Nominal Error Rates in Computerized Classification Testing

Nathan A. Thompson

Assessment Systems Corporation

Paper presented at the First Annual Conference of the

International Association for Computerized Adaptive Testing

June 7-9, 2010

Arnhem, The Netherlands

Abstract

A common finding in computerized classification testing (CCT) research, though often not explicitly noted, is that observed error rates do not always follow nominal error rates, and in some cases are substantially different. There are likely several contributing factors, including the size of the indifference region, which is typically selected arbitrarily rather than empirically, and the information structure of the item bank. This paper will utilize monte carlo simulations to manipulate these two variables, and investigate their effect on the observed error rates in CCT. Both approaches to the likelihood ratio criterion (point and composite) will be utilized.

Nominal Error Rates in Computerized Classification Testing

A common application for computer-based testing is to classify examinees into mutually exclusive groups, known as computerized classification testing (CCT). A large body of research exists on CCT, dating back to Ferguson (1967). Currently, the predominant psychometric algorithm for designing CCTs is the likelihood ratio approach based on item response theory (IRT). This approach has been shown to be more efficient than confidence intervals around ability estimates (Spray & Reckase, 1996; Rudner, 2002; Thompson, 2009a). The pass/fail decision is formulated as a ratio, namely the likelihood an examinee is above the cutscore divided by the likelihood the examinee is below the cutscore. There are two methods of using the likelihood ratio: a point hypothesis and a composite hypothesis. The sequential probability ratio test (SPRT; Reckase, 1983) operates with a point hypothesis, defining the ratio such that a given examinee's ability value θ is equal to a fixed value below (θ_1) or above (θ_2) the cutscore.

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 rather than specifically at each point (Thompson, 2009b). This formulation is more flexible in that it allows the ratio to adjust its definition 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. Yet the concept of a likelihood ratio test of pass vs. fail remains the same.

Using either approach to the likelihood ratio requires the specification of two pieces of information. The space between the two points θ_1 and θ_2 is referred to as the *indifference region*, as the test developer is indifferent to the classification assigned. The indifference region is typically defined arbitrarily by adding and subtracting a small number δ from the cutscore. The likelihood ratio also requires the specification of nominal error rates (or conversely, accuracy) that represent the amount of classification error the test designer is willing to consider acceptable. A common finding in CCT research, though often not explicitly noted, is that observed misclassification error rates do not always follow nominal error rates specified in the algorithm, and in some cases are substantially different (e.g., Eggen & Straetmans, 2000; Thompson, 2009b). The purpose of this paper is to explore CCT specifications that possibly affect nominal error rates.

There are likely several contributing factors. An important factor is likely the size of the indifference region, which is typically selected arbitrarily or conceptually rather than empirically even though it has been shown in Thompson (2009) and Eggen (1999) to directly affect both the accuracy and efficiency of a CCT. The information structure of the item bank is also likely important, because the bank must contain enough information to produce the required accuracy. Additionally, the nominal error rates should obviously have impact on observed error rates.

This paper will utilize monte carlo simulations to manipulate these three variables, indifference region, item bank structure, and nominal error rates, to investigate their effect on the observed error rates in classification. Both approaches to the likelihood ratio criterion will be utilized: the point hypothesis method (SPRT), and the composite hypothesis method, known as the generalized likelihood ratio (GLR).

The likelihood ratio approach

The likelihood ratio for sequential testing (Wald, 1947) 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 ratio is expressed as the ratio of the likelihood of a response at two points on θ, θ_1 and θ_2,

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

Note that, since the probabilities are multiplied, this 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 = □ (2)
Upper decision point = A = □ (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 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 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.

Point hypothesis formulation

The point hypothesis method suggested by Reckase (1983) is termed the sequential probability ratio test (SPRT), and 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

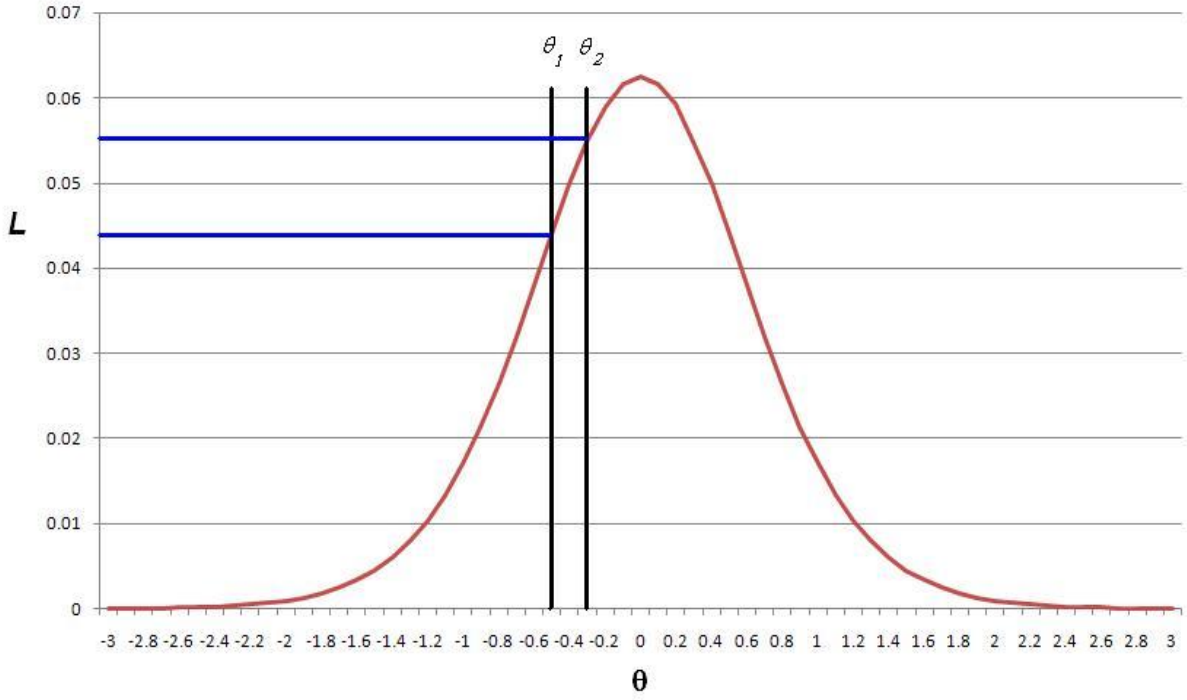
H_0: $\theta = \theta_1$ (4)
H_1: $\theta = \theta_2$. (5)

A graphic representation of this method is shown in Figure 1. In this example, the cutscore is 0.4 and $\delta = 0.1$, 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 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.

Figure 1: Example likelihood function and indifference region



The generalized likelihood ratio

The GLR (Bartroff, Finkelman, & Lai, 2008; Thompson, 2009) 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, so the value of the likelihood at that point will be utilized in the likelihood ratio. The side without the maximum is evaluated the same as with the SPRT.

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.

Nominal error rates

The accuracy of classifications made with the SPRT and GLR is nominally controlled by the two error rates α and β. Adding these two values and subtracting the result from 1.0 produces the nominal accuracy on the proportion metric. However, CCT research typically reports the percentage correctly classified (PCC) rather than the proportion. In addition, CCT research evaluates the average test length (ATL); if a procedure can produce the same accuracy as another procedure but with fewer items, it is more efficient and typically considered preferable.

Unfortunately, the nominal accuracy does not always match PCC observed in simulation studies. Table 1 and Table 2 present such results from past research. Eggen (1999) was comparing two item

selection methods based on Fisher information to two methods based on Kullback-Liebler, while Eggen and Straetmans (2000) was evaluating the effect of content (C) and exposure (E) constraints on maximum information (MI) selection. In both cases, three levels of nominal accuracy were used: 90%, 85%, and 80%. However, all of the observed PCC in Table 1 is near 95%, while all of the observed PCC in Table 2 is near 90%. Results similar to Eggen (1999) were also found by Lin and Spray (2000) and Thompson (2009).

Table 1: Results from Eggen (1999)

Nominal PCC	Item selection method									
	F1		F2		K1a		K1b		K1c	
	ATL	PCC	ATL	PCC	ATL	PCC	ATL	PCC	ATL	PCC
90	16.0	95.6	16.3	94.7	16.1	95.4	16.3	94.6	15.9	95.5
85	14.9	95.0	14.0	95.2	13.9	94.8	13.9	95.2	13.9	95.6
80	13.2	94.9	12.7	94.8	13.2	94.8	12.7	95.3	12.9	94.8

Table 2: Results from Eggen & Straetmans (2000)

Nominal PCC	Item Selection Method							
	MI		MI + C		MI + E		MI + C + E	
	ATL	PCC	ATL	PCC	ATL	PCC	ATL	PCC
90	17.6	90.6	18.4	88.9	18.5	86.8	18.7	88.4
85	16.7	87.8	16.6	91.1	17.0	88.0	17.8	89.2
80	15.2	89.5	15.3	88.5	15.9	87.6	16.1	90.0

Such results suggest that observed accuracy is a function of factors other than nominal accuracy. This study is designed to vary item bank structure, termination criterion, and δ in addition to nominal accuracy to investigate the possible impact of those variables on observed accuracy. These factors were held constant in Eggen (1999) and Eggen and Straetmans (2000).

Method

A monte carlo simulation was designed to evaluate the four independent variables mentioned above. The levels selected to simulate are presented in Table 3.

Table 3: Independent variables and levels

Variable	Levels
Termination	SPRT, GLR
Indifference region	0.0, 0.1, 0.2, 0.3, 0.4
Item bank	Broad, peaked
Nominal accuracy	90-99% in increments of 1%

Parameters were generated both banks of 300 items, with the broad bank determined by generating a standard deviation of *b* parameters of 1.5, as compared to 0.5 for the peaked. The observed descriptive statistics of the item parameters are shown in Table 4. A distribution of 10,000 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 4: Item bank statistics

Statistic	Broad	Peaked
Mean *a*	0.71	0.70
SD *a*	0.20	0.20
Mean *b*	-0.50	-0.50
SD *b*	1.46	0.51
Mean *c*	0.25	0.25
SD *c*	0.04	0.04

Results

The results of this study were highly similar to Eggen (1999) in that the observed PCC was near 95% regardless of the independent variables. This can be seen in Table 6, which presents the marginal means for each independent variable other than nominal PCC. Item bank and δ accounted for some variance in observed PCC, though termination criterion did not. There was a difference of only 0.01% between the accuracy of the GLR and the SPRT, though the GLR used nearly 11 fewer items, on average. The peaked item bank performed better with respect to both ATL and PCC because it provides more information near the cutscore, where the likelihood ratio is being calculated. Finally, an increase in δ provided a great decrease in test length, but at the expense of a small amount of accuracy.

Table 6: Marginal means of ATL and PCC for termination, bank, and δ

Variable	Level	ATL	PCC
Termination criterion	GLR	57.66	95.11
	SPRT	68.52	95.12
Item bank	Broad	66.26	94.85
	Peaked	58.70	95.38
δ	0.0*	80.67	95.24
	0.1	93.95	95.38
	0.2	63.10	95.34
	0.3	46.81	95.09
	0.4	36.98	94.58

*GLR only; not possible with SPRT

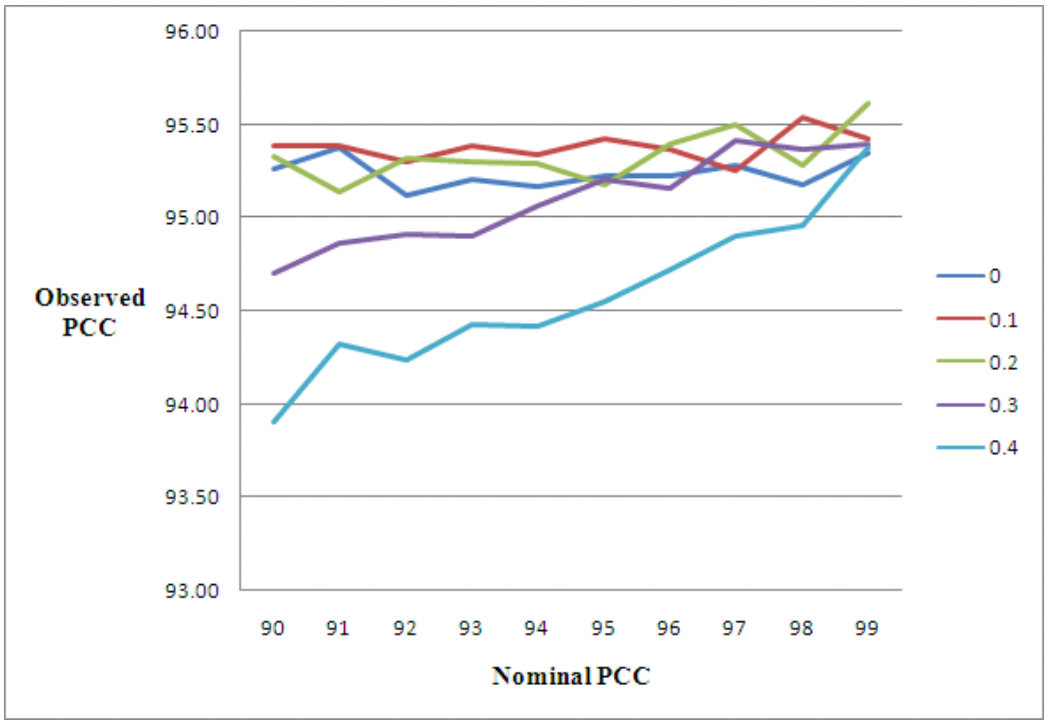
The observed PCC levels remained near 95% even when the nominal PCC was varied (Table 7), though as noted with δ, a slight decrease in observed PCC can bring about a substantial decrease in ATL. Table 7 also presents the difference between the observed and nominal PCC.

Table 7: Marginal means of ATL and PCC for nominal PCC

Nominal PCC	ATL	PCC	PCC Diff
90	52.40	94.88	4.88
91	53.95	94.98	3.98
92	55.65	94.96	2.96
93	57.29	95.03	2.03
94	59.21	95.04	1.04
95	61.76	95.10	0.10
96	64.54	95.17	-0.83
97	67.75	95.27	-1.73
98	72.53	95.27	-2.73
99	79.76	95.44	-3.56

There was little interaction between nominal PCC and termination criterion or item bank, possibly because there were only two levels investigated for each. However, there was a modest interaction between nominal PCC and δ. In Figure 1, the approximate slope of in line decreases as δ decreases. This means that for very small values of δ (0.2 or less), the nominal PCC had virtually no effect on observed PCC. It was always between 95.00 and 95.50. However, for values of δ greater than 0.2, a decrease in nominal PCC brought about a decrease in observed PCC.

Figure 1: Observed PCC as a function of nominal PCC and δ



Discussion

As found in a number of previous studies, nominal PCC had very little effect on observed PCC. This was not surprising, given past evidence; however, it is notable that the three additional variables hypothesized to have an effect were found to have little effect either. It is not that the likelihood ratio approach is so robust that it simply provides a high level of accuracy regardless of specifications; when the nominal PCC was 0.96 or greater, the procedure actually underperformed, unable to attain that level of accuracy.

Therefore, further research is necessary to investigate other possible variables that might affect observed PCC. One possibility is the location of the cutscore. This has an effect on observed PCC (Lin & Spray, 2000), as it is far easier to make classifications if the cutscore is in the extremes. Typically, only a few items might be needed to classify an examinee above a cutscore of -2.0 or below +2.0. However, very few testing applications utilize such extreme cutscores, so a complete understanding of the procedure in the context of typical cutscores in the range of -1.0 to +1.0 is still not achieved. Item selection method is another relevant independent variable, but Eggen (1999) and Eggen and Straetmans (2000) found little effect on PCC.

Moreover, these secondary variables remain simply that: secondary. Only δ and the nominal PCC are specifications of the likelihood ratio approach. So perhaps the issue is in the likelihood ratio procedure itself, either in the calculation of the ratio or in the calculation of the upper and lower bounds used to make decisions. There are still improvements to be made, as the GLR was only applied to educational assessment recently (Bartroff, Finkelman, & Lai, 2008; Thompson, 2009), as it is still fairly recent in the statistical literature (e.g., Huang, 2004). One possibility is an integrated likelihood ratio (Thompson & Ro, 2007), which did not perform adequately, implying that some type of adjustment is needed, such as adjustments for the A and B bounds. This was also suggested by Bartroff, Finkelman, and Lai (2008) in the context of the GLR.

In summary, the GLR remains the most efficient method of classifying examinees into two groups, with the SPRT slightly less efficient. The apparent lack of control over observed accuracy is somewhat troubling, though it is not an issue if simulation studies show that it meets the goals of a particular testing program. Further research is needed to explore this issue.

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.

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

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.

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.