilized in the likelihood ratio. The side without the maximum is evaluated the same as with the SPRT, as the highest likelihood to the left of θ_1 is at θ_1.

In the example of Figure 1, this modification to the likelihood ratio now produces a value of 0.062/0.044 = 1.41. Because this ratio is further from a ratio of 1.0 than the fixed SPRT value of 1.25, the classification can be made with more confidence given the same number of items, or with equal confidence given a fewer number of items. The primary research question of this paper is whether this increase in efficiency comes with an increase in classification error (false positives and false negatives) as compared to other methods of pass/fail decisions, and if the efficiency is moderated by the width of the indifference region.

Ability confidence intervals

Ability confidence intervals (ACI) is an alternative method of using the likelihood function to make a classification decision. However, rather than considering the entire likelihood function, it makes a confidence interval around the maximum likelihood (or Bayesian) estimate of ability using the conditional standard error of measurement (SEM). This can be expressed as (Thompson, 2009b; Hambleton & Swaminathan, 1985, Eq. 5.28):

$$\hat{\theta}_j - z_\varepsilon(SEM) \le \theta_j \le \hat{\theta}_j + z_\varepsilon(SEM) \qquad (7)$$

where z_ε is the normal deviate corresponding to a 1 - ε confidence interval, given α + β = ε for nominal error rates α and β. For example, a 95% confidence interval entails $z_\varepsilon = 1.96$, with α =0.025, β = 0.025, and ε = 0.05. While the SPRT and GLR differentiate examinees only at the cutscore, ACI evaluates across the spectrum of θ, wherever the current estimate lies. Therefore, previous research (Spray & Reckase, 1996; Eggen & Straetmans, 2000; Thompson, 2009b) has shown that ACI operates more efficiently when items are selected adaptively at the current estimate, while the SPRT and GLR operate more efficiently when items are selected to maximize information at the cutscore.

For this study, the confidence intervals were calculated with two methods: theoretical and observed. For the theoretical approach, model-predicted SEM is calculated using the test information function evaluated at the relevant θ regardless of response pattern (Embretson & Reise, 2000, Eq. 7A.6),

$$SEM = 1/\sqrt{TI(\theta)} \qquad (8)$$

and θ is estimated using brute force methods by directly evaluating the likelihood function from -3.0 to +3.0 in intervals of 0.01 to find the empirical maximum. In practice, it is more common to estimate θ with efficient Newton-Raphson methods (Embretson & Reise, 2000, p. 164), and calculate an observed SEM based on the second derivative of the likelihood function (Baker & Kim, 2004, Eq. 3.16):

$$SEM = \sqrt{\frac{1}{-E(\partial^2 L / \partial \theta_j{}^2)}}\,. \qquad (9)$$

Study 1

The study utilized a monte carlo simulation methodology, with 10,000 examinees simulated under each testing condition, to evaluate differences in efficiency and accuracy. The population of examinees was randomly selected from a N(0,1) distribution. With monte carlo simulation, item responses are generated by comparing a randomly generated number $0.0 < r < 1.0$ to the probability of a correct response for each examinee to each item. The probability is calculated using the item response function and the true examinee θ, which is known because it was generated. For example, if there is a 0.70 probability of a correct response, an $r = 0.65$ would produce a response of "correct' and an $r = 0.75$

would produce a response of "incorrect." Responses are generated as each item is administered to the examinee in the simulation.

The independent variable was the design of the test. The three primary levels investigated were the ACI, SPRT, and GLR variable-length termination criteria. Fixed-form tests of 200 items, 100 items, and 50 items, with both number-correct and IRT maximum likelihood scoring, were included as a baseline. The fixed forms were constructed by selecting items from the bank of 500 with the most information at the cutscore, producing tests with a highest possible level of differentiating capability. The dependent variables are average test length (ATL), and percentage of correct classifications (PCC). If a test is performing well, it will produce high PCC but low ATL, namely accurate decisions with only a few items.

Because the value of δ affects the results of the SPRT and GLR, it must be manipulated to provide an opportunity for adequate comparison. Namely, a wide range of values was not arbitrarily selected, but methods were rather matched on observed PCC. The ACI simulations were completed first with a 95% confidence interval, and then the SPRT and GLR simulations completed with δ varied until a similar PCC (95.7) was reached, which was 0.3. Simulations were also completed with δ= 0.2 for an additional comparison.

The cutscore for the simulations was θ = -0.5, which corresponds to a pass rate of approximately 69%, representing a mastery test where the majority of students typically pass. For the fixed-form tests with number-correct scoring, this was converted to a raw cutscore using the test response function (Parshall, Spray, Kalohn, & Davey, 2002): 122.5 for the 200-item test, 63.85 for the 100-item test, and 32.49 for the 50-item test. The variable-length tests were constrained to have a minimum of 20 items and a maximum of 200 items. A maximum is necessary to prevent the entire bank from being administered to examinees with true ability at the cutscore, because a decision would never be able to be made with confidence. A minimum is not psychometrically necessary, but has a public relations function in that it protects against examinees failing after only a few items, possibly reducing complaints.

The bank for the test consisted of 500 items with IRT parameters to represent plausible values for a test designed to differentiate at a cutscore of -0.50. The difficulty of the bank was centered on the cutscore, and the discrimination values were generated with a target mean of 0.70, which is typical for achievement tests. The guessing parameter *c* was generated to have a mean of 0.25, representing 4-option multiple choice items. The summary statistics for the generated parameters are presented in Table 1.

Table 1: Summary statistics of item bank

Statistic	*a*	*b*	*c*
Mean	0.716	-0.480	0.251
SD	0.204	0.552	0.041

The results of the simulations are presented in Table 2. ATL refers to the average number of items seen by each examinee; for the fixed-form tests, this is of course equal to the test length. PCC is the percentage of examinees correctly classified, comparing the results of the test to the generated person parameter θ. Type I errors are examinees that passed but should have failed, having a true generated θ below the cutscore, and Type II failed but should have passed.

As hypothesized, the variable-length methods produced short tests, with ATL ranging from 37.62 to 55.77, while maintaining the level of accuracy produced by the longer fixed form tests that delivered two to four times as many items. Specifically, the two likelihood ratio methods with δ = 0.2 had PCC approximately equivalent to the 200-item fixed test, but with only 48.41 and 55.77 items. The 50-item fixed test entailed approximately as many items as the variable-length methods, but with notably decreased accuracy.

Also notable are the differences between the variable-length methods. The SPRT and GLR produced shorter tests than ACI while maintaining accuracy. The two ACI conditions required more than

50 items, with the intentionally matched PCC of approximately 95.75. The SPRT and GLR produced equivalent accuracy with less than 40 items.

Table 2: Average test length (ATL) and percent correctly classified (PCC) for each condition

Test design	Scoring	ATL	PCC	Type I	Type II
200 item fixed	Number-correct	200.00	96.10	1.81	2.09
200 item fixed	IRT	200.00	96.19	2.07	1.74
100 item fixed	Number-correct	100.00	95.19	2.56	2.25
100 item fixed	IRT	100.00	95.13	2.62	2.25
50 item fixed	Number-correct	50.00	93.62	3.60	2.78
50 item fixed	IRT	50.00	93.46	3.24	3.30
Ability confidence intervals (ACI)	Theoretical SEM	51.65	95.73	2.57	1.70
Ability confidence intervals (ACI)	Observed SEM	54.61	95.78	2.51	1.71
Sequential probability ratio test (SPRT)	δ= 0.3	39.30	95.74	1.85	2.41
Generalized likelihood ratio (GLR)	δ= 0.3	37.62	95.73	2.03	2.24
Sequential probability ratio test (SPRT)	δ= 0.2	55.77	96.21	1.81	1.98
Generalized likelihood ratio (GLR)	δ= 0.2	48.41	96.06	2.01	1.93

There was a small but recognizable difference between the two methods of calculating ACI. The two produced equivalent PCC, but utilizing the observed SEM and the more computationally efficient Newton-Raphson θ estimation required three more items, on average, than the model-predicted SEM calculated with the test information function and the empirical likelihood estimate.

The GLR was slightly more efficient than the SPRT; this gain in efficiency increases with a decrease in δ because a wide δ forces the GLR and SPRT to utilize the same calculations. For example, the GLR produced a larger ratio with Figure 1, but if the indifference region in Figure 1 was -0.8 to 0.0, then the GLR and the SPRT would be equivalent. Therefore, the GLR utilized only two fewer items on average with the wider indifference region (δ= 0.3), but there was a difference of seven items when δ = 0.2.

This study demonstrates that the GLR performs as expected, namely highly similar to the point hypothesis SPRT, but with several fewer items, indicating an increase in efficiency. This increase is greater with a narrower indifference region. Study 2 will examine this effect further, focusing on the GLR and SPRT. Both are much more efficient than fixed-form tests.

Study 2

A similar monte carlo simulation was designed to further compare the GLR with the SPRT, while investigating the effect of indifference region width on efficiency by simultaneously comparing the observed classification error rates to the nominal rate. Parameters were generated for a bank of 300 items; fewer items were necessary because 200-item tests were not being created. The descriptive statistics of the item parameters are shown in Table 1, and reflect the fact that the bank was again intended to provide a substantial number of items with difficulty near the cutscore of -0.50. A distribution of examinees was also randomly generated, from a N(0,1) distribution. PCC and ATL were again the dependent variables, with the practical test length constraints of a minimum of 20 and a maximum of 200.

Table 3: Item parameter statistics

Statistic	*a*	*b*	*c*
Mean	0.70	-0.50	0.25
SD	0.20	0.51	0.04

Besides the comparison of the two termination criteria, the width of the indifference region was an independent variable, manipulated by varying δ from 0.0 to 1.0 in increments of 0.1. The results are presented in Figure 2 for a nominal error rate of 1% and in Figure 3 for a nominal error rate of 5%. Note that while a GLR with δ = 0.0 is possible because it will search for values away from the cutscore, an SPRT with δ = 0.0 is impossible because the ratio is always 1.0 (both values exactly at the cutscore).

Figure 2: ATL and PCC for 1% nominal error rate, comparing GLR and SPRT

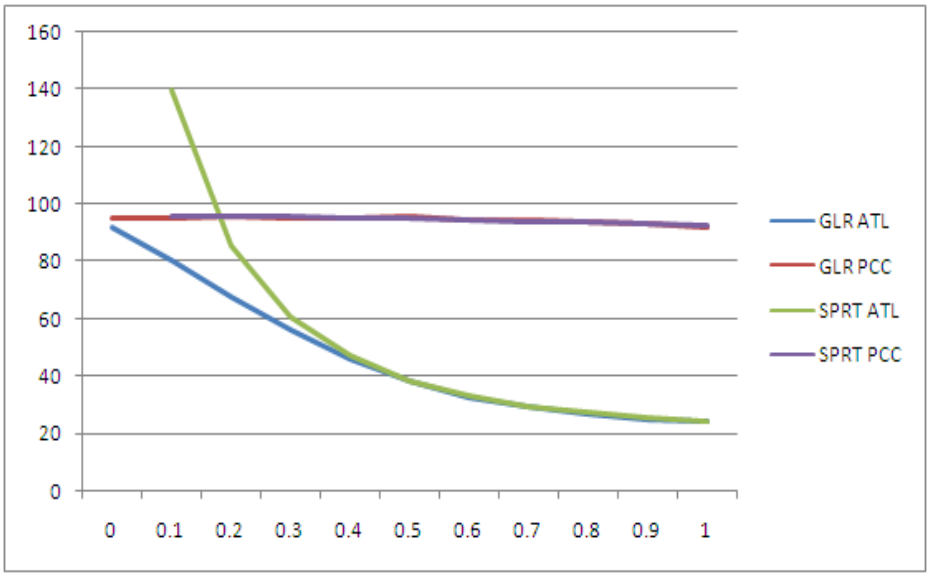
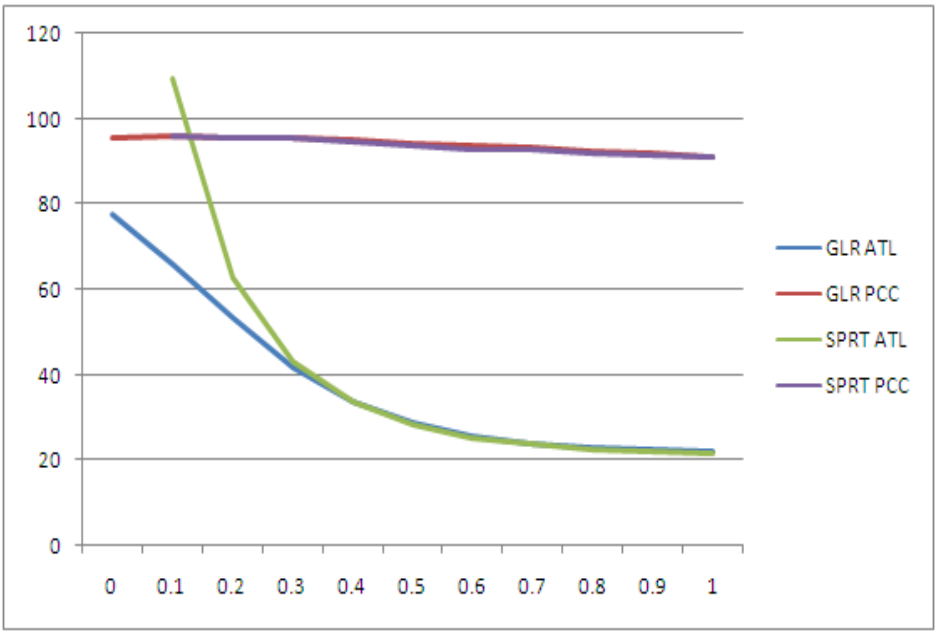


Figure 3: ATL and PCC for 5% nominal error rate, comparing GLR and SPRT



With regards to the termination criteria, the GLR requires fewer items when δ is 0.3 or smaller, while the two methods perform equivalently with larger values of δ. The detailed values for δ = 0.2 are presented in Table 4 to show this effect; here, the GLR required substantially fewer items (ATL) while maintaining accuracy (PCC). This is concordant with the results of the first simulation study, as large values of δ force the same values to be selected for input into the likelihood ratio.

Table 4: Results with δ = 0.2

Termination	Nominal	δ	ATL	PCC
GLR	99	0.2	67.90	96.02
GLR	95	0.2	53.61	95.32
SPRT	99	0.2	85.55	95.68
SPRT	95	0.2	62.62	95.65

However, note that PCC appears stable when δ is 0.3 or smaller, it decreases beyond 0.3. Moreover, ATL dropped dramatically with smaller values of δ, but successively larger values decreased ATL only slightly. Increasing the size of the indifference region will greatly decrease the number of items needed to make classifications, but will also marginally decrease the accuracy of the test, and this effect depends on the range of δ in question. It is therefore imperative that testing programs which employ the likelihood ratio as a termination criterion perform extensive simulation research to ensure that the drop in ATL is maximized while maintaining PCC at nominal levels. Not doing so could lead to tests longer than necessary, or accuracy less than nominal levels.

The maintenance of PCC near nominal levels is itself a substantial issue. In Figure 3, the accuracy remained near nominal levels for the 5% condition while δ < 0.3. However, for the 1% condition, observed accuracy was *always* lower than the nominal accuracy. In fact, the highest observed PCC in Figure 2 was only 96.02 (in Table 4), well short of the nominal 99%. Furthermore, as δ increased, the observed PCC dropped to approximately 92%. This extreme disconnect between observed and nominal accuracy has been found in past research and warrants further research. For example, Eggen (1999, Table 1) reported observed accuracy of approximately 95% with nominal levels of 90%, 85%, and 80%.

Discussion

The results of the first study demonstrate the well-known (Kingsbury & Weiss, 1983; Spray & Reckase, 1996; Eggen, 1999) result that variable-length testing methods are highly efficient in the context of pass/fail decisions. While 100-item fixed-form tests produced approximately 95% accuracy, the SPRT and GLR could do so with less than 40 items on average. While 200-item fixed-form tests produced more than 96% accuracy, the SPRT and GLR could do so with approximately 50 items on average.

Moreover, the likelihood-ratio approaches (SPRT and GLR) produced even shorter tests than ACI, as has been shown in previous research (Eggen & Straetmans, 2000; Thompson, 2009b). However, the SPRT and GLR have one substantial disadvantage: the selection of items at the cutscore for each examinee means that each examinee receives the same test, as they would with a fixed-form approach. The adaptive item selection of ACI means that nearly every examinee sees a different set of items, aiding in test security by reducing overexposure of items. Nevertheless, for many applications this disadvantage might be irrelevant.

Additionally, the GLR is always at least as efficient as the fixed-point SPRT while maintaining equivalent levels of accuracy. If the value of δ is relatively large (> 0.3) then the two procedures are equivalent, but for smaller values there is a notable increase in efficiency with the GLR. This suggests that the GLR be used in applied assessment programs rather than the SPRT, especially since the difference in algorithm is small.

However, the most important result of this study is the strong effect that δ has on both the accuracy and efficiency of the test. For this reason, the width of the indifference region should never be specified by the arbitrary methods often suggested: attempting to estimate the θ values corresponding to a minimal pass or a maximal failure, or even worse, simply adding and subtracting an arbitrarily chosen number δ. Instead, a study such as this one should be conducted, designed based on actual characteristics of a testing program like bank size and examinee distribution, to determine the value of δ that produces the shortest test lengths while still maintaining the desired level of accuracy. This is especially true given the finding that observed accuracy is not necessarily equal to, or even near, the nominal accuracy of the procedure.

While the variable-length approaches investigated in this study require the use of IRT, similar tests can also be designed with classical test theory (Rudner, 2002). That has the advantage of smaller sample sizes for calibrating the item bank while still producing highly efficient CCTs, but has the drawback that it requires an independent verification of pass/fail for examinees in the calibration sample.

In summary, the GLR approach is optimal for testing programs that need to make a classification decision with as few items as possible, though fixed-form tests are still appropriate for many testing programs due to practical or content-distribution constraints. However, the design of any CCTs require extensive simulation research to ensure that the test is as efficient as possible.

Future Direction

A possibility for future research is an integrated likelihood ratio. Bayesian expectation *a posteriori* (EAP) estimation of θ takes into account asymmetry in the likelihood function by utilizing a weighted average, essentially an integration of the likelihood function. Because using the 3PL will produce asymmetric likelihood functions, EAP can be useful. This approach could be extended to classification by integrating the area below the curve to the left and right of the cutscore, then comparing the areas as a ratio. However, the raised lower asymptote of the 3PL could affect this detrimentally.

Future research should also consider more complex underlying models, such as mixture models or models that incorporate response time. Additional comparisons to other approaches could be made, such as decision theory methods (Rudner, 2002).

References

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* Monticello, New York: Marcel-Dekker.

Bartroff, J., Finkelman, M. & Lai, T.L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73*, 473-486.

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.

Frick, T. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187-213.

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

Huang, W. (2004). Stepwise likelihood ratio statistics in sequential studies. *Journal of the Royal Statistical Society, 66*, 401-409.

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.

Parshall, C.G., Spray, J.A., Kalohn, J.C., Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

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.

Thompson, N.A. (2009a). *Utilizing the generalized likelihood ratio as a termination criterion*. Paper presented at the 2009 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Thompson, N.A. (2009b). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793.

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

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

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