ACT Research Report Series 2000-8

Effects of Item-Selection Criteria on Classification Testing with the Sequential Probability Ratio Test

Chuan-Ju Lin

Judith Spray

ACT

June 2000

For additional copies write:

ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Effects of Item-Selection Criteria on Classification Testing with the Sequential Probability Ratio Test

Chuan-Ju Lin
Judith Spray

ABSTRACT

This paper presents comparisons among three item-selection criteria for the sequential probability ratio test. The criteria were compared in terms of their efficiency in selecting items, as indicated by average test length (ATL) and the percentage of correct decisions (PCD). The item-selection criteria applied in this study were the Fisher information function, the Kullback-Leibler information function, and a weighted log-odds ratio. We also examined the effects of the cutoff scores, the width of the indifference region, the item pool size, and the item exposure rate under the different item-selection criteria. The results of the computer simulations showed that the three criteria yielded very small differences in the outcome measures, regardless of the conditions imposed.

EFFECTS OF ITEM-SELECTION CRITERIA ON CLASSIFICATION TESTING WITH THE SEQUENTIAL PROBABILITY RATIO TEST

Introduction

Computerized adaptive testing (CAT) is receiving more attention and has been applied more commonly over the last few years. Adaptive testing can yield more efficient tests by saving testing time (i.e., shorter tests) and increasing measurement precision. If the purpose of a test is to classify examinees into one of two or more mutually exclusive categories rather than estimating ability levels, the CAT procedure can be applied to make efficient decisions of classification by selecting and administering optimal items with algorithms based on statistical hypothesis testing, such as the sequential probability ratio test or SPRT (Spray & Reckase, 1994, 1996). The main purpose of this study was to compare three item-selection criteria in terms of average test length (ATL) and percentages of correct decisions (PCD) in the context of item selection with the SPRT. Variables hypothesized to affect ATL and PCD included the choice of the item-selection criteria, position of cutting points on the ability metric, the width of the indifference region, item pool size, and item exposure rate. Three types of selection criteria, three different cutting points, 11 indifference regions, two different item pool sizes, and three item exposure rates were examined.

The SPRT

Wald's (1947) SPRT has been applied for classifying examinees into two mutually exclusive categories using a computerized adaptive test (Eggen, 1999; Spray & Reckase, 1996). In order to distinguish the computerized SPRT from conventional CAT, the SPRT is usually regarded as a computerized classification test or CCT (Spray, Abdel-fattah, Huang, & Lau, 1997). In criterion-referenced testing situations, it is necessary to decide between two hypotheses, H_1 and H_2, which can be written arbitrarily as

$$H_1: \theta \le \theta_0 - \delta = \theta_1$$
$$\text{vs.}$$
$$H_2: \theta \ge \theta_0 + \delta = \theta_2,$$

where θ represents the ability of an examinee, θ_0 is a given cutting point or passing criterion, θ_1 and θ_2 refer to the lower and upper bounds, respectively (i.e., we assume that $\theta_2 > \theta_1$), of a particular decision threshold, and where δ forms a small region, called an indifference region, on both sides of the cutting point. The width of the indifference region or interval of θ_2 - θ_1 usually equals 2δ.[1] Two decision error rates, α (i.e., type I error rate or false positive) and β (i.e., type II error rate or false negative) can be defined as follows: P(choosing $H_2 | H_1$ is true) = α vs. P(choosing $H_1 | H_2$ is true) = β. The test statistic used in SPRT is a likelihood ratio, which is a ratio of the likelihood functions under the alternative (H_2) and null hypotheses (H_1), or

$$LR(\underline{x}) = \frac{L(\theta_2;\underline{x})}{L(\theta_1;\underline{x})} = \frac{\prod_{i=1}^{k} L(\theta_2;x_i)}{\prod_{i=1}^{k} L(\theta_1;x_i)} = \frac{\prod_{i=1}^{k} p_i(\theta_2)^{x_i}[1-p_i(\theta_2)]^{1-x_i}}{\prod_{i=1}^{k} p_i(\theta_1)^{x_i}[1-p_i(\theta_1)]^{1-x_i}}, \tag{1}$$

where L denotes the likelihood function, k represents the number of items or the test length, $\underline{x}$ contains observed dichotomous item responses, $x_1, x_2, . x_i, \ldots, x_k$, and $p_i(\theta_1)$ and $p_i(\theta_2)$ define the probabilities of a correct response to item i, conditional on θ_1 and θ_2. Equation (1) indicates that the higher the ratio, the more likely an examinee would be above the cutting point; the smaller the ratio, the more likely an examinee would be below the cutting point. According to Wald (1947), the nominal error rates, α and β, can be determined before test administration because

[1] The width of the indifference region around θ_0 need not be symmetrical (i.e., need not be equal to 2δ).

the upper and lower bounds of the likelihood ratio test are defined as functions of α and β. The actual observed error rates, α^* and β^*, may be different from those predetermined, where usually $\alpha^* \leq \alpha / (1 - \beta)$ and $\beta^* \leq \beta / (1 - \alpha)$. With the specified nominal error rates, the decision (or stopping) rules used can be defined as follows (Wald, 1947):

Continue selecting another item when:	$\beta / (1 - \alpha) < LR(\underline{x}) < (1 - \beta) / \alpha$;
Accept H_1 when:	$LR(\underline{x}) \leq \beta / (1 - \alpha)$;
Accept H_2 when:	$LR(\underline{x}) \geq (1 - \beta) / \alpha$.

Any test administered using SPRT is adaptive in terms of test length. The items are administered, one by one, to an examinee until a classification decision is made, so that examinees with different ability levels obtain different average lengths of tests. Examinees with ability $\theta_1 < \theta < \theta_2$ are expected to have longer tests than those with ability $\theta \leq \theta_1$ or $\theta \geq \theta_2$, because it is more difficult to make decisions about those examinees with ability levels in the indifference region, especially those near the cutting score.

In practice, a minimum and maximum test length are usually specified. Even though a decision may not be achieved after the specified maximum number of items have been administered from the item pool, a forced classification can be made: reject H_1 if $LR(\underline{x})$ is greater than the midpoint of the interval $[\beta / (1 - \alpha), (1 - \beta) / \alpha]$; otherwise accept H_1.

Item-selection criteria

(Fisher) Item Information

In computer-based classification tests, the items in the item pool are usually ranked from maximum to minimum in terms of some item-selection criteria at the specified cutting point.

Fisher (item) information is the item-selection criterion that is most often used and is defined for item i as (Eggen, 1999)

$$I_i(\theta) = \mathrm{E}\left(\frac{\frac{\partial}{\partial\theta} L(\theta; x_i)}{L(\theta; x_i)} \right)^2 . \tag{2}$$

The three-parameter logistic model (3-PL) is defined as follows:

$$p_i(\theta) = c_i + \frac{(1 - c_i)}{1 + \exp\{-1.7a_i(\theta - b_i)\}}. \tag{3}$$

The term, $p_i(\theta)$, represents the probability of a correct response to item i (i.e., the 3-PL) and a_i, b_i, and c_i are item parameters. Equation (2) may be rewritten (Lord, 1980) for the three-parameter logistic model (3-PL) as:

$$I_i(\theta) = \frac{1.7^2 a_i^2 [1 - p_i(\theta)][p_i(\theta) - c_i]^2}{[1 - c_i]^2 p_i(\theta)}. \tag{4}$$

Within the context of a computerized adaptive test for classification with the SPRT procedure, items with the largest Fisher information at the cutting point are selected for administration first.

Kullback-Leibler (K-L) Information

Another item-selection criterion is Kullback-Leibler (K-L) information, which is a concept somewhat related to SPRT. The K-L information is a measure of the difference between the two likelihood functions and is indicative of the expected information for discriminating between the two functions. In theory, the larger the K-L information, the earlier the test is terminated based on the SPRT criterion. The K-L information function for an item is defined as follows (Eggen, 1999):

$$K_i(\theta_2 \| \theta_1) = \mathrm{E}_{\theta_2} \log\left(\frac{L(\theta_2; x_i)}{L(\theta_1; x_i)} \right), \tag{5}$$

where $K_i(\theta_2||\theta_1)$ denotes an item information index for item i for any two θ values (θ_2 and θ_1), and E is the expected value operator, taken relative to θ_2. The K-L test information function (i.e., $K(\theta_2||\theta_1)$) is the sum of the K-L information functions over all k items in the test, which equals

$$K(\theta_2\|\theta_1) = \sum_{i=1}^{k} K_i(\theta_2\|\theta_1). \tag{6}$$

The items with maximum K-L information are selected sequentially. The discrepancy between the likelihood function under the null and alternative hypotheses is a maximum when the K-L information is maximized. Therefore, testing is expected to be quite efficient because K-L information is, itself, a likelihood ratio; thus, the number of items needed to make decisions is expected to be minimized. With the dichotomously-scored IRT model, K-L item information can be computed as:

$$K_i(\theta_2\|\theta_1) = p_i(\theta_2)\log\frac{p_i(\theta_2)}{p_i(\theta_1)} + q_i(\theta_2)\log\frac{q_i(\theta_2)}{q_i(\theta_1)}, \tag{7}$$

where $p_i(\theta_2)$ and $p_i(\theta_1)$ are the probabilities of a correct response to item i at θ_2 and θ_1, respectively, and $q_i(\theta_2)$ and $q_i(\theta_1)$ are the complement probabilities.

Weighted Log-odds Ratio

An alternative measure on which to rank items for selection using the SPRT procedure is a weighted log-odds ratio criterion. This value is based on the following premise:

The likelihood ratio, $LR(\underline{x})$, is equal to 1.θ at the beginning of the testing session. The value, $p_i(\theta_2)/ p_i(\theta_1)$, is multiplied to the likelihood ratio if the item is answered correctly or when x = 1. Likewise, $LR(\underline{x})$ is multiplied by $q_i(\theta_2)/q_i(\theta_1)$ when x = θ, or when the item is answered incorrectly. As testing continues, $LR(\underline{x})$ is compared to the two boundaries, β / (1 - α) and

$(1 - \beta) / \alpha$, to determine if testing should terminate or another item administered. $LR(\underline{x})$ will make its largest gains (and therefore move closer to a boundary most quickly) whenever $p_i(\theta_2)/ p_i(\theta_1)$ or $[q_i(\theta_2)/ q_i(\theta_1)]^{-1}$ is greatest. This also implies that items with the steepest slopes of $p_i(\theta)$ between θ_2 and θ_1 will be best at discriminating between pass and fail status. Therefore, it is desirable to find items in the pool with the largest values of $p_i(\theta_2)/ p_i(\theta_1)$ when x = 1, **and** those with the largest values of $[q_i(\theta_2)/ q_i(\theta_1)]^{-1}$ when x = θ.

In other words, it is desirable to locate items with the largest values of

$$\left(\frac{p_i(\theta_2)}{p_i(\theta_1)}\right)^X \div \left(\frac{q_i(\theta_2)}{q_i(\theta_1)}\right)^{1-X}. \tag{8}$$

Because θ_j for the j^{th} examinee is neither known nor estimated, the expected value of (8) or the expected value of the log of (8), where the expectation is taken over the entire population of examinees, is considered, or

$$E_\theta\left\{\left(\frac{p_i(\theta_2)}{p_i(\theta_1)}\right)^X \div \left(\frac{q_i(\theta_2)}{q_i(\theta_1)}\right)^{1-X}\right\}.$$

Equivalently, we want to find items with large values of

$$E_\theta\left(\log\left\{\left(\frac{p_i(\theta_2)}{p_i(\theta_1)}\right)^X \div \left(\frac{q_i(\theta_2)}{q_i(\theta_1)}\right)^{1-X}\right\}\right), \text{ or}$$

$$E_\theta(X)\left\{\log\left(\frac{p_i(\theta_2)}{p_i(\theta_1)}\right)\right\} - E_\theta(1-X)\left\{\log\left(\frac{q_i(\theta_2)}{q_i(\theta_1)}\right)\right\}. \tag{9}$$

This also can be written as

$$E_\theta(X)\{\log p_i(\theta_2) - \log p_i(\theta_1)\} - E_\theta(1-X)\{\log q_i(\theta_2) - \log q_i(\theta_1)\}, \tag{10}$$

where $E_\theta(X) = \int_{-\infty}^{\infty} p(\theta|X)g(\theta)d(\theta)$, the expected p-value for this item.

The rationale for using this value to select items within the SPRT framework is that we are searching for items that will cause the SPRT likelihood ratio to cross the decision boundaries, $(1-\beta)/\alpha$ and $\beta/(1-\alpha)$, or $\log[(1-\beta)/\alpha]$ and $\log[\beta/(1-\alpha)]$, most quickly. Therefore, it makes sense to find the value of (10) for all items in the item pool. Thus, in theory, those items with greater weighted log-odds ratios should be selected earlier so that a decision will be made as soon as possible with the fewest number of items.

Item Exposure Control

With computerized adaptive testing, the *best* items will be frequently selected, which is undesirable for test security reasons. Therefore, in order to protect the item pool, many item-exposure control strategies have been developed (e.g., Davey & Parshall, 1995; McBride & Martin, 1983; Sympson & Hetter, 1985). Item-exposure control is not only an important issue in CAT but also in CCT. Within the context of the current study, the *best* or optimal items refer to those with the best criterion values (e.g., highest Fisher information) at the cutting point. Without item-exposure control, the item-overlap rate between two CCT examinations would be very high because optimal items would be selected first in the test administration sequence and would eventually lead to overexposure.

A randomization scheme is a typical approach to controlling item exposure for CCT examinations (Spray et al., 1997; Way, Zara, & Leahy, 1996), especially in simulation studies This approach for CCT is similar to the *5-4-3-2-1* randomization procedure used in CAT for ability estimation (McBride & Martin, 1983). The randomization methods indirectly control item exposure by randomly selecting an item from a group of a particular number (e.g., m) of items. This usually results in longer tests to achieve the desired measurement precision, a necessary trade-off to protect the integrity of an item pool and the validity of a test.

With CCT randomization, an item is randomly selected from a group of *m* top-ranked items. All items in the pool are ordered based on the magnitude of the item-selection criterion (from maximum to minimum) at the specified cutting point(s). A *stack* of items is thus ranked at the cutting score with *m* items in a cell. For example, the top five items are grouped into the first cell of the stack, the second top five into the second cell, and so on. The first item administered to an examinee is one that is randomly selected from the first cell, the second item from the second cell, and so on. If the stack is exhausted for a particular examinee, the algorithm will continue selecting items from the top of the stack, avoiding those items that have been administered previously.

Purpose of Study

The traditional SPRT item-selection criterion of choosing items that provide the most Fisher item information at the cutting score, θ_0, may be questionable because the SPRT does not depend on θ_0. Because the location of θ_0 within 2δ is arbitrary, it has been hypothesized that using selection criteria that are functions of θ_2 and θ_1 might produce better results than the use of the traditional Fisher information at θ_0, especially as the width of the indifference region, 2δ, increases.

Eggen (1999) conducted a study concerning the effects of Fisher and Kullback-Leibler information with the SPRT procedure on two- and three-category classification problems and found that item-selection procedures based on maximum K-L information performed as well as those based on Fisher information in terms of testing efficiency and classification errors The purpose of the current study was to investigate the efficiency of these two item-selection criteria more thoroughly by including several manipulations hypothesized to maximize possible differences in the criteria, as well as to include the weighted log-odds ratio criterion in the comparison.

Method

Item Pools

This study utilized two sizes of item pools – a *whole* pool and a *half* pool. The *whole item pool* used in this study was the ACT Assessment Mathematics Usage Test containing six equivalent (i.e., previously administered, intact) test forms. Each form was composed of 60 items, and thus, 360 items comprised the pool. Although two dimensions have been identified for each form based on previous multidimensional studies, the unidimensional SPRT procedure can be used with this item pool because it is robust to the violation of the unidimensionality assumption (Spray et al., 1997). The items were calibrated with the 3-PL IRT model.

In addition to using the whole pool, the item pool was split into two similar pools, each of which included three equivalent test forms and, thus, 180 items. One of these smaller pools was subsequently used for this study and was labeled as the *half pool*.

Item-selection Criteria

Three item-selection criteria or functions were used for item selection:

1. Fisher information function.
2. Kullback-Leibler information function.
3. Weighted log-odds ratio.

Design

In this study, the randomization scheme was used to control item-exposure rate, and different *stratum depths* were used. A stratum depth referred to the number of items grouped together to yield the stratum within the randomization scheme. The minimum test length was set to one, and the maximum test length was set to 360 for the whole pool, and to 180 for the half pool situations (i.e., there were no test-length constraints for these simulations).[2] The five conditions under which the effects of item-selection criteria on ATL and PCD were investigated were as follows:

1. Whole pool, stratum depth = 1 item (i.e., no exposure control).
2. Whole pool, stratum depth = 5 items.
3. Whole pool, stratum depth = 10 items.
4. Half pool, stratum depth = 1 item (i.e., no exposure control).
5. Half pool, stratum depth = 5 items.

Simulation Procedure

The comparisons among the different item-selection criteria were conducted through a simulation study. The simulations were performed as follows: (1) a simulee with ability θ was randomly selected from a standard normal distribution, N(0,1); (2) based on the SPRT, items were administered sequentially to a simulee using one of the three item-selection criteria, and the response vector for a simulee was generated by comparing $p_i(\theta)$ to a random deviate (e.g., u) drawn from a uniform [0,1] distribution. If $p_i(\theta) \geq u$, the item was scored as correct; otherwise, it was scored as incorrect; (3) the same procedure was then repeated for 100,000 simulees.

For this study, α and β were .05. The cutting points (i.e., θ_0) and δ (i.e., half the distance between θ_1 and θ_2) varied within the item-selection procedures:

θ_0 = -.32, .81, and 1.79, which corresponded to proportion-correct scores of .41, .61, and .82, and $.20 \leq \delta \leq .30$ with increments of .01. The various item-selection procedures were then

[2] It was thought that the possibility of finding differences among the three different selection criteria might be maximized if the tests were allowed to run without length constraints.

compared on the outcome variables, average test length (ATL) and the percent of correct decisions made (PCD). Therefore, there were 99 possible conditions (i.e., 3 information criteria times 3 θ_0 values times 11 δ values) under each of five combinations of pool size and exposure-control conditions listed previously.

Results and Discussion

The ATL and PCD for three item-selection criteria with various indifference regions under five conditions are presented in Tables 1-5. It appeared that, for a particular cutting score with a given δ, there were almost no differences in either ATL or PCD among the three item-selection criteria. This was especially surprising when δ was largest around the cutting point, θ_0 (i.e., when θ_1 and θ_2 were farthest apart). See Table 1 vs. 2 vs. 3 and Table 4 vs. 5.

These tables also showed several expected results, namely that (1) as θ_0 moved farther away from the mean of the θ distribution, the PCD increased; (2) ATL increased when δ decreased and when item-exposure control increased (i.e., when a larger stratum depth was used); and (3) when a smaller pool was used, the ATL and PCD decreased. The latter finding resulted from more optimal items being administered more frequently under the half-pool condition (and, thus, the test lengths were shorter for all simulees). However, those simulees near the cutting point were missclassified at slightly higher rates **because** of the shortened test lengths. Thus, a decrement in classification accuracy occurred.

Further evidence of the similar behavior of the three item-selection criteria was exhibited by the rank correlations of the items at the cutting point. Table 6 provides the rank-order correlations among the three criteria and for three values of δ (representing small, medium, and large indifference regions) at the three different cut-off scores. All of the correlation coefficients were greater than .832, which indicated that there were not substantial differences in the rank order of items for the three selection criteria.

See Tables 1-6 at end of report

In terms of these simulation results, there was no evidence indicating that Fisher information, K-L information and weighted log-odds ratio performed differently on item selection with SPRT for two-category decision problems. Thus, the current practice of selecting items via the "maximum (Fisher) information at the cutting score criterion" appears to have been validated by these results. Nevertheless, some factors, such as content balancing not incorporated in the present study might have some effects on item selection and yield different results. Content-balancing issues should be considered in future studies.

References

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Eggen, T. (1999). Item selection in adaptive testing with the Sequential Probability Ratio Test. *Applied Psychological Measurement*, 23(3), 249-261.

Lord, F. M. (1980). *Applications of item response theory to practical problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York, Academic Press.

Spray, J. A., Abdel-fattah, Abdel-fattah A., Huang, C. Y., & Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional.* (Research Report 97-5). Iowa City, IA: American College Testing.

Spray, J. A. & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

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 and Behavioral Statistics,* 21(4), 405-414.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the 17th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

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

Way, W. D., Zara, A. R., & Leahy, J. (1996, April). *Modifying the NCLEX CAT item selection algorithm to improve item exposure.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

TABLE 1
Average Test Length and Percentage of Correct Decisions for All Possible Item Selection Procedures with Whole Pool and Stratum Depth = 1.

		$\theta_0 = -.32$		$\theta_0 = .81$		$\theta_0 = 1.79$	
		ATL	PCD	ATL	PCD	ATL	PCD
$\delta = .20$	Fisher	64.66	.945	16.64	.969	6.28	.991
	LR	65.74	.946	16.70	.969	5.86	.991
	K-L	64.72	.946	16.48	.968	6.37	.991
$\delta = .21$	Fisher	60.49	.945	15.38	.967	5.99	.991
	LR	60.42	.945	15.04	.968	5.48	.991
	K-L	61.39	.944	15.40	.967	6.02	.991
$\delta = .22$	Fisher	56.47	.945	14.10	.967	5.69	.991
	LR	56.59	.944	13.98	.967	5.23	.991
	K-L	56.48	.943	13.98	.966	5.69	.991
$\delta = .23$	Fisher	52.49	.944	13.08	.965	4.54	.990
	LR	53.14	.943	12.48	.965	4.81	.990
	K-L	52.61	.943	12.96	.965	4.49	.991
$\delta = .24$	Fisher	49.48	.942	11.83	.963	4.26	.991
	LR	49.93	.941	11.72	.964	4.52	.990
	K-L	49.73	.941	12.06	.963	4.30	.990
$\delta = .25$	Fisher	46.21	.941	10.64	.962	3.85	.990
	LR	45.99	.941	11.00	.963	4.36	.990
	K-L	46.43	.940	11.18	.963	3.89	.990
$\delta = .26$	Fisher	43.72	.939	10.10	.962	3.70	.989
	LR	43.26	.940	10.41	.961	4.09	.990
	K-L	44.01	.939	10.49	.962	3.71	.989
$\delta = .27$	Fisher	40.14	.938	9.73	.959	3.49	.989
	LR	40.43	.937	9.59	.960	3.92	.990
	K-L	40.23	.936	9.26	.961	3.51	.988
$\delta = .28$	Fisher	37.67	.936	8.96	.959	3.29	.989
	LR	37.60	.936	8.88	.959	3.03	.989
	K-L	37.52	.937	8.86	.958	3.28	.989
$\delta = .29$	Fisher	35.34	.935	8.42	.958	3.14	.988
	LR	35.23	.935	8.58	.958	2.87	.989
	K-L	35.36	.936	8.55	.958	3.19	.988
$\delta = .30$	Fisher	33.15	.935	8.03	.956	2.99	.988
	LR	33.28	.934	8.05	.955	2.76	.988
	K-L	33.47	.933	8.18	.957	3.05	.988

Note: Fisher: Fisher Information
LR: Weighted Log-Odds Ratio
K-L: Kullback-Leibler Information
ATL: Average Test Length
PCD: Percentage of Correct Decisions

TABLE 2
Average Test Length and Percentage of Correct Decisions for All Possible Item Selection Procedures with Whole Pool and Stratum Depth = 5.

		$\theta_0 = -.32$		$\theta_0 = .81$		$\theta_0 = 1.79$	
		ATL	PCD	ATL	PCD	ATL	PCD
$\delta = .20$	Fisher	101.18	.944	29.08	.968	11.13	.991
	LR	100.64	.945	29.16	.968	10.67	.991
	K-L	100.05	.946	29.51	.968	11.41	.991
$\delta = .21$	Fisher	95.12	.944	27.22	.967	10.48	.991
	LR	95.11	.944	27.19	.966	10.07	.991
	K-L	94.93	.945	27.38	.967	10.62	.991
$\delta = .22$	Fisher	89.98	.945	25.18	.966	9.87	.991
	LR	89.74	.944	25.13	.966	8.94	.991
	K-L	89.77	.944	25.47	.966	9.88	.990
$\delta = .23$	Fisher	85.28	.942	23.66	.964	9.25	.990
	LR	85.70	.942	23.38	.964	8.38	.990
	K-L	84.92	.943	23.58	.964	9.36	.990
$\delta = .24$	Fisher	80.68	.941	21.92	.964	8.63	.990
	LR	80.90	.941	21.84	.963	7.96	.990
	K-L	80.39	.940	22.20	.963	8.59	.990
$\delta = .25$	Fisher	76.58	.940	20.63	.962	7.98	.989
	LR	76.77	.941	20.31	.962	7.38	.990
	K-L	75.83	.941	20.44	.961	8.32	.989
$\delta = .26$	Fisher	72.72	.938	19.01	.960	7.63	.988
	LR	72.45	.939	19.10	.961	6.99	.990
	K-L	72.73	.939	19.23	.961	7.50	.989
$\delta = .27$	Fisher	69.21	.938	17.76	.959	7.22	.989
	LR	69.00	.937	17.63	.960	6.60	.989
	K-L	68.81	.939	18.00	.959	7.15	.989
$\delta = .28$	Fisher	65.66	.936	16.61	.957	6.56	.988
	LR	65.37	.936	16.63	.957	6.08	.989
	K-L	65.21	.937	17.02	.958	6.84	.988
$\delta = .29$	Fisher	62.57	.935	15.82	.956	6.40	.988
	LR	61.76	.936	15.66	.956	5.84	.989
	K-L	62.01	.934	15.83	.958	6.33	.988
$\delta = .30$	Fisher	58.98	.932	14.77	.955	5.88	.988
	LR	59.27	.932	14.85	.956	5.50	.988
	K-L	58.80	.933	14.86	.955	6.04	.988

Note: Fisher: Fisher Information
LR: Weighted Log-Odds Ratio
K-L: Kullback-Leibler Information
ATL: Average Test Length
PCD: Percentage of Correct Decisions

TABLE 3

Average Test Length and Percentage of Correct Decisions for All Possible Item Selection Procedures with Whole Pool and Stratum Depth = 10.

		$\theta_0 = -.32$		$\theta_0 = .81$		$\theta_0 = 1.79$	
		ATL	PCD	ATL	PCD	ATL	PCD
$\delta = .20$	Fisher	108.27	.945	34.67	.968	13.33	.991
	LR	108.47	.946	34.57	.968	12.43	.991
	K-L	108.28	.946	34.93	.968	13.50	.991
$\delta = .21$	Fisher	103.50	.945	32.34	.967	12.48	.990
	LR	103.08	.944	32.04	.966	11.65	.991
	K-L	102.64	.944	32.28	.967	12.73	.991
$\delta = .22$	Fisher	98.10	.943	30.01	.966	11.75	.991
	LR	97.79	.943	30.17	.965	10.90	.990
	K-L	97.39	.943	30.36	.967	11.67	.990
$\delta = .23$	Fisher	93.06	.942	28.37	.964	11.02	.991
	LR	92.72	.943	28.08	.965	10.33	.990
	K-L	93.67	.941	28.28	.965	11.21	.990
$\delta = .24$	Fisher	88.88	.941	26.45	.962	10.37	.989
	LR	88.95	.941	26.30	.964	9.60	.990
	K-L	88.63	.943	26.38	.963	10.52	.990
$\delta = .25$	Fisher	84.15	.941	24.79	.961	9.87	.989
	LR	84.82	.941	24.52	.962	9.21	.990
	K-L	84.78	.939	24.88	.962	9.81	.990
$\delta = .26$	Fisher	80.86	.940	23.36	.960	9.12	.989
	LR	80.37	.938	23.17	.960	8.52	.990
	K-L	80.59	.939	23.32	.961	9.31	.989
$\delta = .27$	Fisher	76.62	.937	21.77	.958	8.65	.988
	LR	76.76	.938	21.72	.958	8.10	.989
	K-L	76.49	.938	21.92	.959	8.80	.989
$\delta = .28$	Fisher	72.83	.937	20.69	.957	8.27	.988
	LR	73.23	.937	20.51	.957	7.58	.989
	K-L	72.93	.937	20.75	.958	8.21	.989
$\delta = .29$	Fisher	69.98	.935	19.35	.958	7.79	.988
	LR	69.93	.935	19.34	.956	7.08	.988
	K-L	69.85	.936	19.56	.957	7.96	.988
$\delta = .30$	Fisher	66.81	.933	18.49	.955	7.45	.988
	LR	66.69	.934	18.37	.955	6.90	.988
	K-L	66.11	.933	18.72	.954	7.66	.987

Note: Fisher: Fisher Information
LR: Weighted Log-Odds Ratio
K-L: Kullback-Leibler Information
ATL: Average Test Length
PCD: Percentage of Correct Decisions

TABLE 4
Average Test Length and Percentage of Correct Decisions for All Possible Item Selection Procedures with Half Pool and Stratum Depth = 1.

		$\theta_0 = -.32$		$\theta_0 = .81$		$\theta_0 = 1.79$	
		ATL	PCD	ATL	PCD	ATL	PCD
$\delta = .20$	Fisher	62.40	.927	22.80	.963	10.82	.989
	LR	62.35	.926	22.69	.963	9.60	.990
	K-L	62.14	.928	22.76	.963	11.07	.990
$\delta = .21$	Fisher	59.65	.926	21.30	.962	9.87	.989
	LR	58.96	.927	21.58	.963	9.06	.990
	K-L	58.99	.926	21.43	.964	9.82	.990
$\delta = .22$	Fisher	55.63	.926	20.36	.962	9.47	.989
	LR	56.09	.928	20.03	.963	8.75	.990
	K-L	55.91	.926	20.23	.962	9.40	.989
$\delta = .23$	Fisher	52.99	.927	18.54	.962	8.96	.990
	LR	53.20	.926	18.99	.961	8.46	.989
	K-L	53.06	.926	18.62	.960	9.02	.989
$\delta = .24$	Fisher	50.36	.926	17.42	.960	8.35	.989
	LR	50.80	.925	18.02	.960	8.07	.989
	K-L	50.27	.925	17.44	.960	8.46	.988
$\delta = .25$	Fisher	48.34	.925	16.66	.959	7.96	.989
	LR	47.85	.926	16.28	.960	7.64	.989
	K-L	48.16	.926	16.61	.958	8.13	.988
$\delta = .26$	Fisher	45.61	.926	15.65	.959	7.62	.988
	LR	45.51	.925	15.12	.959	7.52	.989
	K-L	45.83	.924	15.71	.959	7.78	.989
$\delta = .27$	Fisher	43.27	.924	14.74	.958	6.64	.989
	LR	43.36	.925	14.36	.956	6.40	.988
	K-L	43.38	.924	14.59	.959	7.50	.988
$\delta = .28$	Fisher	41.32	.924	13.68	.957	6.55	.988
	LR	41.09	.926	13.64	.958	6.20	.988
	K-L	41.13	.925	13.81	.957	7.29	.988
$\delta = .29$	Fisher	39.13	.922	12.94	.956	6.26	.987
	LR	39.19	.925	13.07	.955	5.79	.988
	K-L	39.22	.924	13.07	.955	7.15	.988
$\delta = .30$	Fisher	37.28	.922	12.32	.953	6.17	.987
	LR	37.15	.923	12.62	.955	5.62	.988
	K-L	37.14	.924	12.20	.954	6.92	.987

Note: Fisher: Fisher Information
LR: Weighted Log-Odds Ratio
K-L: Kullback-Leibler Information
ATL: Average Test Length
PCD: Percentage of Correct Decisions

TABLE 5
Average Test Length and Percentage of Correct Decisions for All Possible Item Selection Procedures with Half Pool and Stratum Depth = 5.

		$\theta_0 = -.32$		$\theta_0 = .81$		$\theta_0 = 1.79$	
		ATL	PCD	ATL	PCD	ATL	PCD
$\delta = .20$	Fisher	86.57	.926	34.86	.963	14.71	.990
	LR	86.88	.926	34.31	.963	13.78	.989
	K-L	86.51	.928	34.87	.964	14.78	.990
$\delta = .21$	Fisher	83.00	.927	33.14	.963	14.23	.989
	LR	83.05	.927	32.49	.962	12.88	.989
	K-L	83.30	.928	32.92	.963	14.24	.990
$\delta = .22$	Fisher	79.89	.927	31.06	.963	13.44	.989
	LR	79.94	.927	30.91	.961	12.04	.990
	K-L	80.00	.927	31.30	.961	13.47	.989
$\delta = .23$	Fisher	76.89	.928	29.50	.962	12.60	.990
	LR	76.98	.927	29.09	.962	11.60	.989
	K-L	76.60	.927	29.43	.961	12.58	.990
$\delta = .24$	Fisher	73.57	.927	27.95	.960	11.94	.989
	LR	73.93	.926	27.35	.960	11.02	.989
	K-L	73.62	.926	27.83	.960	12.21	.988
$\delta = .25$	Fisher	71.03	.926	26.42	.959	11.42	.988
	LR	70.98	.927	26.03	.959	10.55	.988
	K-L	71.20	.925	26.56	.959	11.49	.988
$\delta = .26$	Fisher	68.27	.925	24.93	.958	10.88	.988
	LR	68.08	.925	24.92	.958	9.90	.989
	K-L	68.26	.926	25.10	.958	10.95	.989
$\delta = .27$	Fisher	65.87	.924	23.57	.957	10.34	.988
	LR	65.59	.925	23.48	.957	9.48	.988
	K-L	65.56	.924	23.77	.958	10.41	.989
$\delta = .28$	Fisher	63.29	.924	22.40	.956	9.86	.988
	LR	63.55	.924	21.94	.956	9.04	.988
	K-L	63.61	.923	22.58	.957	9.94	.988
$\delta = .29$	Fisher	60.98	.924	21.17	.956	9.55	.987
	LR	60.86	.923	21.24	.955	8.54	.988
	K-L	60.84	.923	21.39	.955	9.54	.987
$\delta = .30$	Fisher	58.46	.922	20.24	.953	8.97	.988
	LR	58.30	.923	20.15	.954	8.08	.987
	K-L	58.79	.922	20.40	.955	9.23	.987

Note: Fisher: Fisher Information
LR: Weighted Log-Odds Ratio
K-L: Kullback-Leibler Information
ATL: Average Test Length
PCD: Percentage of Correct Decisions

TABLE 6
Rank Correlations among Three Selection Criteria for Three δ Values at Each of the Three Cutting Points (θ_0).

		$\delta = .20$		$\delta = .25$		$\delta = .30$	
		Fisher	K-L	Fisher	K-L	Fisher	K-L
θ_o=-.32	LR	.976	.994	.975	.832	.975	.929
	K-L	.983		.853		.950	
θ_o=.81	LR	.985	.979	.985	.978	.986	.976
	K-L	.999		.999		.998	
θ_o=1.79	LR	.910	.896	.910	.893	.910	.891
	K-L	.999		.999		.999	

Note: Fisher: Fisher Information
LR: Weighted Log-Odds Ratio
K-L: Kullback-Leibler Information