RUNNING HEAD: Item Selection in CCT

Item Selection in Computerized Classification Testing

Nathan A. Thompson

Prometric

Address:
Nathan A. Thompson
Prometric
1260 Energy Lane
St. Paul, MN 55108

Abstract

Several alternatives for item selection algorithms based on item response theory (IRT) in computerized classification testing (CCT) have been suggested (Spray & Reckase, 1994; Eggen, 1999; Lin & Spray, 2000; Weissman, 2004) with no conclusive evidence on the substantial superiority of a single method. It is argued that the lack of sizable effect is due to the fact that some of the methods actually assess the same concept through different calculations, and the simplest method is therefore the most appropriate. Moreover, the efficiency of item selection methods depend on the termination criteria that is used, which is demonstrated through didactic example and monte carlo simulation. Item selection at the cutscore, which seems conceptually appropriate for CCT, is not always the most efficient option.

Item Selection in Computerized Classification Testing

Computerized administration of psychoeducational tests offers many advantages. A well-documented advantage is the reduced number of items required per examinee when variable-length testing is applied. Variable-length testing refers to tests where not every examinee receives the same number of items; if the purpose of the test has been satisfied after only a small number of items, the test is concluded. The most common method of variable length testing, when it is used for ability/trait (θ) estimation, is computerized adaptive testing (Weiss & Kingsbury, 1984). When the purpose of the test is to assign an examinee to two or more mutually exclusive categories along the θ continuum, it is called computerized classification testing (CCT: Lin & Spray, 2000).

A major component within CCT that realizes the advantage of reduced test length is that of intelligent item selection algorithms. Contrary to a fixed-form conventional test, items are selected throughout the test, either in testlets (Luecht & Nungester, 1998) or after each individual item (Spray & Reckase, 1994; Lin & Spray, 2000). The item selection process can be termed "intelligent" because the computer attempts to evaluate which item remaining in the bank would be the "best" item to administer next, within certain constraints. Item selection methods differ in how they perform this evaluation.

Several intelligent item selection options have been suggested based on item response theory (IRT), including the maximization of item information at the current θ estimate (Reckase, 1983), the cutscore (Spray and Reckase, 1994; 1996), across a region of θ (Eggen, 1999), a log-odds ratio (Lin & Spray, 2000), and across θ (Weissman, 2004). However, these tend to fall into two general types: estimate-based (EB) and cutscore-based (CB).

The other major component of a CCT that produces shorter tests is the application of variable-length termination criteria. The termination criterion is an algorithm that determines whether the examinee is able to be classified within certain parameters at each point in the test. When the examinee is able to be classified, they are assigned a category and the test is terminated. There are three primary termination criteria in CCT. The sequential probability ratio test (SPRT; Wald, 1947; Eggen, 1999) formulates the decision process as a hypothesis test that the examinee's θ is equal to a specified point above the cutscore or another specified point below the cutscore. Ability confidence intervals (ACI; Thompson, 2006), originally termed adaptive mastery testing (Kingsbury & Weiss, 1983), terminate the test when a confidence interval for the examinee's θ is completely above or below the cutscore. Lastly, loss/utility structures from Bayesian decision theory can be used to assign classifications (Rudner, 2002). The decision theoretic approach often uses random item selection and classical test theory, so it is not considered here.

The purpose of this study is threefold:

(1) Demonstrate that item selection methods can be generally classified as EB or CB;
(2) Demonstrate that, within the EB and CB paradigms, various item selection methods assess the same concept;
(3) Demonstrate that EB selection is appropriate for ACI, and CB selection is appropriate for the SPRT.

Item Response Theory

The third necessary component of a CCT is the adoption of a psychometric model. While CCTs can be developed using classical test theory (Frick, 1992), methods discussed herein only make use of item response theory (IRT; Hambleton & Swaminathan, 1985; Embretson & Reise, 2000) for several reasons. First, item banks for large-scale testing programs are often calibrated with IRT. Second, CCT methods that use classical test theory assume that examinees can be neatly divided into categories such as “masters” and “nonmasters” before item calibration, which is not always possible. The ACI criterion for CCT and the item selection methods discussed in this study require that items be calibrated with IRT because IRT places items and examinees on the same scale. Lastly, the majority of research on CCT is based on IRT.

This study assumed that the data can be efficiently modeled with the three-parameter logistic model. The probability of an examinee with a given θ_j correctly responding to an item *i* is (Hambleton & Swaminathan, 1985, Eq. 3.3):

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

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.

D was assumed to be 1.0 in this study to make the example calculations simpler.

Item Selection Criteria

Although CCT research dates back to the 1960s (Ferguson, 1969), the first intelligent item selection method used in CCT was the maximization of Fisher information (FI) at the current estimate of θ (Reckase, 1983; Kingsbury & Weiss, 1983). FI is broadly defined as the conditional slope squared divided by the conditional variance, given the probability of a correct response to item *i* $P_i(\theta)$, or (Embretson & Reise, 2000, Eq. 7 A.1)

$$I_i(\theta) = P_i'(\theta)^2 \big/ P_i(\theta)(1 - P_i(\theta)). \qquad (2)$$

The information function for the three-parameter model is specifically defined as (Embretson & Reise, 2000, Eq. 7 A.2)

$$I_i(\theta) = \left[a_i^{\,2} \frac{1 - P_i(\theta)}{P_i(\theta)} \right] \left[\frac{(P_i(\theta) - c_i)^2}{(1 - c_i)^2} \right]. \qquad (3)$$

FI, while a function of θ, is evaluated at a single point for each item. Reckase (1983) and Kingsbury & Weiss (1983) chose the current θ estimate as this point. Spray and Reckase (1994) suggested that items instead be selected to maximize Fisher information at the cutscore, arguing that this makes more conceptual sense since the goal of the test is only to determine if an examinee is above or below that point.

Eggen (1999) advocated the use of information across a region of θ. This regional, rather than point, type of information is known as Kullback-Liebler information (KLI). KLI can be described as the expectation over observed responses x_i of the log-likelihood ratio for each item, or

$$K_i(\theta_2 \| \theta_1) = E_{\theta_1} \log \left[\frac{L_i(\theta_2 ; x_i)}{L_i(\theta_1 ; x_i)} \right] \qquad (4)$$

with θ_1 and θ_2 representing two points on θ chosen by the test user that define the region that KLI is calculated on and

$$L_i(\theta; x_i) = P_i^{x_i}(\theta)\left[1 - P_i(\theta)\right]^{1-x_i} \qquad (5)$$

denoting the likelihood function for the *i*th item. The double vertical bars are used in this context to emphasize that θ_1 and θ_2 are separated, and not viewed as the conditional relationship indicated by a single vertical bar. With a dichotomous IRT model (Eggen, 1999; Lin & Spray, 2000), this simplifies to

$$K_i(\theta_2 \parallel \theta_1) = P_i(\theta_2)\log\frac{P_i(\theta_2)}{P_i(\theta_1)} + Q_i(\theta_2)\log\frac{Q_i(\theta_2)}{Q_i(\theta_1)} \qquad (6)$$

where $P_i(\theta_2)$ is the probability of a correct response at θ_2 and $Q_i(\theta_2)$ is the complementary probability of an incorrect response. The values θ_1 and θ_2 can be chosen above and below the cutscore or the current estimate.

Lin and Spray (2000) developed another item selection criterion, selecting items by maximizing the log of the ratio of the item response probabilities at θ_1 and θ_2,

$$R = \left(\frac{P_i(\theta 2)}{P_i(\theta 1)}\right)^X \div \left(\frac{Q_i(\theta_2)}{Q_i(\theta_1)}\right)^{1-X} \quad . \qquad (7)$$

Lin and Spray suggest that θ_1 and θ_2 be chosen above and below the cutscore, but similar to KLI, it is also possible for them to be specified at an interval above and below the current θ estimate. While the log calculations performed by Lin and Spray were more involved, they were still dependent on maximizing the difference $P(\theta_2) - P(\theta_1)$.

The most general item selection method, mutual information (Weissman, 2004), evaluates information across all responses and all values in θ. It is equivalent to the KLI between the distributions of θ and *X*. This is expressed as

$$I(X_i, \Theta) = \sum_{x \in X} \sum_{\theta \in \Theta} f(x_i, \theta)\log\left[\frac{f(x_i, \theta)}{f(x_i)\,f(\theta)}\right]. \qquad (8)$$

Because the function in this situation is the probability of a response, the numerator of the bracketed ratio, the joint function of the item response and θ, is the item response function (IRF). The function $f(x_i)$ is, in psychometric notation, P(X = 1), the classical difficulty value. The function $f(\theta)$ is the assumed distribution of θ. Weissman proposed that this expression be evaluated at discrete points on θ, such as a point above and below a cutscore. The simulation in that study involved multiple cutscores, but if there is only one cutscore, mutual information can be reduced to KLI.

Conceptually, mutual information then quantifies the difference between the IRF and the classical difficulty across θ at the discrete points specified. An item with very low a_i results in IRF values that differ very little from the classical difficulty across θ. An item with high a_i results in low IRF values with low θ and high IRF values with high θ, resulting in higher mutual information. What makes mutual information different from a mere transformation of the a_i parameter is the application of an assumed θ distribution. Discrimination values equal, an item with a difficulty near the mean of the θ distribution will have greater mutual information than an item with extreme difficulty, because it provides its information to a greater proportion of examinees.

One drawback to mutual information is the assumption of a prior. If nothing is known concerning the examinee distribution, a uniform prior may be appropriate, analogous to the use of maximum likelihood θ estimation rather than Bayesian θ estimation procedures. If this were to be applied, mutual information would simply be a function of a_i. This was the approach used by Weissman (2004).

Weissman (2004) also suggested another new criterion for item selection in CCT, Fisher information with a weight function of the posterior θ distribution or likelihood function. Similar to other methods, this was evaluated at the current θ estimate by Weissman, but can also be evaluated at the cutscore. This method does not contribute anything more than FI because, at a given point such as the cutscore or current θ estimate, the FI for all remaining items in the bank is being multiplied by the same value – whatever is the weight for that θ value. Such multiplication does not change the ranking of the items in terms of information.

EB and CB Selection, and Equivalence Within Category

As evident from the outline of the evaluative processes above, each method can be designed to assess the item with regards to the cutscore or to the current θ estimate. For the three methods with a constrained evaluative locale (FI, KLI, log-odds ratio), the evaluative process takes place at either a single point or two endpoints of a region, and either approach can be specified to occur at the cutscore or θ estimate. Even mutual information, which evaluates information across any number of points on θ, can be specified with points relative to cutscore(s) or a θ estimate. For this reason, item selection methods in CCT can be broadly categorized as EB or CB. This categorization is outlined in Table 1.

Moreover, the three constrained methods are essentially equivalent, and tend to select the same item. All three assess the information provided by an item at the evaluative locale, cutscore or estimate, and simply difference in the calculation used to perform the evaluation. For example, suppose a CCT is designed to use CB item selection at the cutscore $\theta_c = 0.5$. FI will select the item with the highest information at 0.5, which is the item with the greatest a_i, and b_i nearest to 0.5. This same item will also provide the highest average information across a region around 0.5. Additionally, an item with these characteristics will also produce the greatest difference $P(\theta_2) - P(\theta_1)$ for a θ_1 below the cutscore and θ_2 above the cutscore, which maximizes Lin and Spray's (2000) criterion.

Consider the following example with three items, shown in Table 2. These three items have similar IRT parameters. Item 1 represents an appropriate item for administration at $\theta_c = 0.5$; the location parameter matches exactly, and the discrimination parameter is moderately high. Item 2 is also appropriate, but the location parameter is not exactly the same, while item 3 has the same location parameter as item 1 but slightly less discrimination. Even though the item parameter differences are quite small, the rankings of the items on the three item selection criteria are equivalent because they are assessing the same concept. Moreover, the values are proportionate with each method.

It is because of this fact that very little difference has been found in empirical studies. Eggen (1999) compared several formulations of FI and KLI, and found that the best formulations of each method were approximately equal in terms of ATL and PCC. Lin and Spray (2000) found that the log-odds ratio also performed similarly.

Weissman (2004) found that mutual information had higher PCC but lower ATL than FI, but used EB FI rather than CB FI. As described below, this is an unfair comparison when the SPRT is the termination criterion, as it was in the study. Furthermore, the difference was only evident for very short tests such as a maximum test length of 10, in which case EB FI selection is inefficient because the test does not have a chance to obtain an accurate θ estimate (Chang & Ying, 1996). Even so, the difference was only two to three items in terms of average test length.

In the interests of parsimony, since the three approaches produce the same ranking of items, the method with the least computational burden and specification of parameters should be used when designing a CCT.

Termination Criteria and Item Selection

As previously mentioned, there are three termination criteria available for CCT, with the SPRT and ACI the most commonly used in application. ACI classifies an examinee by estimating θ after each item in the test and constructing a confidence interval around the estimate $\hat{\theta}$ using the conditional standard error of measurement (SEM), if maximum likelihood estimation is applied, or the square root of the Bayesian posterior variance, if Bayesian estimation is applied. The confidence interval is represented mathematically as

$$\hat{\theta} - z_{\alpha}(SEM) < \theta < \hat{\theta} + z_{\alpha}(SEM)\,. \qquad (9)$$

where z_α is the normal deviate for a 1-α confidence interval, such as 1.96 for a 95% interval. If the confidence interval falls completely above the cutscore, the examinee can be classified as "pass." If the confidence interval falls completely below the cutscore, the examinee is classified as "fail." If the confidence interval contains the cutscore, another item is administered.

The SPRT structures the decision as a hypothesis test with $H_0\!: \theta = \theta_1$ and $H_1\!: \theta = \theta_2$. The likelihood of each hypothesis is compared in the form of a likelihood ratio:

$$L = \frac{L(\theta = \theta_2 \mid u)}{L(\theta = \theta_1 \mid u)} = \frac{\prod_{i=1}^{n} P_i(\theta_2)^X \, Q_i(\theta_2)^{1-X}}{\prod_{i=1}^{n} P_i(\theta_1)^X \, Q_i(\theta_1)^{1-X}}\,. \qquad (10)$$

This ratio is then compared to two decision points *A* and *B*. Wald (1947) suggested the following as approximations, assuming a nominal Type I error rate of α and Type II rate of β:

Lower decision point = $B = \beta/(1 - \alpha)$ (11)

Upper decision point = $A = (1-\beta)/\alpha$. (12)

If $L > A$, the examinee is classified as above the cutscore and the test is terminated. If $L < B$, the examinee is classified as below the cutscore. If $B < L < A$, another item is administered. Note that $\hat{\theta}$ is not involved.

The SPRT was originally applied to CCT with classical test theory item parameters (Ferguson, 1969). Reckase (1983) developed a procedure to apply item response theory to the specification of P_i and Q_i. Reckase suggested that two points θ_1 and θ_2 be chosen on the θ metric, where the value of θ_1 represents the lowest level that the test developer is willing to pass, while θ_2 represents the highest θ that the test developer is willing to fail. The space between the two is called the indifference region, and is often specified by adding and subtracting a user-defined value δ from the cutscore. For example, if the cutscore is 1.0 the test user could define, according to their interpretation of the testing situation, a small indifference region with $\delta = 0.1$ ($\theta_1 = 0.9$ and $\theta_2 = 1.1$) or a large indifference region with $\delta = 1.0$ ($\theta_1 = 0.0$ and $\theta_2 = 2.0$).

This introduces a certain amount of arbitrariness into the procedure, which is notable because the size of δ has a direct effect on the performance of the termination criterion. Because an IRT item response function is strictly increasing, a large value of δ will lead to a greater disparity between P_i (θ_1) and P_i (θ_2) if the item was answered incorrectly or Q_i (θ_1) and Q_i (θ_2) if the item was answered incorrectly. This in turn causes the value of the likelihood ratio to depart from 1.0 with fewer items.

The same reasoning applies to the efficiency of item selection criteria. Because the SPRT operates with regards to the cutscore only, while ACI is constructed with regards to $\hat{\theta}$, their respective information needs differ. With the SPRT, a decision is made more quickly when L is maximized or minimized by items that maximize the difference $P(\theta_2) - P(\theta_1)$. As demonstrated previously, this is equivalent to selecting items with the greatest FI at the cutscore. Conversely, ACI makes a decision more quickly when an accurate estimate of θ is obtained. This occurs when the conditional SEM or Bayesian posterior variance is minimized. As the SEM is inversely related to information at the current θ estimate (Embretson & Reise, 2000), ACI makes a decision more quickly when information is maximized at $\hat{\theta}$ rather than the cutscore.

Simulation Study

This interaction between item selection method and termination criterion was evaluated with a brief monte carlo simulation. CCTs were simulated for 10,000 examinees that were randomly generated from a N(0,1) distribution. The item bank consisted of 400 dichotomously scored items, with $a \sim N(1,0.2)$, $b \sim N(0,1)$, and $c = 0.25$. Because the SPRT introduces the additional arbitrariness in the parameter of IR width, the simulation was completed for the two ACI conditions, and then the IR width systematically varied until an approximately equivalent PCC was obtained. This allows a more direct comparison between ACI and the SPRT in terms of ATL. The IR width for the CB condition was 0.40, and the width for the EB condition was 0.39.

The interaction is evident in the ATL for each condition, presented in Table 3. For the SPRT, CB item selection used 1.44 *fewer* items on average. For ACI, CB item selection required 9.66 *more* items, on average, while also having slightly less classification accuracy.

No item exposure constraints were employed in this simulation, as past research is conclusive that this simply decreases differences between competing methods of other CCT aspects, or has negligible effects (Spray, Abdel-fattah, Huang, and Lau, 1997; Lau, 1998; Eggen, 1999; Eggen & Straetmans, 2000; Lin & Spray, 2000; Jiao, Wang, & Lau, 2004). This simulation followed Lin and Spray (2000), who specifically did not include constraints because they reduce the visibility of comparisons among other variables.

Discussion

This simulation, as well as the previous examples, demonstrates how a single item selection method is not the most efficient for all uses. Contrary to initial perception, cutoff item selection is not always appropriate in variable-length CCT. Because of the way that the SPRT and ACI utilize information differently in the classification of examinees, the most appropriate item selection method can vary. Specifically, CB item selection is more efficient when the SPRT is the termination criterion, and EB item selection is more efficient when ATL is the termination criterion. While it may be initially intuitive that CB item selection is more appropriate for all two-classification CCT because the test is only interested in if the examinee is above or below the cutoff, this is not true.

No previous research has made an even comparison with a complete crossing of item selection and termination criterion methods, but this simulation is supported by two earlier simulation studies. Eggen and Straetmans (2000) found that ATL was two to three items lower for EB selection than CB selection, with ACI as the termination criterion. Spray and Reckase (1994) supported CB selection with the SPRT but found little difference between CB and EB selection in terms of ATL with ACI. However, PCC was not investigated in that study.

Also note that CB item selection entails greater exposure of items with location parameters near the cutscores. Item exposure is less of an issue with EB selection; only the first few items tend to be the same for every examinee when no restrictions are imposed, whereas the entire set is the same for CB selection. If item exposure constraints were used in the current study, they would have a greater effect on ATL for CB selection than EB selection, for this reason. This would lead to more similar ATL for CB and EB selection with the SPRT, and increase the ATL advantage for EB selection with ACI. An evaluation of the extent of this effect offers a good target for future research, given the widespread use of exposure constraints in high-stakes testing.

Additional independent variables also offer opportunities for future research. Because of the different information needs of ACI and the SPRT, the shape of the item bank function must be appropriate. Namely, there must be sufficient information near the cutscore for CB selection with the SPRT, and there must be sufficient information across θ for EB selection with ACI. Efficiency would decrease with an inappropriate item bank.

Because item selection method has a direct effect on the efficiency of the examination, the specification of this aspect of CCT as argued herein has direct significance for practitioners. Depending on the remaining parameters of the CCT design, application of an item selection algorithm that is more appropriate for a given termination criterion will reduce the ATL for examinees, thereby modestly reducing test seat time and the required size of the item bank. Given that this is the intent of variable-length CCT, proper design of CCTs that enhances this effect is of practical importance.

Conversely, correct specification of item selection method increases the accuracy of decisions if test length is held constant. In high-stakes testing, accuracy is more important than efficiency, as an examinee is likely to be more upset with a perceived misclassification than with being administered more items. Because the utilization of information is important for maximizing classification accuracy, it should be given substantial consideration when a CCT is designed.

References

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing.*Applied Psychological Measurement, 20*, 213-229.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249-261.

Eggen, T.J.H.M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713-734.

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

Ferguson, R. L. (1969). The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh.

Frick, T.W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*(2), 187-213.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications.* Norwell, MA: Kluwer Academic Publishers.

Jiao, H., Wang, S., & Lau, C. A. (2004). An Investigation of Two Combination Procedures of SPRT for Three-category Classification Decisions in Computerized Classification Test. Paper presented at the annual meeting of the American Educational Research Association, San Antonio, April 2004.

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 test theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Lau, C. A. (1998). Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data. Unpublished doctoral dissertation, University of Iowa, Iowa City IA.

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

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.

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, April). An examination of decision-theory adaptive testing procedures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Spray, J. A., Abdel-fattah, A. A., Huang, C., and 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, Iowa: ACT, Inc.

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 national meeting of the National Council on Measurement in Education, 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 and Behavioral Statistics, 21*, 405-414.

Thompson, N.A. (2006). Variable-length computerized classification testing with item response theory. *CLEAR Exam Review, 17*(2), 13-18.

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361-375.

Weissman, A. (2004). Mutual information item selection in multiple-category classification CAT. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego CA.

Table 1: Classification of item selection methods

Approach	Point (FI)	Region (KLI)
Estimate-Based	1. Maximum FI at current estimate (Reckase, 1983)	1. KLI or MI in region around current estimate (not used yet) 2. Log-odds ratio around current estimate (not used yet)
Cutscore-Based	1. Maximum FI at cutscore(s) (Spray & Reckase, 1994)	1. KLI or MI in region around cutscore(s) (Eggen, 1999; Weissman, 2004) 2. Log-odds ratio (Lin & Spray, 2000)

Table 2: Example calculations

Item	*a*	*b*	*c*	P(θ = 0.0)	P(θ = 0.5)	P(θ = 1.0)	FI	KLI	P(θ = 1.0) - P(θ = 0.0)
1	1.00	0.50	0.25	0.5332	0.6250	0.7168	0.0844	0.0307	0.1836
2	1.00	0.60	0.25	0.5158	0.6063	0.6990	0.0824	0.0301	0.1832
3	0.90	0.50	0.25	0.5420	0.6250	0.7080	0.0683	0.0251	0.1660

Table 3: ATL results from simulation

Termination Criterion	Item Selection	ATL	PCC
SPRT	CB	13.63	96.61
	EB	15.07	96.70
ACI	CB	43.61	96.55
	EB	33.95	96.82