AMERICAN INSTITUTES FOR RESEARCH®

Oregon’s Adaptive Algorithm

DRAFT, Version 1.5

April 28, 2007

Introduction

The first state to bring its statewide testing on line, Oregon came to electronic testing with a history of adaptive tests on paper. For years, the Oregon state assessment system included multiple forms of each test, each targeting a different proficiency level. These tests were placed on a common scale, allowing educators a developmental approach to testing—they could administer tests at a comfortable difficulty level for students that would yield an informative result. When Oregon moved to on-line testing, it was natural that the test would be adaptive.

The move to a more standards-based environment brought changes in Oregon's testing program. Clear articulation of curricular content standards provided a stronger framework for the content of the test. Opportunities to test below grade-level were abandoned in favor of requiring grade-level tests that clearly tested the curriculum.

Currently, Oregon is redesigning the adaptive algorithm used for selecting items to be administered to examinees with the goals of

- improving fidelity to the content standards;
- improving fidelity to the test blueprint; and
- providing as much diagnostic information from the test as possible;

This document outlines an adaptive algorithm designed to meet these goals.

The algorithm described here is under development for implementation in the 2007-08 operational testing window. Additional revisions are underway that may introduce a significant simplification and provide improvements in efficiency; future versions of the algorithm may change to reflect these improvements.

Specific objectives

The item selection algorithm seeks to meet five objectives:

1. Ensure an appropriate distribution across content strands, or Score Reporting Categories (SRCs);
2. Ensure an appropriate distribution across common curriculum goals (CCGs);
3. Ensure an appropriate distribution across depth of knowledge categories;
4. Optimize the classification of proficient/not proficient on content strands; and
5. Optimize the measurement variance.

The first three objectives ensure the alignment of the test administered. The fourth objective targets instructionally relevant information from the test. The fifth seeks an optimally reliable score for each student.

In addition, the algorithm must prevent students taking the test on multiple opportunities from receiving the same items.

Objectives 1-3: Alignment

The first three objectives are required to meet the blueprint of the test. These characteristics of a test's correspondence with its blueprint are evaluated by federal peer-review committees. The typical evaluation requires an independent review, commonly drawing on the methods proposed by Webb (1997) and implemented in the Web Alignment Tool (WAT, http://www.wcer.wisc.edu/WAT). The tool measures:

- The alignment of items to the content standards that they are designed to measure (*Categorical Concurrence*) and whether the difficulty of the item stems from that knowledge (*Source of Challenge)*;
- The range of knowledge required by the test and relative emphasis on content within a standard (*Range of Knowledge Correspondence* and *Balance of Representation*); and
- The degree to which the items require knowledge or skills at the complexity level identified in the standards (*Depth of Knowledge Consistency*).

Meeting these criteria, and demonstrating that they have been met in an adaptive test, requires a three step process:

- First, the items in the item pool must each be identified in terms of content and depth of knowledge categories, and these classifications must be shown to be accurate and the source of difficulty of the items.
- Second, the test blueprint must require that the items administered be distributed appropriately across content and depth of knowledge categories; and
- Third, the algorithm must ensure that the items administered match the blueprint.

Objective 4: Instructionally Relevant Feedback

Students and educators devote substantial time and energy into testing. They would like to receive some actionable information back from the test. While psychometric practice has traditionally shied away from providing diagnostic feedback from a summative test, the formative and summative assessment are not necessarily at odds. A criterion-referenced test designed to meet the alignment objectives outlined above can often hold useful information at the content strand level.

Psychometricians note that content-strand level scores from a summative assessment typically entail substantial measurement error. In most cases, the true correlations among strand scores is high enough and the measurement error large enough that cross-strand differences that may exist are obscured by measurement error.

Large measurement error and high correlations among strands combine to create a risk of misinterpretation. The measurement error will lead to scores that look different, even when there is no real difference in proficiency across content strands. The high correlations among content strands ensures that there will be few real differences. Combined, these factors can lead to many "false positives," where educators may believe that students have a particular strength or weakness that is not really supported by the data.

Reports from the tests can guard against these misinterpretations while still providing content-strand level information. Rather than report scores for content strands, we recommend reporting content-strand-level information in three categories:

- Clearly above proficient;
- Too close to proficient to tell whether the examinee has met the mark or not; and
- Clearly below proficient.

Examinees can be categorized as clearly above proficient or below proficient if it is highly improbable that their observed vector of item responses would result if their true scores were exactly at the threshold value. Failure to reach the pre-established confidence level places the examinee in the middle, indeterminate category. Readers of the report can validly contrast performance on "clearly above" strands to "clearly below" strands. Careful wording can lead them to undertake further evaluation and assessment before drawing any conclusions about the content strands in the middle category.

An adaptive algorithm can select items to minimize the probability of an examinee falling into the middle category, thereby maximizing the diagnostic value of the assessment.

Objective 5: Measurement Variance

Minimization of measurement variance is the classic application of an adaptive algorithm. Typical algorithms seek to select the item that maximizes the information about an examinee's score. Using the current estimate of the examinee's score, the information is maximized by the item with the difficulty closest to the examinee's score.

Typically, a graph of the standard error of measurement of adaptive tests is relatively flat across a wide range of ability, while a static test will take on a "U" shape, indicating less information near the ends of the scale.

Scores known more precisely yield fewer classification errors, and provide a more stable basis for growth estimates.

Objective Function

The item selection algorithm is designed to meet these multiple objectives. We begin by defining an objective function. Each time an item is selected, it is selected to maximize the objective function. Depending on the size and constitution of the item pool, it may be desirable to select randomly from among the highest-scoring items, rather than selecting the highest scoring item with certainty.

The objective function is given by

$$f_{ijt} = \frac{1}{\sum_{k=0}^{K+1} w_{kt}} \left(w_{0ijt} u_{ijt} + \sum_{k=1}^{K} w_{kijt} v_{kijt} + w_{(k+1)ijt} d_{ijt} \right) \tag{1}$$

where f_{ijt} is the value of the item *j* for examinee *i* for selection number *t*, and *i={1,2,…,N}, j={1,2,…J}*, and *t={1,2,…,T}*. The content strands are indexed by *k* for *k*={0,1,2…K} content strands, with Strand 0 as the overall score. The value function for the reduction of measurement variance is represented as *u*, and *v* represents the value of the item for the purpose of optimizing strand-level classification.

The variable d_{ijt} indicates the number of minimum requirements that the item contributes to satisfying. In the early iterations, this will be a constant because every item will meet at least requirements on every dimension that requirements exist; later it will be a mix. Inclusion of this term will cause the algorithm to favor items that are required for the blueprint. Once all of the minimums have been met, this term drops out.

The weights, *w*, serve a dual role 1) establishing the relative importance of the two value functions and 2) enforcing the alignment objectives. Below, we will define *u*, *v* and *w* to be greater than or equal to zero.

Alignment Objectives

Alignment objectives seek to administer a test that matches the blueprint. A blueprint can be specified as a collection of *constraint sets*. A constraint set is a set of exhaustive, mutually exclusive classifications of items. For example, if a subject consists of four content strands and each item measures one and only one of the strands, the content strand classifications constitute a constraint set. Each category is associated with a minimum and maximum number of allowable items.

The alignment objectives as specified in the blueprints are enforced through a combination of weights and the term d_{ijk}:

- Setting weights to zero eliminates items that would exceed the maximum allowable number of items with a given characteristic;
- The final term in Equation 1, $w_{(k+1)ijt}d_{ijt}$, causes to the algorithm to favor items needed to meet minimum requirements.

In addition, the weights also serve to establish the relative importance of the objectives.

Begin by defining a set of constraints *c*, which consists of R_c exhaustive, mutually exclusive categories. Each category *r={1,2,…* R_c *}* represents a feature or classification of an item, and the constraint takes the form of a required minimum number of items with the characteristic administered, and a maximum allowable number of items to be administered with the characteristic. Designate the required minimum number of items in category *r* in constraint set *c* as $n_{r(c)}$ and the maximum for the same category as $m_{r(c)}$. Let

$$r_{cj} = \begin{cases} 1\ if\ \ item\ \ j\ has\ characteristic\ r,\ r \in c \\ 0\ otherwise \end{cases}$$

and $r_{cit} = \sum r_{cj}$ represent the number of items with characteristic r_c already administered to examinee *i*, where the summation is over items administered to examinee *i* through time *t*. The weights are then given by

$$w_{ijt} = \begin{cases} 0\ if\ \ r_{cit} = m_{r(c)}\ for\ any\ c \\ \gamma_m\ otherwise \end{cases}$$

where γ_m is an empirically derived constants that captures the relative importance of the objectives for *m={0,1,…K+1}.*

To value the alignment objectives, define $d_{ijt} = 1 + \sum_{c=1}^{C} d_{cijt}$ for a candidate item with characteristic r_c where $d_{cijt} = \begin{cases} 1 if\ r_{cit} < n_{r(c)} \\ 0\ otherwise \end{cases}$. This will favor items that meet minimum requirements that are not yet met. The constant added to the sum ensures that the third term remains substantial unless $w_{ijt} = 0$, which occurs when the item contains a characteristic that has hit the maximum.

Objectives 1-3 require three constraint sets: content strands, benchmarks, and depth of knowledge classifications.

Content Strand Classification Objective

The value of an item on a content strand is defined as the change in the certainty with which a student can be classified as above or below proficient on the content strand. The chi-square statistic yields an estimate of the confidence with which we can say a student is at a score (θ_{ik}) on a content strand rather than right at the cutscore (θ_k^*). The value function, therefore, is expected change in confidence that student is not at the cut-score on the content strand from administering the item.

This value function corresponds to a likelihood ratio test in which $\theta_{ik} = \theta_k^*$ is taken as the null hypothesis. For each item, we calculate the expected change in the likelihood from administering the item, and transform that to a probability value based on the χ^2 distribution with a single degree of freedom. Letting $p(\theta,b) = \frac{1}{1+e^{b-\theta}}$, the probability of a correct response under a Rasch model,

$$v_{kijt} = \chi_{kijt} - \chi^2[1,2\delta_{kit}]$$

where

$$\chi_{kijt} = \chi^2\left[1,2\left((E(L_{kijt}) - E(L^*_{kij}) + \delta_{kit}\right)\right]$$

$$E(L_{kijt}) = p(\theta_{kit},b_j)Log[p(\theta_{kit},b_j)] + \left(1 - p(\theta_{kit},b_j)\right)Log\left(1 - p(\theta_{kit},b_j)\right)$$

$$E(L^*_{kij}) = p(\theta_{kit},b_j)Log[p(\theta^*_k,b_j)] + \left(1 - p(\theta_{kit},b_j)\right)Log\left(1 - p(\theta^*_k,b_j)\right)$$

$$\delta_{itk} = \sum_{j\in k}\left(z_{ij}Log(p(\theta_{kit},b_{ij})) + (1-z_{ij})Log(1-p(\theta_{kit},b_{ij}))\right)$$

$$-\sum_{j\in k}\left(z_{ij}Log(p(\theta^*_k,b_{ij})) + (1-z_{ij})Log(1-p(\theta^*_k,b_{ij}))\right)$$

$$z_{ij} = \begin{cases} 1\,if\ student\ i\,responded\ correctly\ to\ item\ j \\ \quad 0\,otherwise \end{cases}$$

$\chi^2[1,x]$ *is the cumulative* χ^2 *distribution with* 1 *df evaluated at* x.

θ^*_k *is the cutscore on strand* k

θ_{kit} *is student* i*'s estimated score on strand* k *at iteration* t.

Hence, the value is given by the expected change in the probability of the observed response vector having been generated by a value of θ_{kit} rather than θ^*_k. Maximizing this difference in probability minimizes the probability of a student score landing in the indeterminate range in the middle of the scale.

The working value of θ_{ikt} is given in a subsequent section.

Minimum Variance Objective

Minimizing the measurement variance is the typical objective of a computer adaptive test. Minimizing the variance is the same as maximizing the information (the inverse of the variance). For the Rasch model, this is given by

$$I_{it} = \sum_{r=1}^{t} p(\theta_t, b_{ijr})(1 - p(\theta_t, b_{ijr}))$$
$$I_{ijt} = p(\theta_t, b_j)(1 - p(\theta_t, b_j))$$
$$p(\theta, b) = \frac{1}{1 + e^{(b-\theta)}}$$
$$b_{ijr}\ \mathit{is\,the\ difficulty\ of\ the\,item\ taken\ by\,student}\ i$$
$$\mathit{at\,iteration}\ r.$$

For this multiple-objective application, the value of the smaller measurement variance diminishes as the with the measurement variance. When the measurement variance is very large, it is important to reduce it. If the measurement error is very small, increased information may make a trivial difference. Using the information function directly would yield a value measure that was constant across these two different conditions.

Instead, we propose to use $u_{ijt} = \frac{I_{ijt}}{I_{it} + I_{ijt}}$, which bears a monotonic relationship to the information function, but which diminishes as the total information increases (i.e., as the measurement variance decreases). This function will place diminishing value on reducing measurement error as the measurement error diminishes.

Calculating Scores

The algorithm requires scores for the overall test as well as each content strand. The working θ's will be known with little precision, especially for content strands and for the overall score early in the test. Therefore, we will use *expected a posteriori* (EAP) scores as working scores in the calculations. The priors will be different for the overall θ's and the strand scores.

At the conclusion of the test, the final overall score assigned to the student will be the MLE score. On each content strand, each student will receive a classification and associated probability, rather than a score.

The estimates can be obtained by taking Newton steps towards the optimum. Each step in the Newton algorithm is given by

$$\theta_{is} = \theta_{is-1} + \frac{1}{\left(\sum_{j=1}^{J_t} H_{j\theta_{s-1}} + h(\theta_{is-1};\mu,\sigma)\right)}\left(\sum_{j=1}^{J_s} g_{j\theta_{s-1}} + g(\theta_{is-1};\mu,\sigma)\right)$$

where the summation is over the set of items tat examinee *i* has taken through time *t*.

$$g_{j\theta_{s-1}} = D_\theta Log\left(-\, p(\theta_{is-1},b_j)^{z}\ (1-p(\theta_{is-1},b_j))^{(1-z_{ij})}\right) = z_{ij} - p(\theta_{is-1},b_j)$$
$$H_{\theta_{s-1}} = -D_\theta^2 Log\left(p(\theta_{is-1},b_j)^{z_{ij}}(1-p(\theta_{is-1},b_j))^{(1-z_{ij})}\right) = p(\theta_{is-1},b_j)(1-p(\theta_{is-1},b_j))$$

The functions *g(.)* and *h(.)* represent the first and second derivatives of the prior respectively, which we will take as normally distributed. Hence, $g(\theta_{is-1};\mu,\sigma) = \frac{\mu-\theta_{is-1}}{\sigma^2}$ and $h(\theta_{is-1};\mu,\sigma) = \frac{1}{\sigma^2}$.

Overall Score

For the overall score, the prior μ and σ can be taken as the population mean and standard deviation from the prior year, or be based in some way on the student's performance on an earlier test.

Content Strand Scores

For the content strand scores we propose to use $\tilde{\theta}_{ikt} \sim N(a_{0k} + a_{1k}\theta_{it}, \sigma_k^2 + \frac{1}{I_{it}})$, where the coefficients *a* and the standard deviations derive from a linear regression based on a previous year's data and appear to the algorithm as known constants.

It may be possible to achieve a close approximation with only a single step, or it may be necessary to take multiple steps of this algorithm. This can be determined empirically.

Item Bank Management

The Item bank contains all possible items from which operational item pools will be selected from. While this is not part of the adaptive algorithm, a brief summary of the rules for selecting item pools from the bank may be helpful.

These rules include the following:

1. An item exposure parameter, such that the number of times an item has been presented during testing can be monitored and over-exposure can be prevented.
2. Item bank management will ensure that 50% of the items in an operational item pool will not have been exposed on the previous two year's tests
3. When an item is exposed 50,000 times, it is retired from the pool

A Few Comments about the Algorithm

Sets of Items Following Common a Stimulus

Sets of passages pose no particular problem for this algorithm. Each passage with its attendant items are considered as a set, with the value function equal to the sum of the values of each item included. It may happen that passages contain items that push over the maximum.

Reduction to Common Algorithm

This algorithm selects the same items as the traditional minimum variance algorithm if $\gamma_k = 0$ for all $k > 0$.

Content Strand Calculations and Reporting

The value function for the content strand classification is a likelihood ratio test testing the null hypothesis that the examinee is at the cut-score. As such, it is exactly the calculation needed to coordinate with the recommended content-strand reporting approach (clearly above, indeterminate, clearly below).

References

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states. A study of the State Collaborative on Assessment & Student Standards (SCASS) Technical Issues in Large-Scale Assessment (TILSA). Washington, D. C.: Council of Chief State School Officers.