Utilizing the Generalized Likelihood Ratio as a Termination Criterion

Nathan A. Thompson

Assessment Systems Corporation

Paper Presented at the 2009 GMAC CAT Conference

Utilizing the Generalized Likelihood Ratio as a Termination Criterion

A common application for computer-based testing is to classify examinees into mutually exclusive groups. Currently, the predominant psychometric algorithm for designing computerized classification tests (CCTs) is the sequential probability ratio test (SPRT; Reckase, 1983) based on item response theory (IRT). The SPRT operates by formulating a point hypothesis test 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 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; Rudner, 2002). More recently, it was demonstrated that the SPRT, which only uses fixed values, is less efficient than a generalized form which tests whether a given examinee's θ is *below* θ_1*or above* θ_2 (Thompson, 2007). This formulation allows the indifference region to vary based on observed data. Moreover, this composite hypothesis formulation better represents the conceptual purpose of the exam, which is to test whether θ is above or below the cutscore.

The purpose of this study is to explore the specifications of the new generalized likelihood ratio (GLR: Huang, 2004). As with the SPRT, the efficiency of the procedure depends on the nominal error rates and the distance between θ_1 and θ_2 (Eggen, 1999). This study utilized a monte carlo approach, with 10,000 examinees simulated under each condition, to evaluate differences in efficiency and accuracy due to hypothesis structure, nominal error rate, and indifference region size.

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 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. 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.

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

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

Note that, since the probabilities are multiplied, the SPRT is equivalent to the ratio of the value of the IRT likelihood function at two points. The ratio is then compared to two decision points *A* and *B*, (Wald, 1947):

Lower decision point = $B = \beta / (1 - \alpha)$ (2)
Upper decision point = $A = (1 - \beta)/\alpha$. (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. The point hypothesis method calculates P_1 and P_2 at fixed points selected by the test developer, while the composite hypothesis method 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.

Point hypothesis formulation

The point hypothesis method 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). 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: $\theta = \theta_1$ (4)
H_1: $\theta = \theta_2$. (5)

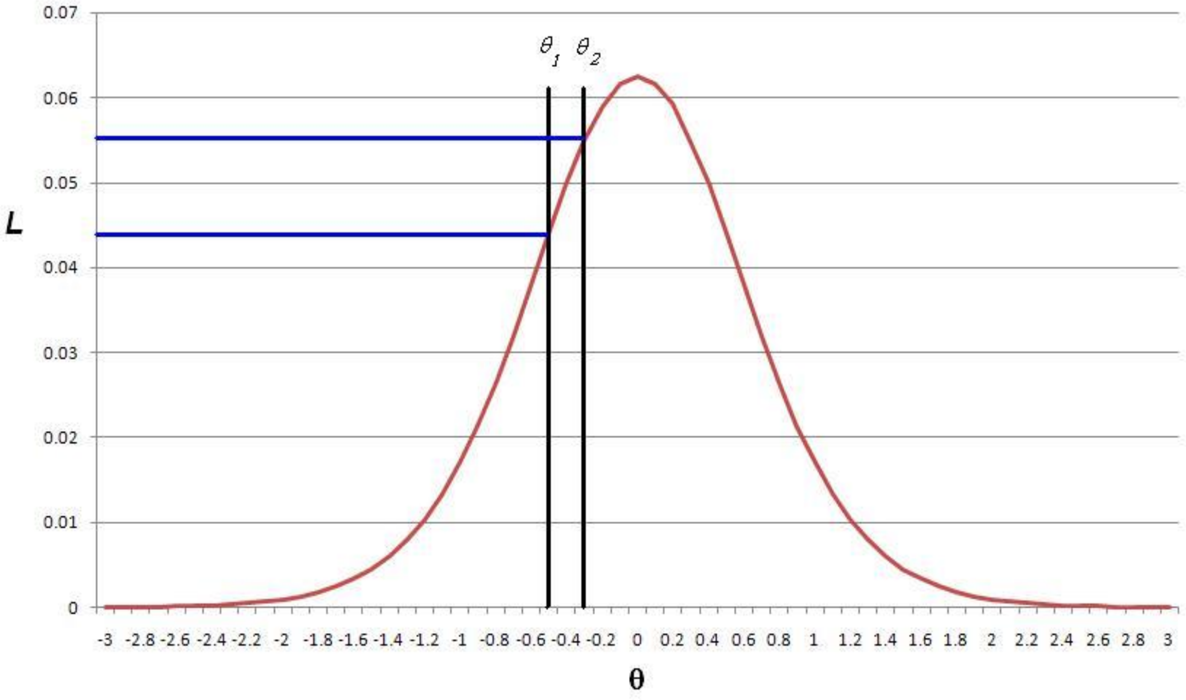
A graphic representation of this method is shown in Figure 1. In this example, the cutscore is 0.4 and δ= 0.1, such that $\theta_1 = 0.3$ and θ_2= 0.5. The likelihood function is evaluated at these two points, producing a ratio of approximately 0.55/0.44 = 1.25. The likelihood that the examinee is a “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.

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.

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, and will be utilized in the likelihood ratio. The side without the maximum is evaluated the same as with the SPRT.

Figure 1: Example likelihood function and indifference region



In the example of Figure 1, this modification to the likelihood ratio now produces a value of 0.62/0.44 = 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 study is whether this increase in efficiency comes with an increase in classification error (false positives and false negatives), and if the efficiency is moderated by nominal error rates or the width of the indifference region.

Method

A monte carlo simulation was designed to evaluate this research question via two substudies. The first investigated the effect of indifference region width on the efficiency of the GLR while simultaneously comparing the observed classification error rates to the nominal rate. The second compared the GLR with the SPRT.

Parameters were generated for a bank of 300 items. The descriptive statistics of the item parameters are shown in Table 1, and reflect the fact that the bank was 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. The study simulated a test for each examinee in each condition of the study, with the practical test length constraints of a minimum of 20 and a maximum of 200.

Table 1: Item parameter statistics

Statistic	*a*	*b*	*c*
Mean	0.70	-0.50	0.25
SD	0.20	0.51	0.04
Min	0.03	-1.98	0.13
Max	1.24	0.75	0.37

The dependent variables quantify both the efficiency and the accuracy of the simulated tests. The efficiency is indexed by the average test length (ATL), or mean number of items required to make a classification. The accuracy of the test is indexed by the percentage of examinees correctly classified (PCC), because the results of the test can be compared to the known true classification.

As mentioned, the first substudy only investigated the performance of the GLR. The independent variable was the width of the indifference region, manipulated by varying δ from 0.0 to 0.50 in increments of 0.02. The results are presented in Figure 2 for a nominal error rate of 5% and in Figure 3 for a nominal error rate of 1%.

Figure 2: ATL and PCC for 5% nominal error rate with the GLR

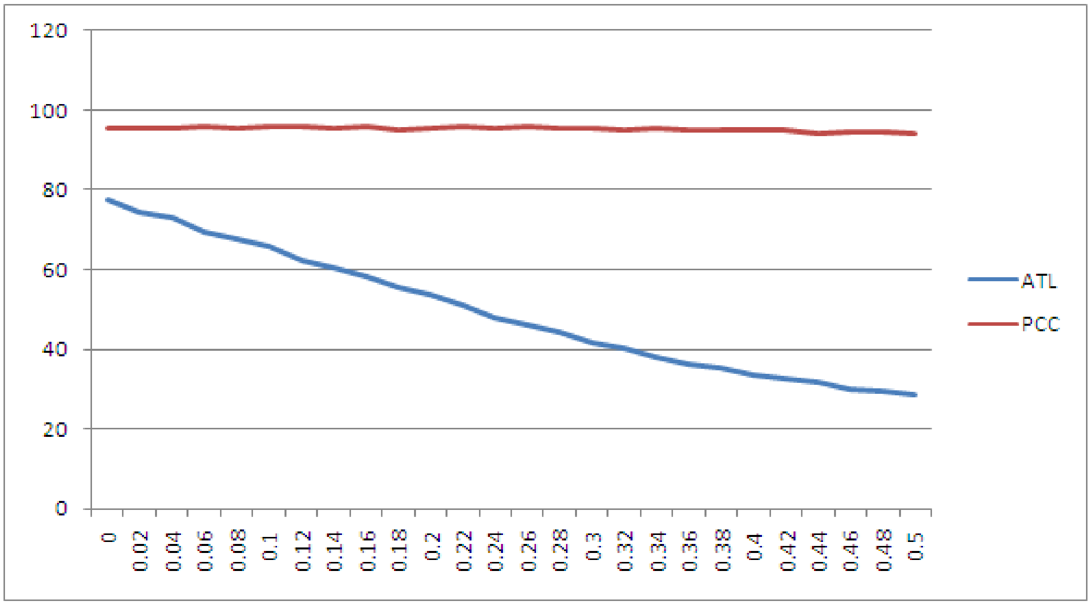
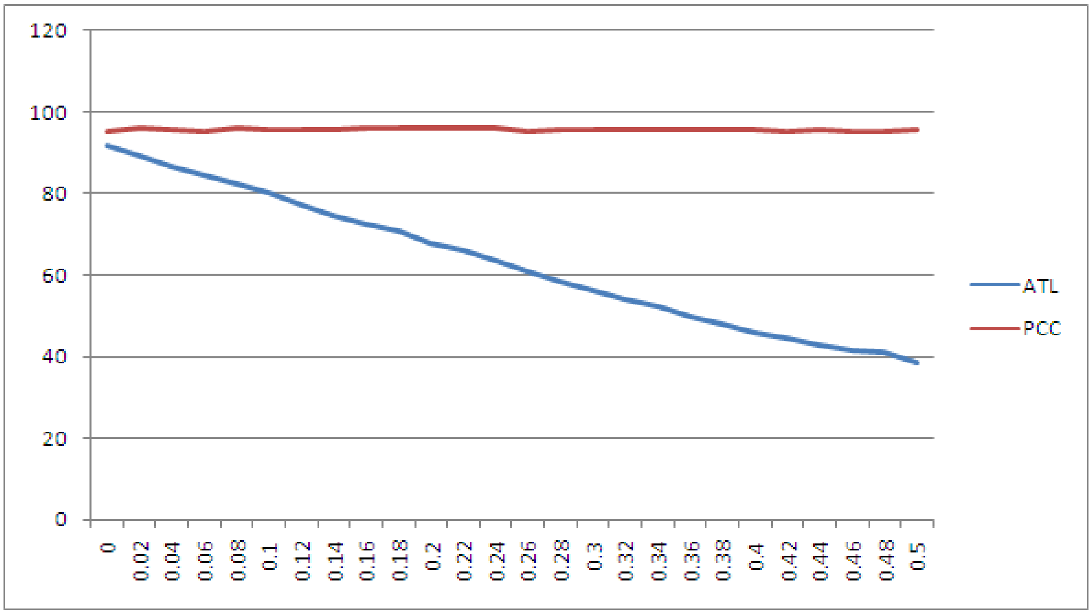


Figure 3: ATL and PCC for 1% nominal error rate with the GLR



The PCC decreased minimally as δ increased, while ATL dropped dramatically. Increasing the size of the indifference region will greatly decrease the number of items needed to make classifications, but will also decrease the accuracy of the test. This effect was true in both conditions. However, the accuracy remained near nominal levels for the 5% condition; for the 1% condition, observed accuracy was always lower than the nominal accuracy.

Having demonstrated the ability of the GLR to classify examinees regardless of indifference region width, the next step was to compare the efficiency of the GLR to the fixed-point SPRT. The same study was completed, but with δ intervals of 0.1 rather than 0.02. The results are presented in Figures 4 and 5.

Figure 4: ATL and PCC for 5% nominal error rate, comparing GLR and SPRT

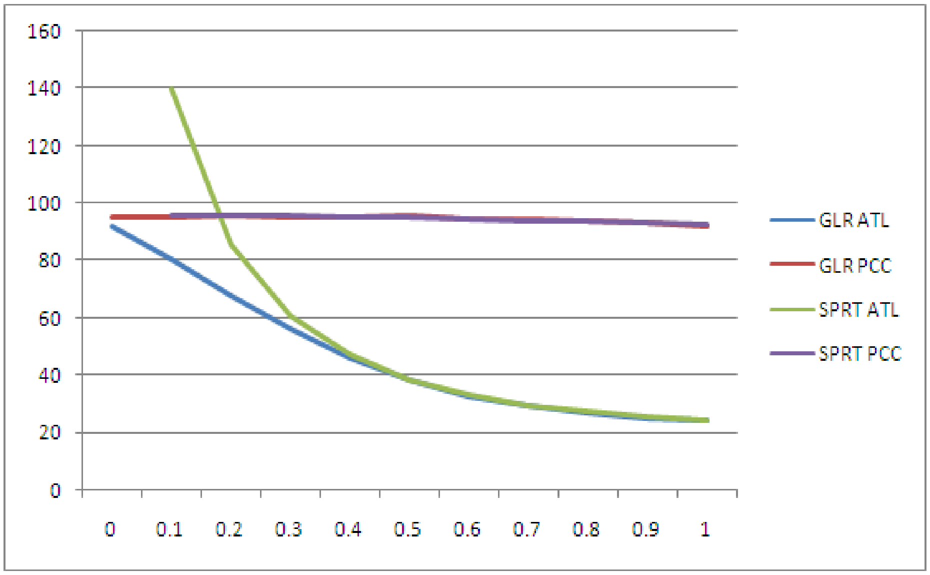
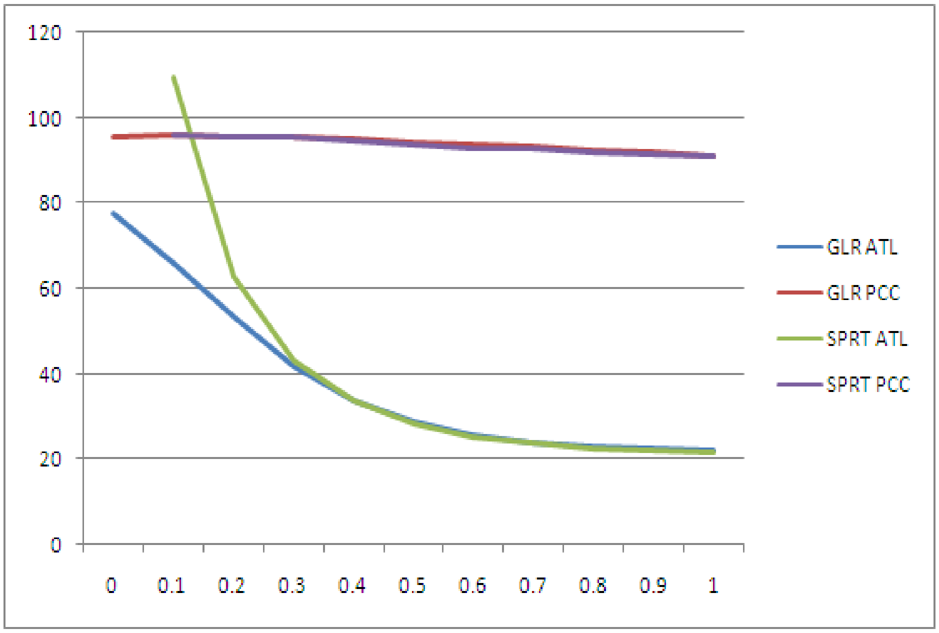


Figure 5: ATL and PCC for 5% nominal error rate, comparing GLR and SPRT



Figures 4 and 5 present a similar shape to Figures 2 and 3, with the ATL dropping sharply while the PCC decreases minimally. The GLR requires fewer items when δ is 0.3 or smaller, while the two methods perform equivalently with larger values of δ. However, note that PCC appears stable when δ is 0.3 or smaller, but decreases afterwards.

Discussion

As is evident in Figures 4 and 5, the GLR is always at least as efficient as the fixed-point SPRT while maintaining equivalent levels of accuracy. 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 message 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.

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.

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

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., & Ro, S. (2007). *Computerized classification testing with composite hypotheses*. Presentation at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

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