Advanced Methods of Designing Tests for Pass/Fail Decisions

Nathan A. Thompson

Assessment Systems Corporation

Poster presented at the 2010 Annual Conference of the
Council for Licensure, Enforcement, and Regulation

September 23-25, 2010

Nashville, TN

Abstract

Most examinations in the realm of professional regulatory testing are designed with the purpose of pass/fail decisions for each examinee. Historically, these decisions have been made by administering a fixed number of items to each examinee, and assigning a "pass" decision if the examinee observed score is equal to or above a cutscore.

However, this method is inefficient, in that it requires a large number of items to be administered before the decision is made. It is for this reason that computerized adaptive testing (CAT) methods were developed. Yet even CAT methods are not optimal for regulatory testing, as they are designed to obtain precise scores, and precise scores are not needed – only a pass/fail decision. A related methodology, computerized classification testing (CCT), specifically designs tests to provide this decision with as few items as possible, but retaining the decision accuracy of both fixed-form and CAT methods.

This paper will explain differences between these approaches and provide a comparison of them to demonstrate the efficiency of the CCT approach for pass/fail testing. Simulations will be conducted for tests under each approach, and results compared in terms of decision accuracy (percentage correctly classified) and efficiency (average test length). The applicability of each method in the licensure/certification context will be discussed. Participants will be able to recognize the advantages and disadvantages to help determine if CCT or CAT methods might be useful for their testing program.

Advanced Methods of Designing Tests for Pass/Fail Decisions

In licensure and certification testing, the purpose of the test is primarily to classify examinees into mutually exclusive groups rather than obtain accurate estimates of individual scores. 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 ability confidence intervals (ACI: Kingsbury & Weiss, 1983; Thompson, 2007) 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 (Bartroff, Finkelman, & Lai, 2008; Thompson, 2009b). 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 compare the generalized likelihood ratio (GLR: Huang, 2004) to the SPRT, ACI, and baseline fixed-format approaches. Because the three sophisticated methods are *variable-length*, meaning they can stop the test as soon as a confident decision can be made, they can produce much shorter tests. For example, if a high-ability examinee answers 45 out of the first 50 questions correctly, they can be classified as a "pass" without administering more items. The GLR, SPRT, and ACI methods use this approach, but quantify the decision within the framework of IRT.

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 variable-length methods produce tests that are much shorter but just as accurate; Kingsbury and Weiss (1984) suggest that the reduction is typically 50% with no loss of accuracy.

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):

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

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,

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

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):

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

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 θ, then adding and subtracting it from the cutscore (e.g., Eggen, 1999; Eggen & Straetmans, 2000).

Therefore, the hypothesis test is structured as

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

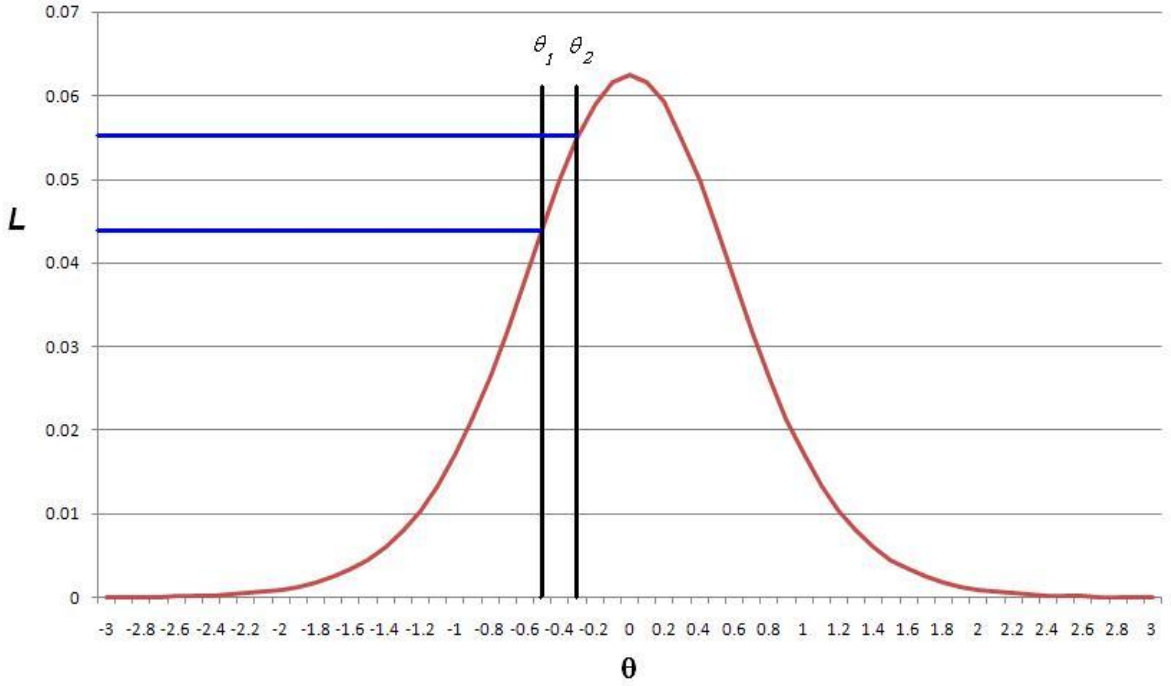
A graphic representation of this method is shown in Figure 1. In this example, the cutscore is 0.4 and $\delta = 0.1$, such that $\theta_1 = 0.3$ and $\theta_2 = 0.5$. The likelihood function is evaluated at these two points, producing a ratio of approximately $0.55/0.44 = 1.25$. The likelihood that 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 θ 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.

ACI

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, θ, using the conditional standard error of measurement (SEM). This can be expressed as (Thompson, 2009a):

$$\hat{\theta}_j - z_\varepsilon(SEM) \le \theta_j \le \hat{\theta}_j + z_\varepsilon(SEM) \qquad (6)$$

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_ε = 1.96, with α =0.025, β = 0.025, and ε = 0.05. While the SPRT and GLR differentiate only at the cutscore, ACI evaluates across the spectrum of θ, wherever the current estimate lies. Therefore, previous research (Thompson, 2009a) 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. The theoretical SEM is calculated using the item information functions evaluated at the relevant θ regardless of response pattern, and θ is estimated using brute force methods. In practice, it is more common to estimate θ with Newton-Raphson methods, and estimate an SEM using the observed likelihood function.

Method

The independent variable of this study is the design of the test. The three primary levels investigated are ACI, SPRT, and GLR variable-length termination criteria. Fixed-form tests of 200 items and 100 items, with both number-correct and IRT scoring, are used 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 high 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.

Because the value of δ affects the results of the SPRT and GLR, it was varied to provide a better opportunity for comparison. 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 is reached, which was 0.3. Simulations were also completed with $\delta = 0.2$ for comparison.

The cutscore for the simulations was $\theta = -0.5$, which corresponds to a pass rate of approximately 69%. For the fixed-form tests with number-correct scoring, this was converted to a raw cutscore using the test response function: 122.5 for the 200-item test and 63.85 for the 100-item test. The variable-length tests were constrained to have a minimum of 20 items and a minimum of 200 items. A maximum is necessary, otherwise the entire bank would be administered to examinees with true ability right at the cutscore, because a decision would never be able to be made with confidence.

The bank for the test consisted of 500 items with IRT parameters to represent plausible values for a credentialing test. The difficulty of the bank was centered on the cutscore of -0.5, 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

Results

The results of the simulations are presented in Table 2. As hypothesized, the variable-length methods produced much shorter tests, with ATL ranging from 37.93 to 55.53, while maintain the level of accuracy produced by the fixed form tests that delivered two to four times as many items.

Table 2: ATL and PCC for each condition

Test design	**ATL**	**PCC**
200 item fixed (number correct)	200.00	96.09
200 item fixed (IRT)	200.00	96.17
100 item fixed (number correct)	100.00	95.18
100 item fixed (IRT)	100.00	95.29
ACI (model SEM)	51.65	95.73
ACI (observed SEM)	54.61	95.78
SPRT ($\delta = 0.3$)	39.30	95.74
GLR ($\delta = 0.3$)	37.93	95.66
SPRT ($\delta = 0.2$)	55.53	96.35
GLR ($\delta = 0.2$)	48.44	96.03

Also notable is the differences between the variable-length methods. The SPRT and GLR produced shorter tests than ACI while maintaining accuracy. The GLR was slightly more efficient than the SPRT; this gain in efficiency increases with a decrease in δ. This is because a wide δ forces the GLR and SPRT to utilize the same calculations.

Conclusions

The results demonstrate that variable-length testing methods are far more efficient than fixed forms for 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 show in previous research (Eggen, 1999; 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 over-exposure of items.

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 approach has the drawback that is requires an independent verification of pass/fail for examinees in the calibration sample. IRT-based approaches do not require this, although they have a requirement of greater sample size.

In summary, credentialing examinations can utilize variable-length approaches to drastically reduce test length while maintaining accuracy of decisions. Additional psychometric expertise is required to implement these effectively, though that expertise is also necessary for IRT-based fixed-form examinations. However, the benefits of shorter tests with equivalent accuracy can easily offset the cost of additional expertise. Therefore, this type of approach is optimal for testing programs with the volume necessary to implement IRT.

Contact information:
Nathan Thompson
Vice President, Assessment Systems Corporation
www.assess.com
nthompson@assess.com

References

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.

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

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

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