Cutscore Location and Classification Accuracy in Computerized Classification Testing

Shungwon Ro

Nathan A. Thompson

Prometric

Correspondence:
Nathan A. Thompson
Prometric
1260 Energy Lane
St. Paul, MN 55108
Nathan.thompson@thomson.com

Abstract

A common dependent variable in studies of computerized classification testing (CCT; Parshall, Spray, Kalhon, & Davey, 2006) is the amount of classification error or accuracy. This is controlled in part by nominal values that are chosen as parameters for the termination criterion of the CCT. Past research, however, has focused on a comparison of observed classification error to assess the efficiency of various CCT algorithms, and has not evaluated the relationship between observed and nominal error rates. This study explored classification accuracy with CCT as a function of cutscore location with relation to the examinee distribution, as a cutscore near the center will lead to much more error than a very high or low cutscore. In addition to classification accuracy, the average number of items required to make a classification decision is also considered. A cutscore that is very high or low will fail or pass, respectively, most examinees with very few items.

The purpose of this monte carlo simulation study is to evaluate the extent of this effect and its relationship to several relevant independent variables. Because the relationship between nominal and observed accuracy is being evaluated, nominal accuracy rates are varied at 99% and 95%. Three CCT termination criterion are utilized: the sequential probability ratio test (SPRT; Wald, 1947; Reckase, 1983), the composite likelihood ratio (Thompson & Ro, 2007), and IRT confidence intervals (Kingsbury & Weiss, 1983). The additional *a priori* parameter for the SPRT, indifference region, is also varied due to its effect on test length and accuracy.

This topic is of great practical importance because a test user should be aware of the fact that an operational CCT may have classification error greater than or less than the nominal rates selected a priori, and other aspects of the CCT might need to be adjusted to achieve desired levels of accuracy. The effects of cutscore location on test length may also require the application of practical constraints.

Cutscore Location and Classification Accuracy in Computerized Classification Testing

Computerized classification testing (CCT; Parshall, Spray, Kalohn, & Davey, 2006) refers to computerized tests that are designed to classify examinees into groups rather than obtain a precise estimate of ability. Often, CCTs utilize the variable-length testing paradigm, where the test is terminated when a statistical criterion is met rather than administering a fixed set of items, though the generality of the name does not constrain it to be so. The current study investigates classification accuracy with the more psychometrically sophisticated approach of variable-length CCT.

CCT research commonly makes use of two dependent variables to compare competing methods of test design (e.g., Spray & Reckase, 1996): classification accuracy (or inversely, error) and the number of items needed to satisfy the termination criterion. The former is frequently operationalized as a percentage or proportion of examinees correctly classified (PCC). In monte carlo studies, where the true ability (θ) value is known for each examinee, this is the percentage of correct classifications. In studies involving real data, where the true θ value is not known for examinees, the analogous index is the proportion of examinees consistently classified.

The amount of observed classification error is ostensibly a function of the *a priori* nominal error rates that set as parameters of the termination criterion. However, this is not necessarily the case. A great example of this is found in Eggen (1999; p. 257), who found that observed PCC tended to always be near 95%, whether the nominal accuracy was 90%, 85%, or 80%.

Additionally, observed accuracy also a function of several additional characteristics of the test. Important among these is the location of the cutscore in relation to the examinee distribution. If the cutscore is very high or low, there will be relatively few examinees near the cutscore, decreasing the opportunity for the test to make a classification error. For example, if the cutscore is very low, the test will easily be able to classify most examinees as above cutscore, likely with few items. This is important for the test designer to take into account; if the cutscore is at the center of the distribution, there is more opportunity for error, and observed error may actually be higher than nominal levels.

The observed error is also affected by other characteristics. If the termination criterion used is the sequential probability ratio test (SPRT: Wald, 1947; Reckase, 1983), a parameter known as the indifference region can affect the length and accuracy of the test. This small region around the cutscore is defined by the test developer as a range of ability where they are indifferent as to which classification is made. If this region is wider, there are more examinees that the test is "indifferent" about, increasing classification error.

If test length constraints are applied, they will also affect the observed classification accuracy. A maximum test length will terminate a test before a termination criterion is met for some examinees. Classification error is more likely for such examinees. If the maximum test length is relatively small, such as 10 items, classification accuracy will drop substantially (Rudner, 2002).

Some practical constraints, such as content distribution or item exposure controls, have little effect on classification error (Kalohn & Spray, 1998; Lin & Spray, 2000; Eggen & Straetmans, 2000). These two constraints on item selection tend to contribute to the selection of an item that does not necessarily maximize the psychometric contribution. Therefore, they simply require a few more items to meet the psychometric requirements of the termination criterion.

The goal of this study is to evaluate the effect of cutscore location and nominal error on observed classification accuracy, operationalized as PCC. In addition, the average test length (ATL) for each condition will be evaluated in addition to PCC. Three termination criteria are considered: the SPRT, ability confidence intervals (ACI: Kingsbury & Weiss, 1983), and the composite likelihood ratio (CLR: Thompson & Ro, 2007).

Method

This study utilized a monte carlo simulation methodology. A sample of 10,000 examinee θ values was generated, and used to determine a true classification. A CCT was simulated for each examinee in each condition, and the result compared to the true classification to determine PCC. The number of items needed to make a classification was recorded to determine ATL.

Altogether, four independent variables were included in the study, creating a total of 50 conditions. They were not completely crossed, as indifference region is nested within the SPRT condition. ACI and the CLR have no analogous parameter and therefore have no corresponding nested independent variable. The variables were (levels in parentheses):

1. Nominal accuracy (95%, 99%)
2. Cutscore location (-1, -0.5, 0, 0.5, 1)
3. Termination criterion (ACI, SPRT, CLR)
4. Indifference region for SPRT (δ = 0.2, 0.3, 0.4).

The independent variable that theoretically should have the most effect on observed error is nominal error rates. Two levels were examined, 1% and 5%, which correspond to a 99% and 95% nominal PCC. For the SPRT, Type I error rate α and Type II error rate β are specified separately, which corresponds to $\alpha = \beta = 0.005$ (0.5%) and $\alpha = \beta = 0.025$ (2.5%), respectively.

Because, as mentioned above, a cutscore that is closer to a greater number of examinees provides more opportunity for misclassification, the distance of the cutscore from the mean of the examinee distribution was varied. It was hypothesized that the location of the cutscore would have a somewhat lesser effect than nominal error rates on observed PCC, with PCC being lower for a cutscore of $\theta_c = 0.0$ and higher for $\theta_c = -1$ and $\theta_c = 1$.

Each of the termination criteria in this study employed dichotomous IRT, though polytomous IRT can be applied (Lau & Wang, 1998; 1999; Thompson, 2007a). For the multiple-choice achievement items that commonly compose CCTs, the most appropriate dichotomous model is the three-parameter model. With this model, the probability of a correct response to an item is a function of θ called the item response function (Hambleton & Swaminathan, 1985, Eq. 3.3):

$$P_i(X = 1 \mid \theta_j) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \quad (1)$$

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 first of the termination criteria, ACI, is similar to adaptive testing for point estimation of θ and has therefore been referred to as the statistical estimation approach (Eggen & Straetmans, 2000). Items are selected to maximize information at the current estimate of ability and a likelihood function of θ is determined. However, while point estimation adaptive testing terminates when the conditional standard error of measurement (CSEM) falls below an arbitrarily chosen constant, a CCT with ACI terminates when a confidence interval using the CSEM falls above or below the cutscore. Specifically, a confidence interval is constructed around the current ability estimate $\hat{\theta}_j$ using (Thompson, 2006)

$$\hat{\theta}_j - z_\alpha(CSEM) \le \theta_j \le \hat{\theta}_j + z_\alpha(CSEM) \quad (3)$$

where z_α is the normal deviate corresponding to a 1-α confidence interval. For instance, if the interval used is 95%, an examinee would pass when $\hat{\theta}_j$ is 1.96(CSEM) above θ_c. The CSEM used was the model-predicted calculation, where

$$CSEM = 1/\sqrt{I(\theta)} \quad (4)$$

and test information $I(\theta)$ is (Embretson & Reise, 2000, Eq. 7A.2)

$$I(\theta) = \sum D^2 a_i{}^2 P(\theta)(P(\theta) - 1). \quad (5)$$

The SPRT, on the other hand formulates the classification problem as a hypothesis test that $\theta_j = \theta_1$ or $\theta_j = \theta_2$, where θ_1 is an arbitrarily chosen point below the cutscore and θ_2 above the cutscore, delineating the indifference region. The test is calculated as a likelihood ratio comparing the two aforementioned hypotheses, where X is the observed response to item i:

$$LR = \frac{\prod_{i=1}^{n} P_{2i}{}^X (1 - P_{2i})^{1-X}}{\prod_{i=1}^{n} P_{1i}{}^X (1 - P_{1i})^{1-X}}. \quad (6)$$

This ratio is compared to two decision points A and B (Wald, 1947)

Lower decision point = $B = \beta / (1 - \alpha)$ (7)
Upper decision point = $A = (1-\beta)/\alpha$. (8)

If LR is below B, the examinee is classified as below the cutscore. If LR is above A, they are classified as above the cutscore. If LR is between A and B, another item is administered.

P_{1i} and P_{2i} are the probability of a correct response to item i for an examinee with $\theta_j = \theta_1$ or $\theta_j = \theta_2$, respectively. The distance between θ_1 and θ_2 is referred to as the indifference region because the test user is indifferent as to whether examinees in the region are classified in either group. If the indifference region is wide, the values of P_{1i} and P_{2i} will be more disparate than if the region was narrow, because the item response function is continuously increasing. This causes LR to diverge more quickly, which in turn causes examinees to be classified with fewer items. This can cause observed classification error to increase. Therefore, several indifference region widths will be examined for the SPRT: 0.2, 0.3, and 0.4.

Recently, a termination criterion was suggested that combines ACI and the SPRT into a single method, the composite likelihood ratio (CLR: Thompson & Ro, 2007). The IRT-based likelihood function that is used as a basis for the confidence interval is evaluated using a likelihood ratio approach. Specifically, the likelihood function is integrated below and above the cutscore, and the two values are compared as a ratio. With Riemann midpoint integration, this is presented as

$$LR_c = \frac{L(\theta \in \Theta_2)}{L(\theta \in \Theta_1)} = \frac{\sum_{\theta_c}^{\theta_c+k} L(u \mid \theta + 0.5\Delta\theta)\Delta\theta}{\sum_{\theta_c-k}^{\theta_c} L(u \mid \theta + 0.5\Delta\theta)\Delta\theta} \quad (9)$$

where u is the observed response vector, and

$$L(X_{1j}, X_{2j}, ..., X_{nj} \mid \theta_j) = \prod_{i=1}^{n} P_i(\theta_j)^{X_{ij}} (P_i(\theta_j) - 1)^{1-X_{ij}} . \quad (10)$$

Thompson and Ro (2007) proposed an adjustment of the boundaries A and B for this type of likelihood ratio 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:

$$\text{Lower decision point} = B = \frac{\beta}{1-\alpha} \times \frac{1}{\gamma\sqrt{n}} \quad (11)$$

$$\text{Upper decision point} = A = \frac{1-\beta}{\alpha} \times \frac{1}{\gamma\sqrt{n}} . \quad (12)$$

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. For this study, $\gamma = 1.0$, so that the adjustment was only relative to n.

No test length constraints were applied to the monte carlo simulations because the purpose of the study was to investigate the performance of the termination criteria. As mentioned above, applying a maximum test length decreases ATL and PCC. Conversely, adding a minimum test length increases the ATL and PCC from the level "normally" produced. In an extreme case, a minimum test length could be specified at a relatively large value, such as 50 items; this would lead to tests that required more items on average but were very accurate.

Item selection method was nested within termination criterion. The SPRT performs more efficiently with cutscore-based item selection, maximizing Fisher information at the cutscore, because this produces the greatest $P_{1i} - P_{2i}$ difference (Thompson, 2007b). ACI and the CLR perform more efficiently with estimate-based cutscore selection, as the item with the highest information at the current θ estimate will lead to the greatest reduction in the dispersion of the likelihood function.

Results

The results of the simulation are presented in Table 1 for the primary independent variable, cutscore location. As expected, a cutscore nearer the center of the examinee distribution requires more items while still having reduced accuracy. Also note that, while a higher level of nominal accuracy (99%) produces more accurate classifications, it still does not do so at nominal levels with the specifications used in this study. The average PCC for a nominal accuracy of 99% was 95.59%, while the average PCC for a nominal accuracy of 95% was 94.16%.

Table 1: ATL and PCC as a function of cutscore location and nominal accuracy

Nominal %	*Data*	*Cutscore*					*Total*
		-1	*-0.5*	*0*	*0.5*	*1*	
99%	ATL	30.35	37.30	41.00	35.68	26.61	34.19
	PCC	95.63	95.05	94.72	95.58	96.98	95.59
95%	ATL	14.61	19.94	21.51	18.98	13.99	17.81
	PCC	94.67	93.77	92.82	93.91	95.63	94.16
Total	ATL	22.48	28.62	31.26	27.33	20.30	26.00
	PCC	95.15	94.41	93.77	94.75	96.31	94.88

Table 2 presents the same results, separated by termination criterion. The same trend regarding cutscore location is evident, as is the fact that observed accuracy is not necessarily near nominal accuracy. This is especially true for the SPRT, which has lower ATL than the CLR or ACI, but also lower PCC.

Table 2: ATL and PCC for each termination criterion

Nominal %	*Termination*	*Data*	*Cutscore*					*Total*
			-1	*-0.5*	*0*	*0.5*	*1*	
99	ACI	ATL	44.63	69.06	81.77	71.53	53.67	64.13
		PCC	94.15	95.61	95.78	96.92	97.61	96.01
	SPRT	ATL	20.16	19.97	20.52	18.02	13.86	18.51
		PCC	95.67	94.47	93.97	94.75	96.52	95.08
	CLR	ATL	46.62	57.52	61.69	52.83	37.81	51.29
		PCC	96.96	96.25	95.90	96.76	97.74	96.72
95	ACI	ATL	22.68	45.56	53.26	45.95	32.03	39.90
		PCC	93.21	95.05	95.14	95.73	96.53	95.13
	SPRT	ATL	10.79	10.98	11.73	10.39	7.85	10.35
		PCC	94.77	92.93	91.94	93.24	95.19	93.61
	CLR	ATL	18.01	21.21	19.07	17.81	14.37	18.09
		PCC	95.86	95.01	93.16	94.09	96.05	94.83

The reason for the lower overall ATL and PCC with the SPRT is the width of the indifference region. As mentioned above, a wider indifference region will cause the ratio to diverge quickly, making a decision after fewer items. With a cutscore of 0.0 and $\delta = 0.4$, PCC dropped to only 89.22%, with a nominal accuracy of 95%.

Table 3: ATL and PCC for each level of nominal % and δ

Nominal %	*Termination*	*Data*	*Cutscore*					*Total*
			-1	*-0.5*	*0*	*0.5*	*1*	
99	0.2	ATL	33.64	32.49	33.04	28.87	21.72	29.95
		PCC	95.90	95.42	95.33	95.75	97.20	95.92
	0.3	ATL	16.60	16.82	17.09	15.72	11.67	15.58
		PCC	96.32	94.49	94.01	95.02	96.66	95.30
	0.4	ATL	10.24	10.60	11.43	9.46	8.20	9.99
		PCC	94.80	93.50	92.56	93.47	95.70	94.01
95	0.2	ATL	17.01	17.63	18.15	16.34	12.20	16.27
		PCC	96.19	95.01	94.42	95.23	96.88	95.55
	0.3	ATL	9.66	9.47	10.32	8.73	6.53	8.94
		PCC	94.77	93.15	92.18	93.65	95.09	93.77
	0.4	ATL	5.70	5.85	6.73	6.09	4.83	5.84
		PCC	93.34	90.64	89.22	90.83	93.60	91.53

Conclusions

The results of this study demonstrate the importance of conducting simulations before the production of examinations, do determine if observed accuracy can be expected to be near nominal levels. Without such simulations, it is likely that a 99% nominal accuracy could be specified, but actual accuracy is much lower, translating to classification error rates that are much higher than is acceptable. Observed classification accuracy is a function of several variables, including the nominal accuracy, the location of the cutscore relative to the examinee distribution, and the information structure of the item bank.

An important parameter that deserves attention is the width of the indifference region. The arbitrariness introduced by δ is one reason that a CCT termination criterion without such a parameter, such as the CLR or ACI, is desirable. However, the CLR and ACI do no necessarily produce observed PCC equal to nominal levels, either. Further research must be conducted to determine a termination methodology that involves as little arbitrariness and produces observed accuracy near nominal levels. Otherwise, this arbitrariness is present in all CCT research, even on topics like item selection that do not evaluate termination criteria.

Because of this issue, if the SPRT is to be used operationally, great care must be taken to determine an appropriate value of δ. Ideally, simulation studies should be conducted by the sponsoring organization using realistic data that reflect the actual structure of the item bank and examinee population.

In conclusion, it is dangerous to assume that observed CCT accuracy will be at nominal levels selected during the test development process. It is important for the test developer to consider relevant variables, and hopefully conduct a study similar to this one for each application to ensure that observed accuracy is acceptable to all stakeholders.

References

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

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.

Kalohn & Spray, (1998). Effect of item selection on item exposure rates within a computerized classification test. Paper presented at the Annual Meeting of the National Conference in Education, San Diego, CA.

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.

Lau, C. A., & Wang, T. (1998). *Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Lau, C. A., & Wang, T. (1999). *Computerized classification testing under practical constraints with a polytomous model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

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.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Parshall, Spray, Kalohn, & Davey, 2006

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.

Thompson, N.A. (2006). Variable-length computerized classification testing with item response theory. *CLEAR Exam Review*, 17(2).

Thompson, N.A. (2007a). A Comparison of two methods of polytomous computerized classification testing for multiple cutscores. Unpublished doctoral dissertation.

Thompson, N.A. (2007b). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, 12(3). Available online: http://pareonline.net/getvn.asp?v=12&n=3

Thompson, N.A., & Ro, S. (2007). Computerized Classification Testing with composite hypotheses. Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

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