Berkeley Evaluation & Assessment Research Center
Graduate School of Education
University of California, Berkeley
Berkeley, CA 94720-1670
http://bearcenter.berkeley.edu
Technical Report Series
No. 2005-04-02

Constructing Measurement Models for MRCML Estimation: A Primer for Using the BEAR Scoring Engine

Cathleen A. Kennedy

June 22, 2005

Abstract

Representing complex science inquiry tasks for item response modeling presents a number of challenges for the assessment designer. Typically, such tasks provide evidence of multivariate aspects of learning and involve sequential or interdependent responses. The BEAR Scoring Engine was developed to compute proficiency estimates for such tasks. It produces these estimates using a multidimensional, polytomous extension to the Rasch model known as the Multidimensional Random Coefficients Multinomial Rasch Model (MRCMLM). This paper presents background information on how proficiency estimates are computed using this model and then describes how a number of assessment tasks can be modeled and estimated using the scoring engine.

Keywords: assessment, item response theory, IRT, multidimensional, MRCML, Rasch

Table of Contents

Introduction 1

Measurement Models 5
MRCML Model 13

Measurement Model Examples 18

Unidimensional Models 19
1. A Unidimensional Dichotomous model 19
2. Unidimensional Partial Credit model 20
3. Unidimensional Rating Scale model. 22
4. A Simple Item Bundle example. 26
Multidimensional Models 29
5. Between-Item Multidimensional model. 29
6. Within-Item Multidimensional Partial Credit model. 31
7. Between-item multidimensional bundle example: Bundle is multidimensional with each component item mapping to a single dimension. 35
8. Within-item multidimensional bundle: Bundle is multidimensional with individual component items mapping to multiple dimensions. 37

Conclusions 38

Table of Figures

Figure 1. A partial construct map describing how students are expected to perform at three levels of understanding on the construct. 2
Figure 2. Representation of an assessment model involving two constructs demonstrating the assessment goal of measuring both Science and Math knowledge. Both constructs are needed to solve the assessment task. 3
Figure 3. Sample of an assessment task with one stimulus prompting a sequence of interdependent student responses. 4
Figure 4. Item characteristic curve for a dichotomous (2-category) item. 6
Figure 5. Category probability curves and δ_{ij} values for a 3-category polytomous item. 8
Figure 6. δi, $\tau 1$ and $\tau 2$ representations for the polytomous case with 3 categories. 9
Figure 7. Cumulative probability curves and Thurstonian thresholds for a 4-category polytomous item. Dashed line shows probability = .5. 11
Figure 8. Cumulative probability curves and Thurstonian thresholds for a 4-category polytomous item. 12
Figure 9. Representing probability equations with scoring and design matrices. 15
Figure 10. Between-item and within-item multidimensionality. 16
Figure 11. Example of items measuring different constructs. 30
Figure 12. Example of within-item multidimensionality. 32
Figure 13. A between-item multidimensional bundle. 36

Introduction

An assessment is comprised of a series of tasks that are administered to a respondent to elicit evidence about his or her ability, behavior, or attitude. These targeted abilities, behaviors, or attitudes are referred to as constructs; measuring an individual's locations on the constructs is the goal of assessment. A unidimensional construct can be represented as a continuum from having less of the ability, behavior, or attitude to having more of it, and although a particular assessment may target a narrow range on the continuum, the construct itself is theoretically without bounds. Examples of constructs in an educational setting might include "knowledge of force and motion" or "ability to apply inquiry skills when solving a problem." In a healthcare setting one may find constructs such as "physical function" or "cognitive function." And, in a political survey setting "attitude about voting" or "belief that U.S. troops should be withdrawn from Iraq" could be constructs of interest.

A construct map is a representation of a unidimensional construct, including qualitatively different ordered levels for both the items that tap into the constructs, and the respondents who populate the construct. Figure 1 shows the respondent side of a construct map. When we speak of measuring, we mean identifying the location of a particular respondent at some point on the construct continuum. Aligning all items and respondents on the same continuum enables valid and reliable comparisons between respondents at a specific point in time, and within a respondent at different time points.

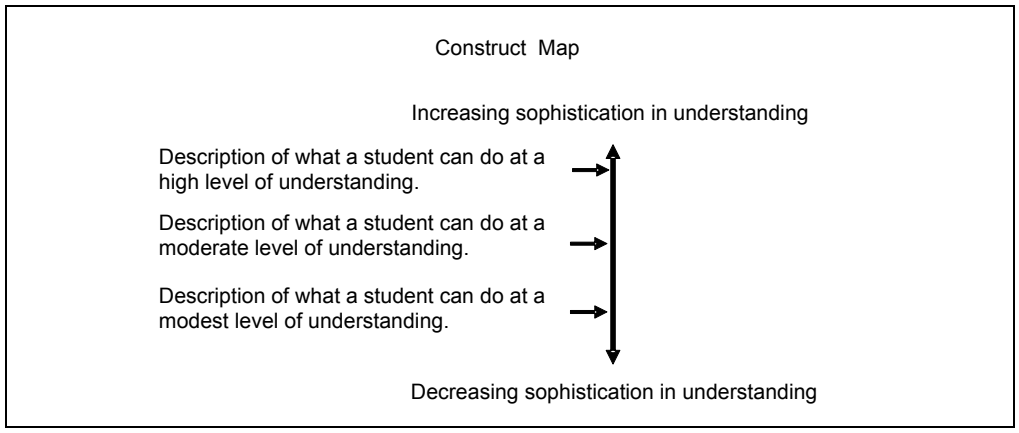


Figure 1. A partial construct map describing how students are expected to perform at three levels of understanding on the construct.

Constructs are considered latent in that they cannot be directly observed; instead, we must draw inferences about a respondent's location on the construct from evidence we gather through observation. Defining one or more relevant construct maps is the first building block for assessment development as advanced in the BEAR (Berkeley Evaluation and Assessment Research center) Assessment System (Kennedy, 2005; Wilson, 2005). In the context of this paper, we refer to the collection of constructs of interest for a particular assessment or for a series of related assessments as the *dimensions* of an assessment model. Figure 2 is a graphical representation of a two-dimensional assessment model. In this example, an assessment designer has theorized that both Science knowledge and Mathematical ability contribute to a person's response to a particular assessment task. In addition, the two constructs are related to one another, as indicated by the curved line between them.

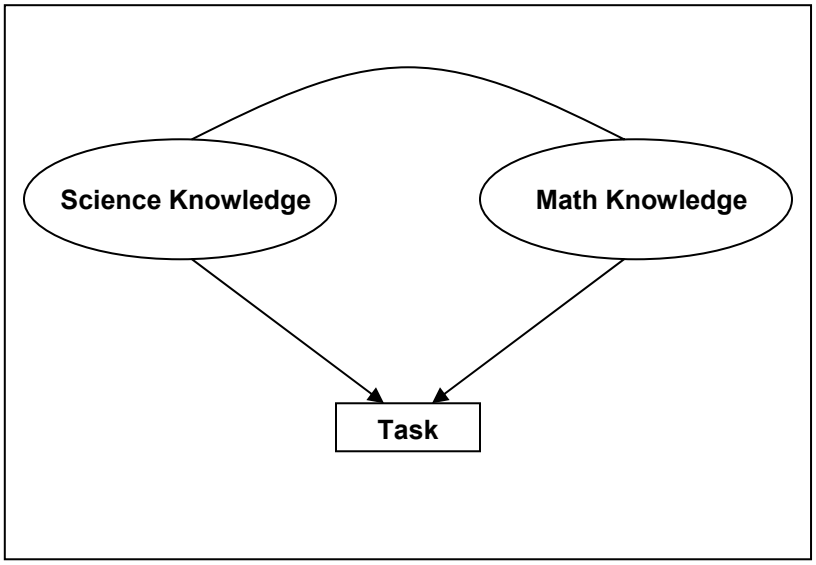


Figure 2. Representation of an assessment model involving two constructs demonstrating the assessment goal of measuring both Science and Math knowledge. Both constructs are needed to solve the assessment task.

An assessment delivery system, whether computerized or manual, is comprised of four interrelated processes, as described in the Four Process Model developed by Almond, Mislevy and Steinberg (2002): 1) Assessment tasks are selected for delivery to the respondent, 2) the tasks are rendered and presented to the respondent and respondent work products are collected, 3) the work products are evaluated and categorized into evidence associated with the targeted constructs, and 4) the evidence is used to draw inferences about the constructs for individual respondents. Within an assessment delivery system, a s*coring engine* is used to implement the fourth process to produce proficiency estimates in accordance with a particular measurement model. The measurement model defines the way evidence is used to draw inferences about respondents' proficiencies in the domains of interest; that is, it connects the evidence to the constructs.

The individual responses from a work product are referred to as item responses (an example is shown in Figure 3) and the evaluations of those responses into

qualitatively different levels (i.e., response categories) are referred to as *scores* in this paper. Thus, one assessment task may include multiple item responses. We note that a single score can be associated with multiple constructs in what we refer to as a *within-item multidimensional model*. Multidimensional models are explained in the next section, and examples of associated measurement models are included in the *Measurement Model Examples* section.

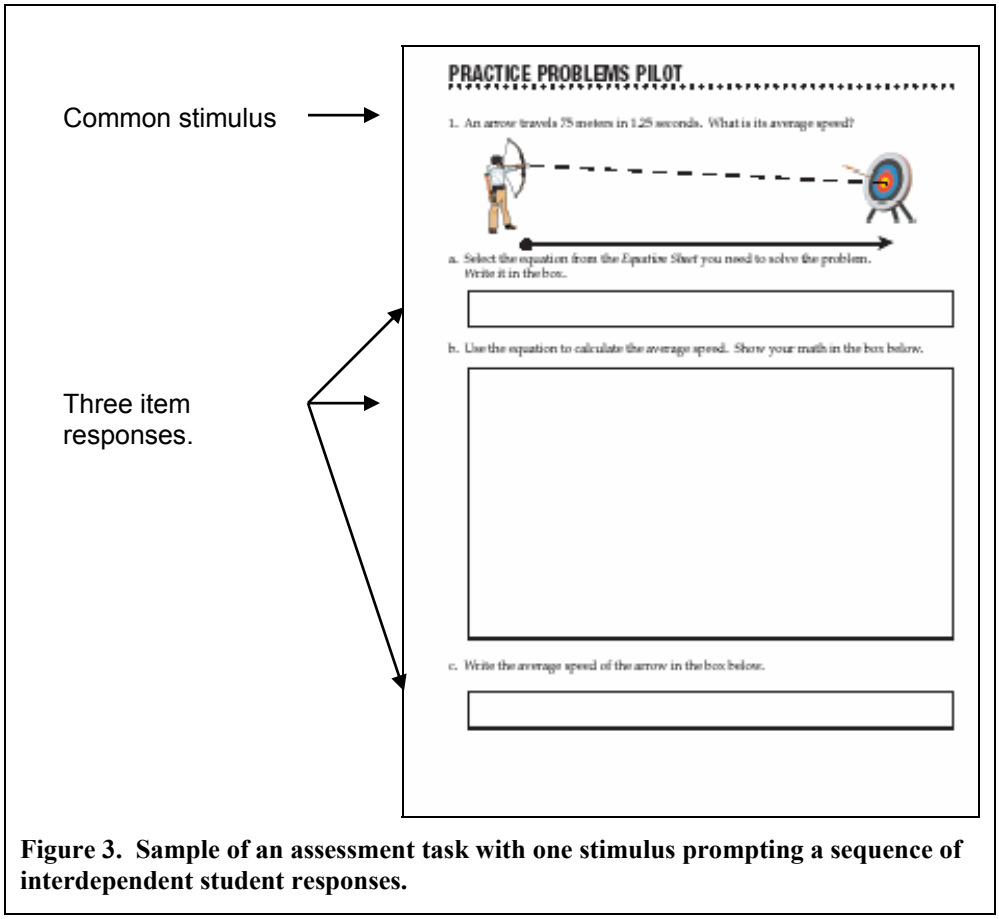


Figure 3. Sample of an assessment task with one stimulus prompting a sequence of interdependent student responses.

In many assessment environments, for example in assessing scientific inquiry abilities, assessment tasks provide evidence of multiple aspects of knowledge and involve sequential and interdependent responses. Designing measurement models that capture the rich information about respondent thinking available in such tasks presents the

assessment developer with a number of challenges. The BEAR Scoring Engine is designed to deal with a number of complexities associated with such tasks. It uses the Multidimensional Random Coefficients Multinomial Logit (MRCML) model (Adams, Wilson & Wang, 1995), which provides a generalized solution for a family of multidimensional, polytomous Rasch-based models. It also accepts a wide array of parameters to define the model, for example, allowing the designer to specify models representing item bundles and/or within-item multidimensionality. Assessment developers specify the model by defining a prior multivariate distribution, scoring and design matrices, and anchored item parameters. These specifications and the response data are sent to the Scoring Engine, which applies the appropriate proficiency algorithm, computes proficiency estimates, and returns updated respondent locations on the constructs to the requesting application. The assessment application determines what to do with the information. For example, reports and charts may be generated, narratives for formative feedback may be provided, or the information may provide input for processing decisions such as which task to deliver next or the selection of an appropriate tutorial segment.

This paper presents background information about how proficiency estimates are computed using the MRCML model and then describes how measurement models might be developed for a number of assessment tasks.

Measurement Models

To compute the probability of achieving a score of 1 rather than 0 on item *i*, given an item difficulty parameter of δ_i, and a specific level on the construct, denoted as θ in the unidimensional case, we use a Rasch formulation (Rasch, 1960) in the form:

$$P(x_i = 1 \mid \theta, \delta_i) = \frac{P(x=1)}{P(x=0) + P(x=1)} = \frac{\exp(\theta - \delta_i)}{1 + \exp(\theta - \delta_i)} \quad (1)$$

For this dichotomous case we have two probability equations:

$$P(x=0) = \frac{1}{1 + \exp(\theta - \delta_1)}; \text{ and}$$

$$P(x=1) = \frac{\exp(\theta - \delta_1)}{1 + \exp(\theta - \delta_1)}.$$

When an item response has only two possible values, correct or incorrect (e.g., in an educational context), the item difficulty is an expression of how much ability a person needs to give a correct answer. By convention, we describe the item difficulty as the proficiency level where the respondent is equally likely to get a correct or incorrect response (that is, both probabilities are .5). In Figure 4, for example, the item difficulty is -0.18; this is the point at which the probability curves for a correct and an incorrect response intersect.

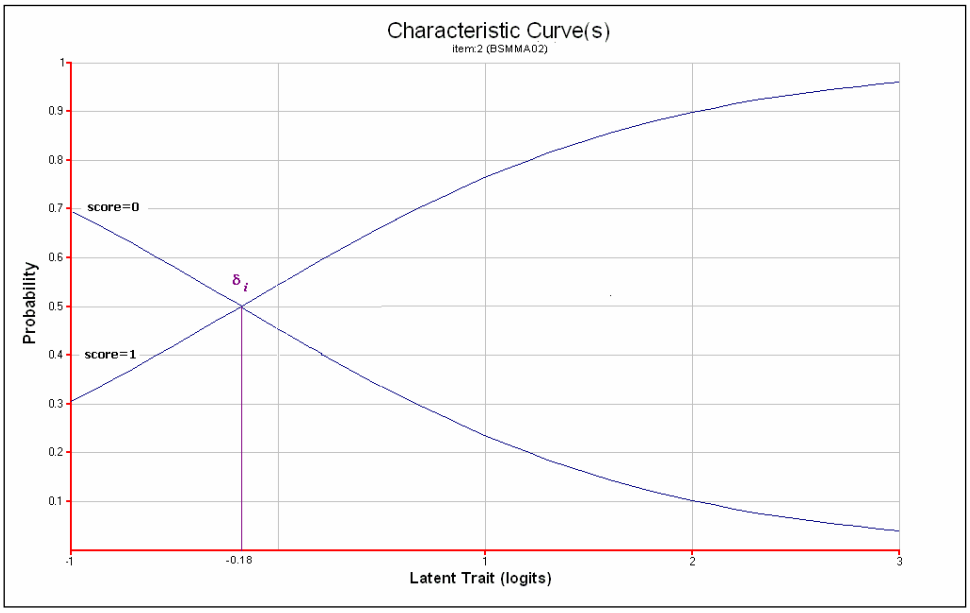


Figure 4. Item characteristic curve for a dichotomous (2-category) item.

For the polytomous case (Partial Credit Model; Masters, 1981), the following equation shows the probability that a person with a proficiency of θ will respond in category *c* rather than in any other category on item *i*, given item difficulty parameters $\xi_i=(\delta_{i1}, \delta_{i2}, \ldots \delta_{im})$.

$$P(x_i = c \mid \theta, \xi_i) = \frac{\exp \sum_{j=0}^{c} (\theta - \delta_{ij})}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} (\theta - \delta_{ij})}, \tag{2}$$

where *m* is the number of steps (number of categories-1) for the item.

For example, for a 3-category item (with 2 steps),

$$P(x_i = 0) = \frac{1}{\sum_{k=0}^{2} \exp \sum_{j=0}^{k} (\theta - \delta_{ij})} = \frac{1}{1 + \exp(\theta - \delta_{i1}) + \exp(2\theta - (\delta_{i1} + \delta_{i2}))};$$

$$P(x_i = 1) = \frac{\exp \sum_{j=0}^{1} (\theta - \delta_{ij})}{\sum_{k=0}^{2} \exp \sum_{j=0}^{k} (\theta - \delta_{ij})} = \frac{\exp(\theta - \delta_{i1})}{1 + \exp(\theta - \delta_{i1}) + \exp(2\theta - (\delta_{i1} + \delta_{i2}))}; \text{and}$$

$$P(x_i = 2) = \frac{\exp \sum_{j=0}^{2} (\theta - \delta_{ij})}{\sum_{k=0}^{2} \exp \sum_{j=0}^{k} (\theta - \delta_{ij})} = \frac{\exp(\theta - \delta_{i1} + \theta - \delta_{i2})}{1 + \exp(\theta - \delta_{i1}) + \exp(2\theta - (\delta_{i1} + \delta_{i2}))} = \frac{\exp(2\theta - (\delta_{i1} + \delta_{i2}))}{1 + \exp(\theta - \delta_{i1}) + \exp(2\theta - (\delta_{i1} + \delta_{i2}))}.$$

Note the conventions $\exp(0) \equiv 1$ and $\sum_{j=0}^{0} (\theta - \delta_{ij}) \equiv 0$; and that

$\sum_{k=0}^{m} \exp \sum_{j=0}^{k} (\theta - \delta_{ij})$ is the sum of the numerators for all categories.

When items have more than two possible outcomes, we need more information than the item difficulty. We need to know how much more of the construct (knowledge, behavior, or attitude) is needed to achieve each possible score on the item. The partial credit case is an extension of the dichotomous case; moving from one category to another implies a dichotomous choice between two levels. For example, consider an item with three categories, scored 0, 1, or 2. The difficulty for step 1, denoted as δ_1, is located at the

location where, if one is considering just categories 0 and 1, one is equally likely of getting the item partially correct (where x=1) or incorrect (where x=0). Note in Figure 5 that this is where the curve for getting a score of 0 intersects with the curve for getting a score of 1. Subsequent steps in difficulty are interpreted in much the same way. The second step difficulty, δ_2, is the proficiency required to have equal probabilities of getting a score of 2 or a score of 1 on the item.

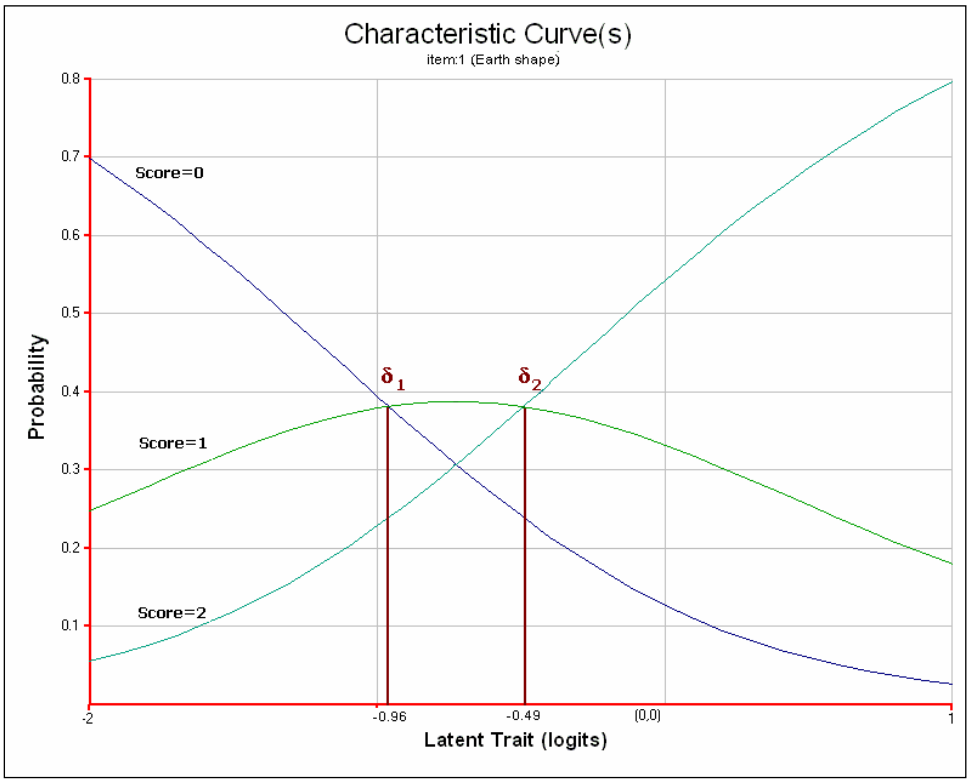


Figure 5. Category probability curves and δ_{ij} values for a 3-category polytomous item.

When using the partial credit model we generally parameterize the difficulty of achieving a score of *j* on item *i* and represent it with δ_{ij}. That is, δ_{ij} is the proficiency level required to expect an equal chance of responding in category *j* or in category *j*-1 on item *i*. Alternatively, we might think of the average of the δ_{ij}'s as an overall item difficulty, and the step difficulties as each step's deviance from the average. In looking at item difficulties in this way we are saying that each δ_{ij} can be formulated as $\delta_i + \tau_{ij}$, where τ_{ij} is the deviance from the average item difficulty for item *i* at step *j*. Note that in this case the last tau parameter is equal to the negative sum of the others so that the sum of all the tau parameters equals zero, $\tau_{im} = -\sum_{k=1}^{m-1} \tau_{ik}$. A graphical representation of this formulation for an item with two steps (and therefore three categories) is shown in Figure 6.

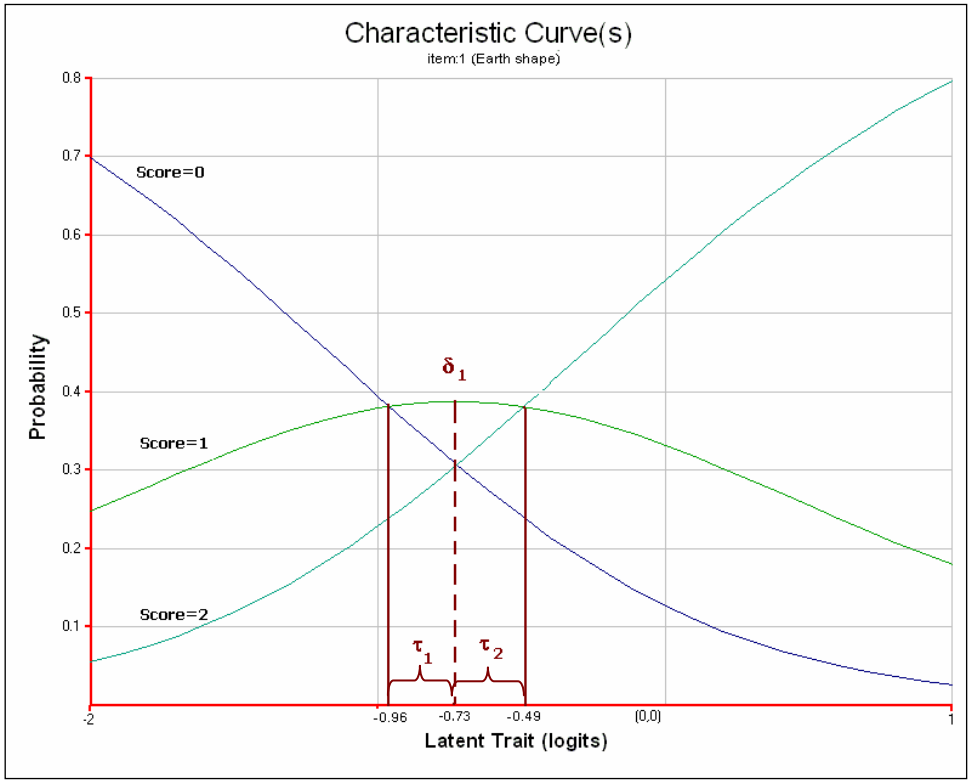


Figure 6. δi, τ1 and τ2 representations for the polytomous case with 3 categories.

The rating scale model is a special case of the partial credit model in which the tau parameters for step *j* are the same for every item. That is, $\tau_{11}=\tau_{21}=\tau_{31}...$, $\tau_{12}=\tau_{22}=\tau_{32}...$, etc. In this formulation, our measurement model becomes

$$P(x_i = c \mid \xi_i, \theta) = \frac{\exp \sum_{j=0}^{c} [\theta - (\delta_i + \tau_j)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\theta - (\delta_i + \tau_j)]}, \quad (3)$$

where $\xi_i=(\delta_i, \tau_1, \tau_2,...\tau_{m-1})$. Again, the final tau parameter, τ_m, is not estimated because it is constrained to make the sum of all the tau parameters equal to zero.

The different parameterization techniques of the step difficulties for partial credit models and the item difficulties and tau parameters for rating scale models is an important distinction in representing the probability equations for Scoring Engine measurement models. If a rating scale model is to be used, then all items that map to the same dimension must use the same tau parameters. These parameterization options are discussed in more detail in the *Measurement Model Examples* section.

Although these step locations are fundamental to defining the models, we have found that their interpretation can lead to misunderstandings among novices. Hence, we have developed an alternative way to display the model. The location at which a person has a 50% probability of achieving a score in that category *or higher* is referred to as the Thurstonian threshold (Wilson, 2005). These locations can be identified on cumulative probability plots at the points where the curves intersect with the probability = .5 line, as shown in Figure 7. These values tend to be more interpretable than δ_j values because they identify levels where individuals are most likely to achieve specific scores. Using the

values displayed in Figure 7, the Thurstonian threshold at step 1 is -0.68, at step 2 is -0.23, and at step 3 is 0.41.

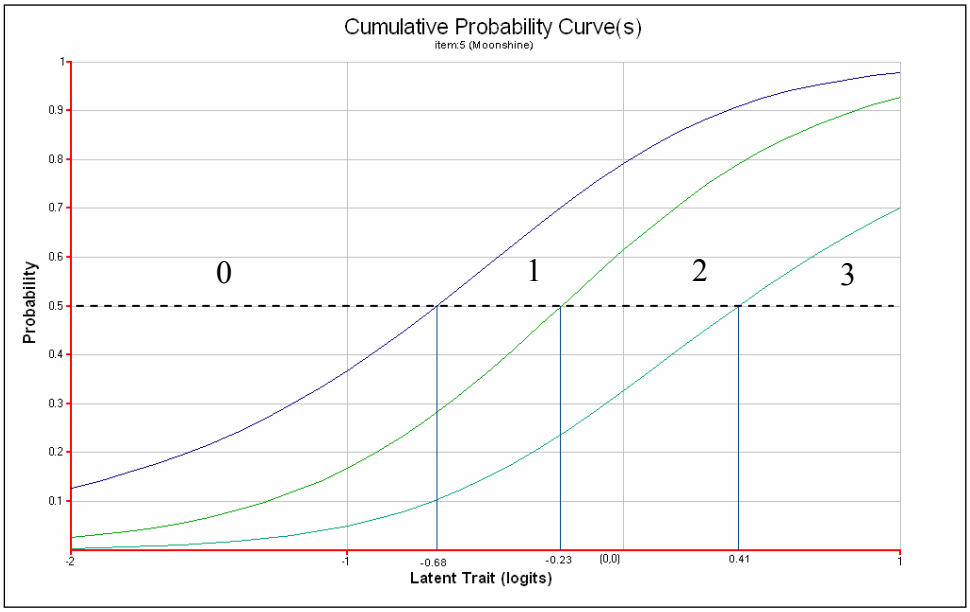


Figure 7. Cumulative probability curves and Thurstonian thresholds for a 4-category polytomous item. Dashed line shows probability = .5.

Figure 8 shows the same item, illustrating that a person with a proficiency located at 0.60 is more likely to achieve a score of 3 than a lower score, while a person with a proficiency located at 0.00 is more likely to achieve a score of 2 or 3 than a score of 1 or 0. This can be determined by examining the vertical lines at logit values of 0.00 and 0.60. For example, at a logit value of 0.60, the vertical line intersects the probability = .5 line in the area where the most probable score is 3. At a logit value of 0.00, the vertical line intersects the probability = .5 line in the area where the most probable score is 2 or higher.

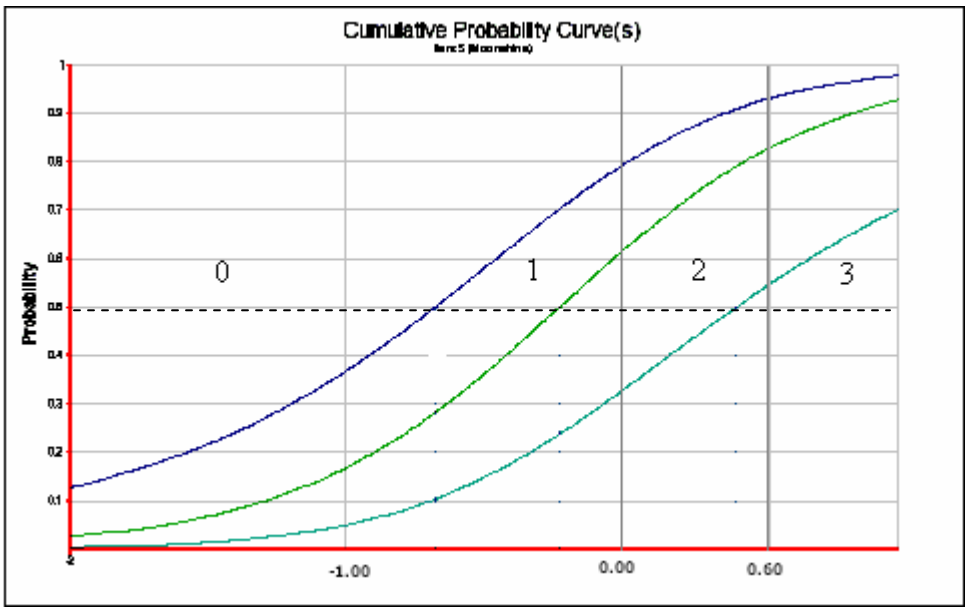


Figure 8. Cumulative probability curves and Thurstonian thresholds for a 4-category polytomous item.

The Random Coefficients Multinomial Logit (RCML) model (Adams & Wilson, 1996) formulates the conditional probability of a response pattern, **x**, as

$$P(\mathbf{X}=\mathbf{x}\mid\theta)=\frac{\exp(\mathbf{x}'(\mathbf{b}\theta+\mathbf{A}\xi)}{\sum_{z\in\Omega}\exp(\mathbf{z}'(\mathbf{b}\theta+\mathbf{A}\xi))}, \qquad (4)$$

where θ is person ability, **b** is the vector of response scores, **A** is the design matrix, ξ is the vector of item parameters, and Ψ is the set of all possible response vectors.

The probability of a particular response pattern occurring is the continued product of the probabilities of the individual responses on an instrument when the items are conditionally independent. When the items are not conditionally independent, item bundles can be constructed to comply with the assumption of item independence in Rasch models (Hoskens & De Boeck, 1997; Wang, Wilson & Cheng, 2000; Wilson & Adams, 1995). Item bundles are described in more detail in Examples 4, 7, and 8 in the *Measurement Models Examples* section.

The measurement models described thus far are all univariate; they measure one construct, or dimension. The types of assessment models one wishes to measure are frequently more complex than this because they involve several aspects of knowledge or proficiency that work together to influence the responses and behaviors observed in respondents. The multidimensional RCML model (MRCML; Adams, Wilson & Wang, 1997) allows us to construct response probabilities and proficiency estimates across multiple dimensions of knowledge, behavior, or attitude.

MRCML Model

The MRCML family of measurement models uses two matrices to define the many different measurement models in the family: A scoring matrix, **B**, that represents the relationships between items (the rows) and dimensions (the columns); and a design matrix, **A**, that represents the relationships between items (the rows) and the model parameters (the columns), such as item difficulties, step difficulties, etc.

The general MRCML formulation for the probability of a response pattern, **x**, to an item is

$$P(\mathbf{x};\boldsymbol{\xi} \mid \boldsymbol{\theta}) = \frac{\exp\left[\mathbf{x}'(\mathbf{B}\boldsymbol{\theta} - \mathbf{A}\boldsymbol{\xi})\right]}{\sum_{\mathbf{z}\in\Omega}\exp\left[\mathbf{z}'(\mathbf{B}\boldsymbol{\theta} - \mathbf{A}\boldsymbol{\xi})\right]} \tag{5}$$

where $\boldsymbol{\theta}$ is the vector of specified proficiency levels on each dimension, $\boldsymbol{\xi}$ is the vector of item parameters, and Ω is the set of all possible response patterns. We use **z** to denote a pattern coming from the full set of response patterns while **x** denotes the response pattern of interest. The response pattern, **x**, is comprised of vectors for each item with one element in the vector for each item category, $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_I\} = \{x_{11}, x_{12}, ..., x_{1m1}, x_{21}, x_{22}, ..., x_{2m2}, ..., x_{Im_I}\}$ for m_i = number of categories for item i, and I = number of items.

Note that in this formulation the item parameters, ξ, are considered known and conditioned on θ.

Figure 9 shows an example for how **A** and **B** might be constructed for a single 3-category item that is an indicator of one dimension. Note that the coefficients for θ are collected in the scoring matrix and the coefficients for the δs are collected in the design matrix. There are two columns in the design matrix because there are two parameters for the item, the two steps δ_{i1} and δ_{i2}; the first column is associated with δ_{i1} and the second with δ_{i2}. There is only one column in the scoring matrix because this is a unidimensional example. In a multidimensional example there would be a column for each dimension. The **A** and **B** matrices will always have an equal number of rows, with one row for each response category in each item. Note that there will always be more than one item in a given assessment context, so both **A** and **B** will usually be larger than shown in this small example.

Note the patterns in the numerators for the series of equations for the response categories of a single item. For example, in the partial credit case with a 3-category item:

P(x=0): $\exp(0\theta - 0) = \exp(0) = 1$

P(x=1): $\exp(1\theta - 1\delta_{i1}) = \exp(\theta - \delta_{i1})$

P(x=2): $\exp(2\theta - (1\delta_{i1} + 1\delta_{i2})) = \exp(2\theta - (\delta_{i1} + \delta_{i2}))$

The θ coefficients comprise the entries in a Scoring Matrix, **B**:

$$\mathbf{B} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

The δ coefficients comprise the entries in a Design Matrix, **A**:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix},$$

where column 1 represents the δ_{i1} coefficients and column 2 represents the δ_{i2} coefficients.

Figure 9. Representing probability equations with scoring and design matrices.

When an assessment is intended to measure multiple dimensions, individual items may measure a single dimension or multiple dimensions. As shown in Figure 10, we refer to the case in which each response category for an item provides evidence about a single dimension as *between-item* multidimensionality and the case in which a single response category provides evidence about multiple dimensions as *within-item* multidimensionality.

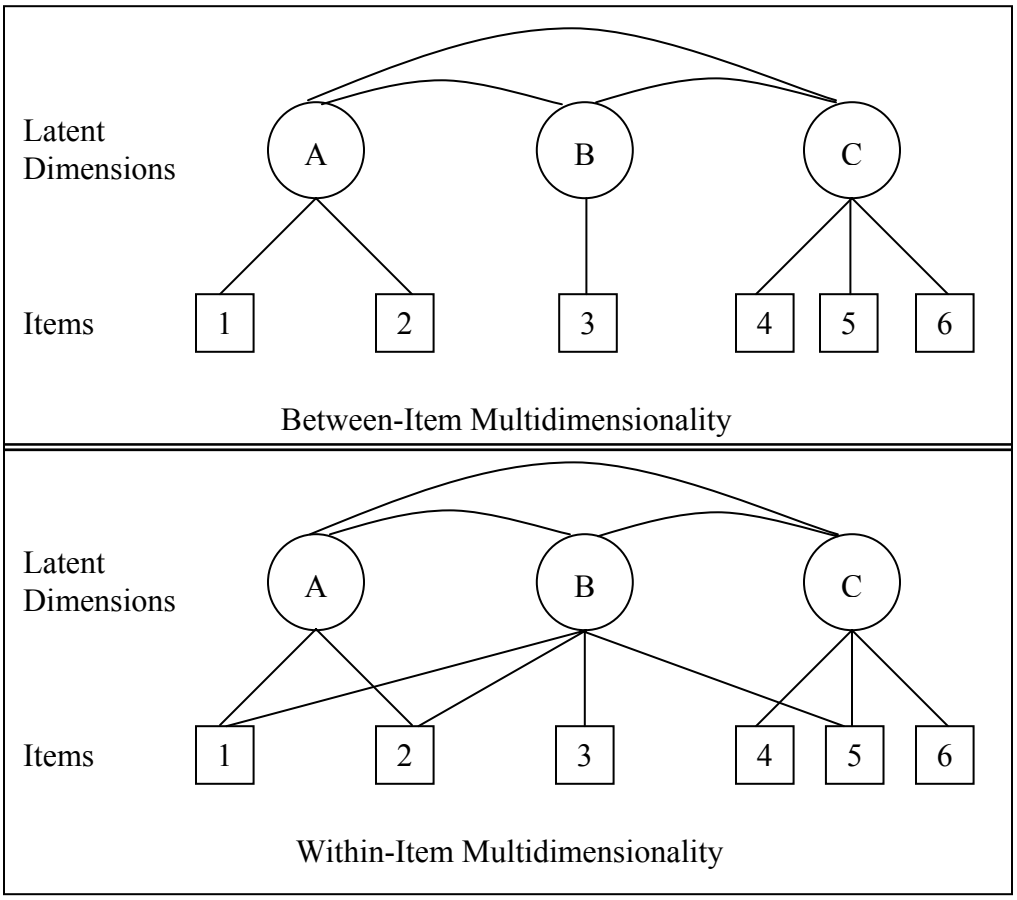


Figure 10. Between-item and within-item multidimensionality.

Take, for example, the case of an assessment comprised of three dichotomous items in which the first two items are indicators of the first dimension and the third item is an indicator of the second dimension (i.e., between-item multidimensionality). The probability of a response pattern of 1, 0, and 1 on the items (i.e., correct responses on items 1 and 3, and an incorrect response on item 2) is computed from the following:

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \ \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \text{ and } \boldsymbol{\xi} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}.$$

Then, $\mathbf{B\theta} = \begin{bmatrix} 0 \\ \theta_1 \\ 0 \\ \theta_1 \\ 0 \\ \theta_3 \end{bmatrix}$, $\mathbf{A\xi} = \begin{bmatrix} 0 \\ \delta_1 \\ 0 \\ \delta_2 \\ 0 \\ \delta_3 \end{bmatrix}$, $\mathbf{B\theta} - \mathbf{A\xi} = \begin{bmatrix} 0 \\ \theta_1 - \delta_1 \\ 0 \\ \theta_1 - \delta_2 \\ 0 \\ \theta_2 - \delta_3 \end{bmatrix}$, and

$$\mathbf{x}'(\mathbf{B\theta} - \mathbf{A}\boldsymbol{\xi}\boldsymbol{\xi} = (\theta_1 - \delta_1) + (\theta_2 - \delta_3).$$

Thus, $P(\mathbf{x};\boldsymbol{\xi} \mid \boldsymbol{\theta}) = \dfrac{\exp\left[(\theta_1 - \delta_1) + (\theta_2 - \delta_3)\right]}{\sum_{\mathbf{z}\in\Omega} \exp\left[\mathbf{z}'(\mathbf{B\theta} - \mathbf{A\xi})\right]}$. Note that **x** is a vector with one element per item-category combination, **B** is the scoring matrix with one column per dimension and two rows per item, **A** is the design matrix with one column per item (no step parameters for dichotomous items), $\boldsymbol{\theta}$ is a vector with one element per dimension (i.e., the same number of elements as columns in **B**), and $\boldsymbol{\xi}$ is a vector with one element per item parameter (i.e., the same number of elements as columns in **A**).

Through MRCML modeling and a flexible model specification structure, the Scoring Engine accommodates assessments that measure multiple aspects of proficiency and which may have item dependencies. The following *Measurement Model Examples* section elaborates on a range of models, from simple to complex, that can be specified for the Scoring Engine.

Measurement Model Examples

The MRCML literature generally considers constructing scoring and design matrices to represent an entire assessment. The Scoring Engine, on the other hand, expects measurement models to be constructed at the item level. This approach encourages reuse of components with similar measurement features. The Scoring Engine constructs a complete assessment measurement model from these individual item-level models. A series of examples demonstrating how one might construct the associated scoring and design matrices follows. Note that examples 1, 2, 3, 5 and 6 assume that each response is independent of any other responses on the assessment, while examples 4, 7, and 8 address issues of conditional dependencies and item bundling. For some examples, both item-oriented matrices and assessment-oriented matrices are shown to assist the reader in differentiating the approach used by the Scoring Engine from that used by assessment-oriented MRCML programs such as ConQuest (Adams, Wu & Wilson, 2005) and GradeMap (Kennedy, Wilson & Draney, 2005).

The Scoring Engine computes student proficiencies using two methods: expected a posteriori (EAP) and maximum likelihood estimation (MLE). The EAP is a Bayesian estimation procedure using both the respondents' scores and the distribution of the respondents, while the MLE approach uses only the respondents' scores. To estimate a respondent's proficiency from his or her responses on an assessment, we use a formulation that evaluates the probability of a person with a known ability responding in each category to an item with known difficulty parameters. The basis of these computations is the estimated item response model for the items.

Unidimensional Models

In addition to the scoring and design matrices described above, the Scoring Engine requires calibrated item parameters to compute response probabilities. These are provided as vectors in which the number of elements is equal to the number of columns in the design matrix. The population parameters must also be defined. A mean vector and a covariance matrix are used to represent these values. For a unidimensional model, the mean vector contains a single value and the variance-covariance matrix contains only the variance.

Unidimensional Means Vector:

$$\begin{matrix} D_1 \\ [0.652] \end{matrix}$$

Unidimensional Covariance Matrix:

$$\begin{matrix} & D_1 \\ D_1 & [0.954] \end{matrix}$$

1. A Unidimensional Dichotomous model.

This model is useful for representing responses that are either correct or incorrect and that measure only one dimension. Examples include making a selection from a list, responding to a true/false or multiple choice question, and fill-in-the-blank items.

Item Scoring Matrix (one dimension, so one column):

$$\begin{matrix} & D_1 \\ \text{Category 1} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \text{Category 2} & \end{matrix}$$

Item Design Matrix (one item difficulty needed, so one column):

$$\begin{matrix} & \delta_1 \\ \text{Category 1} & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ \text{Category 2} & \end{matrix}$$

Item Calibrated Parameters Vector:

$$\begin{array}{ll} & \delta_1 \\ \text{Item difficulty} & \begin{bmatrix} 1.05 \end{bmatrix} \end{array}$$

In the case of an **assessment** with 10 dichotomous items, the associated *assessment* matrices would have the form[1]:

Assessment Scoring Matrix:

$$\begin{array}{ll} & D_1 \\ \text{item 1, category 1} & \begin{bmatrix} 0 \\ \end{bmatrix} \\ \text{item 1, category 2} & 1 \\ \text{item 2, category 1} & 0 \\ \text{item 2, category 2} & 1 \\ \text{item 3, category 1} & 0 \\ \text{item 3, category 2} & 1 \\ \text{item 4, category 1} & 0 \\ \text{item 4, category 2} & 1 \\ \text{item 5, category 1} & 0 \\ \text{item 5, category 2} & 1 \\ \text{item 6, category 1} & 0 \\ \text{item 6, category 2} & 1 \\ \text{item 7, category 1} & 0 \\ \text{item 7, category 2} & 1 \\ \text{item 8, category 1} & 0 \\ \text{item 8, category 2} & 1 \\ \text{item 9, category 1} & 0 \\ \text{item 9, category 2} & 1 \\ \text{item 10, category 1} & 0 \\ \text{item 10, category 2} & \begin{bmatrix} 1 \end{bmatrix} \end{array}$$

Assessment Design Matrix:

$$\begin{array}{ll} & \begin{matrix} \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 & \delta_7 & \delta_8 & \delta_9 & \delta_{10} \end{matrix} \\ \text{item 1, category 1} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\ \text{item 1, category 2} & 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 2, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 2, category 2} & 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 3, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 3, category 2} & 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 4, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 4, category 2} & 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 5, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 5, category 2} & 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0 \\ \text{item 6, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 6, category 2} & 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0 \\ \text{item 7, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 7, category 2} & 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0 \\ \text{item 8, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 8, category 2} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0 \\ \text{item 9, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 9, category 2} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0 \\ \text{item 10, category 1} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ \text{item 10, category 2} & \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

2. Unidimensional Partial Credit model.

This model is used to represent responses that can be scored at more than two levels. A scoring rubric is usually required to describe what a score at each level means relative to the construct being measured. Essay questions are typically scored using this approach, with scores ranging from 0 to 10, for example.

The scoring and design matrices below represent an item with four categories. In this scoring matrix, a response in the third category is represented by a score of 2. Note

[1] Note that this design matrix does not express the constraint which may be needed to identify the model.

that the response data sent to the Scoring Engine indicates which *category* the response is in, using integral values beginning at 0. Thus, a response in the second category is sent to the Scoring Engine as the value 1. The response categories are always positive integers. For simple models, such as that shown below, it is quite common for the response category value to be the same as the score value. However, it is permissible for the scoring matrix to include negative or fractional, or real, values.

Item Scoring Matrix (one dimension, so one column):

$$\begin{array}{l} \\ \text{Category 1} \\ \text{Category 2} \\ \text{Category 3} \\ \text{Category 4} \end{array} \begin{array}{c} D_1 \\ \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} \end{array}$$

Partial Credit Item Design Matrix:

(four categories means three steps, so 3 columns)

$$\begin{array}{l} \\ \text{Category 1} \\ \text{Category 2} \\ \text{Category 3} \\ \text{Category 4} \end{array} \begin{array}{c} \delta_1 \ \ \delta_2 \ \ \delta_3 \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

In this design matrix, the difficulty of achieving a response in the third category is dependent upon the difficulty of advancing from the first category to the second category (the first column), and the difficulty in advancing from the second category to the third category (the second column). That is, the difficulty of achieving a response in the third category is conditioned on being able to earn the lower scores also. This interpretation requires scores to be hierarchically ordered such that each score represents a higher level

of proficiency than the score before. Just as for scoring matrices, entries in the design matrix may also be negative and may be fractional, or real, values.

Item Calibrated Parameters Vector:

$$\begin{matrix} \delta_1 & \delta_2 & \delta_3 \end{matrix}$$
$$\begin{bmatrix} 1.35 & .25 & .86 \end{bmatrix}$$

Note that the number of elements in the calibrated parameters vector is equal to the number of columns in the design matrix.

For an **assessment** with five items in which items 1 through 3 have five categories and items 4 and 5 have three categories, the assessment matrices would take the form[2]:

Assess. Scoring Matrix:

	D_1
item 1, category 1	0
item 1, category 2	1
item 1, category 3	2
item 1, category 4	3
item 1, category 5	4
item 2, category 1	0
item 2, category 2	1
item 2, category 3	2
item 2, category 4	3
item 2, category 5	4
item 3, category 1	0
item 3, category 2	1
item 3, category 3	2
item 3, category 4	3
item 3, category 5	4
item 4, category 1	0
item 4, category 2	1
item 4, category 3	2
item 5, category 1	0
item 5, category 2	1
item 5, category 3	2

Assessment Design Matrix:

	δ_{11}	δ_{12}	δ_{13}	δ_{14}	δ_{21}	δ_{22}	δ_{23}	δ_{24}	δ_{31}	δ_{32}	δ_{33}	δ_{34}	δ_{41}	δ_{42}	δ_{51}	δ_{52}
item 1, category 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 1, category 2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 1, category 3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 1, category 4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
item 1, category 5	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
item 2, category 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 2, category 2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
item 2, category 3	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
item 2, category 4	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
item 2, category 5	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
item 3, category 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 3, category 2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
item 3, category 3	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
item 3, category 4	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
item 3, category 5	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
item 4, category 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 4, category 2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
item 4, category 3	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
item 5, category 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
item 5, category 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
item 5, category 3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

3. Unidimensional Rating Scale model.

This is similar to the unidimensional partial credit model except that the scoring rubric must be consistent with having the same tau parameters for all items on the

[2] Note that, as above, no constraints are built into this example.

assessment; we assume that the scoring rubric performs in the same way for all items. Rating scale models are often used for questionnaires and surveys where a Likert-type set of options has been used (e.g., strongly agree, agree, disagree, strongly disagree). The following scoring matrix could be used for a rating scale item with five categories. In the example below, a response in the second category is represented by a score of 1.

Item Scoring Matrix (one dimension, so one column):

$$\begin{array}{l} \\ \text{Category 1} \\ \text{Category 2} \\ \text{Category 3} \\ \text{Category 4} \\ \text{Category 5} \end{array} \begin{array}{c} D_1 \\ \left[\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\right] \end{array}$$

For a single item, the measurement model could be constructed in the same manner as for the partial credit model. By convention, however, we parameterize the item difficulties differently in the rating scale model (as $(\delta + \tau)$ values), so we construct the design matrix differently also (the need for this is more apparent for an entire assessment than for a single item).

Preliminary Item Design Matrix (average difficulty, δ_i, and four step difficulties, τ_1, τ_2, τ_3, and τ_4, so 5 columns):

$$\begin{array}{l} \\ \text{Category 1} \\ \text{Category 2} \\ \text{Category 3} \\ \text{Category 4} \\ \text{Category 5} \end{array} \begin{array}{c} \delta \quad \tau_1 \quad \tau_2 \quad \tau_3 \quad \tau_4 \\ \left[\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 1 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 \\ 4 & 1 & 1 & 1 & 1 \end{array}\right] \end{array}$$

With this design matrix, the difficulty of achieving a response in the third category is computed from the average difficulty of the item (the first column), the

deviance from the average difficulty to get a response in the second category rather than the first (the second column), and the deviance from the average difficulty to get a response in the third category rather than the second (the third column). These tau parameters have a different interpretation, and are calibrated differently, from the step parameters in the partial credit model, so the formulation of the design matrix looks different from that for the partial credit model.

Note that the total difficulty of getting a response in the third category is:

	average item difficulty + difficulty in going from a category 1 response to a category 2 response
+	average item difficulty + difficulty in going from a category 2 response to a category 3 response
=	2*(average item difficulty) + difficulty in going from a category 1 response to a category 2 response + difficulty in going from a category 2 response to a category 3 response

In MRCML terms, the formulation is denoted as $2\delta_i+\tau_{i1}+\tau_{i2}$. The coefficients 2, 1, and 1 are captured in the design matrix row denoted as "Category 3."

Since the sum of all the tau parameters is 0, the total difficulty of getting a response in the fifth category is $4\delta_i + \Sigma\tau = 4\delta_i$; we simplify the design matrix by setting the tau parameters in the last row to 0. Thus, we do not have to estimate τ_4, and this means that we need only the first four columns (i.e., the average item difficulty, δi, and three step difficulties, τ_1, τ_2, and τ_3).

Final Rating Scale Item Design Matrix

	δ	τ_1	τ_2	τ_3
Category 1	0	0	0	0
Category 2	1	1	0	0
Category 3	2	1	1	0
Category 4	3	1	1	1
Category 5	4	0	0	0

Item Calibrated Parameters Vector:

$$\begin{matrix}\delta & \tau_1 & \tau_2 & \tau_3 \\ \begin{bmatrix}-1.35 & .26 & 1.03 & .64\end{bmatrix}\end{matrix}$$

The rating scale model is considered a special case of the partial credit model. All items noted as Rating Scale for the Scoring Engine use the same parameter estimates for the tau parameters. For an **assessment** comprised of five rating scale items with three categories each, we would need the following scoring and design matrices.

Assessment Scoring Matrix:

	D_1
item 1, category 1	0
item 1, category 2	1
item 1, category 3	2
item 2, category 1	0
item 2, category 2	1
item 2, category 3	2
item 3, category 1	0
item 3, category 2	1
item 3, category 3	2
item 4, category 1	0
item 4, category 2	1
item 4, category 3	2
item 5, category 1	0
item 5, category 2	1
item 5, category 3	2

Assessment Design Matrix:

	δ_1	δ_2	δ_3	δ_4	δ_5	τ_1	τ_2
item 1, category 1	0	0	0	0	0	0	0
item 1, category 2	1	0	0	0	0	1	0
item 1, category 3	2	0	0	0	0	1	1
item 2, category 1	0	0	0	0	0	0	0
item 2, category 2	0	1	0	0	0	1	0
item 2, category 3	0	2	0	0	0	1	1
item 3, category 1	0	0	0	0	0	0	0
item 3, category 2	0	0	1	0	0	1	0
item 3, category 3	0	0	2	0	0	1	1
item 4, category 1	0	0	0	0	0	0	0
item 4, category 2	0	0	0	1	0	1	0
item 4, category 3	0	0	0	2	0	1	1
item 5, category 1	0	0	0	0	0	0	0
item 5, category 2	0	0	0	0	1	1	0
item 5, category 3	0	0	0	0	2	1	1

Notice how the tau parameter columns have entries across different items. This is because these parameters are defined across different items. In this example, we have placed the individual average item difficulties in the design matrix first (i.e., in the leftmost columns), and then the shared tau parameters on the right. This is a convention, but is not required as long as the calibrated item parameters are in the same order as the columns of the associated design matrix. The reconfiguration of individual item matrices into assessment-level matrices is managed by the Scoring Engine.

4. A Simple Item Bundle example.

When a single prompt leads to multiple responses from students it is likely that the responses have some conditional dependencies. For example, in the problem shown in Figure 3 the prompt asks students to compute the average speed, with intermediate responses specifying the equation and the numeric values from the prompt that are to be placed into the equation. If we only use the final response, conditional dependence is not an issue; however, if we wish to capture more of the information available about student thinking, we will want to retain the information from the intermediate responses, and the conditional dependencies must be modeled in some way.

An *item bundle* (Rosenbaum, 1988) can be used to model dependencies between items. The bundling is implemented prior to sending the data to the Scoring Engine. First, individual item responses are evaluated, and then a procedure for combining the intermediate item responses into a new, aggregated (bundled) response is implemented. Only the final bundled response is transmitted to the Scoring Engine and used in estimating proficiencies.

For a simple case, consider 3 dichotomous items in the bundle. One can use a complete model with all possible response combinations, with each mapping to a unique final response, or a reduced model if some of the possible response categories are not needed or if it makes sense to collapse some categories.

The item bundle, rather than individual items, maps to the scoring matrix and the design matrix. In this example, the bundle has eight possible response patterns (the number of representations of three items with two categories each) represented by eight categories. The eight response categories for our example bundle are shown below. The values in the parentheses show the scores for the individual dichotomous items.

Category 1 (0,0,0)
Category 2 (0,0,1)
Category 3 (0,1,0)
Category 4 (0,1,1)
Category 5 (1,0,0)
Category 6 (1,0,1)
Category 7 (1,1,0)
Category 8 (1,1,1)

A *partially ordered* (i.e., we can differentiate between bundle *scores* of 0, 1, 2, or 3, but not between bundle *categories* 2, 3 and 4 or *categories* 5, 6 and 7) item bundle scoring matrix is shown below:

Partially Ordered Bundle Scoring Matrix:

$$\begin{array}{ll} & D_1 \\ \text{Category 1 (0,0,0)} & \begin{bmatrix} 0 \end{bmatrix} \\ \text{Category 2 (0,0,1)} & \begin{bmatrix} 1 \end{bmatrix} \\ \text{Category 3 (0,1,0)} & \begin{bmatrix} 1 \end{bmatrix} \\ \text{Category 4 (0,1,1)} & \begin{bmatrix} 1 \end{bmatrix} \\ \text{Category 5 (1,0,0)} & \begin{bmatrix} 2 \end{bmatrix} \\ \text{Category 6 (1,0,1)} & \begin{bmatrix} 2 \end{bmatrix} \\ \text{Category 7 (1,1,0)} & \begin{bmatrix} 2 \end{bmatrix} \\ \text{Category 8 (1,1,1)} & \begin{bmatrix} 3 \end{bmatrix} \end{array}$$

Note that when using a partially ordered scoring matrix, one cannot derive the original responses from the score, as is possible with a fully ordered scoring matrix.

Partial Credit Bundle Design Matrix:

	δ_1	δ_2	δ_3
Category 1 (0,0,0)	0	0	0
Category 2 (0,0,1)	1	0	0
Category 3 (0,1,0)	1	0	0
Category 4 (0,1,1)	1	0	0
Category 5 (1,0,0)	1	1	0
Category 6 (1,0,1)	1	1	0
Category 7 (1,1,0)	1	1	0
Category 8 (1,1,1)	1	1	1

This is another type of partial credit model, and the design matrix would, again, follow from Example 2, the Unidimensional Partial Credit Model. This design matrix (as shown above) could have three columns, one for each score category. Alternatively, one could design a saturated design matrix with a parameter for each response category. As shown below, a saturated design matrix would have seven columns.

Saturated Bundle Design Matrix:

	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7
Category 1 (0,0,0)	0	0	0	0	0	0	0
Category 2 (0,0,1)	1	0	0	0	0	0	0
Category 3 (0,1,0)	0	1	0	0	0	0	0
Category 4 (0,1,1)	0	0	1	0	0	0	0
Category 5 (1,0,0)	0	0	0	1	0	0	0
Category 6 (1,0,1)	0	0	0	0	1	0	0
Category 7 (1,1,0)	0	0	0	0	0	1	0
Category 8 (1,1,1)	0	0	0	0	0	0	1

Multidimensional Models

For multidimensional models, means are needed for each dimension and the complete variance-covariance matrix is needed.

Multidimensional Means Vector:

$$\begin{matrix} D_1 & D_2 \end{matrix}$$
$$\begin{bmatrix} 0.542 & .865 \end{bmatrix}$$

Multidimensional Covariance Matrix:

$$\begin{matrix} & D_1 & D_2 \\ D_1 & \\ D_2 & \end{matrix} \begin{bmatrix} 1.260 & 0.783 \\ 0.783 & 0.812 \end{bmatrix}$$

Note that this covariance matrix corresponds to the following correlation matrix:

$$\begin{matrix} & D_1 & D_2 \\ D_1 & \\ D_2 & \end{matrix} \begin{bmatrix} 1.000 & 0.774 \\ 0.774 & 1.000 \end{bmatrix}$$

5. Between-Item Multidimensional model.

This is our first example of an assessment that measures proficiency on multiple dimensions, or constructs. We begin with the simple case in which each response is associated with just one of the dimensions. For example, some items may provide evidence of students' knowledge in one area while others provide evidence about a different domain. The items shown in Figure 11 come from the same assessment (Songer, 2000).

1. Some animals change color with the seasons. The snowshoe rabbit is brown during the warm seasons, what color would you expect its coat to be in the winter?

   A. tan
   B. black
   C. white
   D. a mixture of colors

7. Given the food chain: Grain → Mice → Snakes. What will happen when there is a lot of grain?

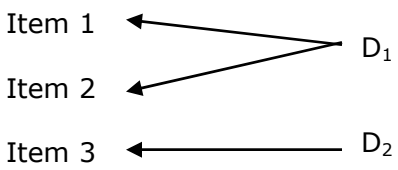
   When there is a lot of grain:

17. Jesse learned that she needs to apply a larger force than the frictional force to move an object. When Jesse applied 5 N to a brick on the desk, the brick did not move. With 10 N the brick moved. Which of the following can explain Jesse's experiment?

   A. 5 N is larger than the frictional force on the brick.
   B. 5 N is smaller than the frictional force on the brick.
   C. 10 N is the same as the frictional force on the brick.
   D. 10 N is smaller than the frictional force on the brick

Figure 11. Example of items measuring different constructs.

In this case, each item maps to a single dimension, so representing a single item is the same as for the Unidimensional Partial Credit Model (Example 2). However, when we construct an *assessment* that measures proficiency on two or more dimensions, we need to relate individual items to specific dimensions. In this example, we have an assessment with three items. Items 1 and 17 have two categories, while item 7 has three categories. Items 1 and 7 map to the first dimension, biodiversity, and item 17 maps to the second dimension, simple machines. The scoring and design matrices below represent an assessment comprised of these three items.



Assessment Scoring Matrix:

	D_1	D_2
Item 1, category 1	0	0
Item 1, category 2	1	0
Item 7, category 1	0	0
Item 7, category 2	1	0
Item 7, category 3	2	0
Item 17, category 1	0	0
Item 17, category 2	0	1

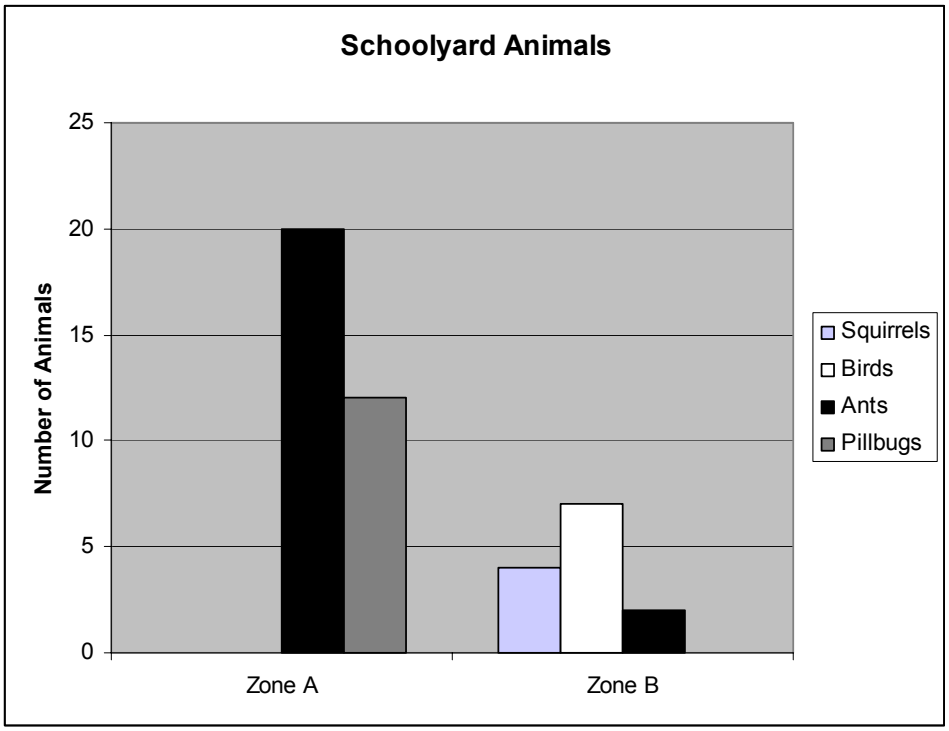
Assessment Design Matrix:

	δ_{11}	δ_{21}	δ_{22}	δ_{31}
Item 1, category 1	0	0	0	0
Item 1, category 2	1	0	0	0
Item 7, category 1	0	0	0	0
Item 7, category 2	0	1	0	0
Item 7, category 3	0	1	1	0
Item 17, category 1	0	0	0	0
Item 17, category 2	0	0	0	1

6. Within-Item Multidimensional Partial Credit model.

This model differs from Example 5 in that a single response can be associated with more than one dimension. For example, a single response to an open ended problem may provide evidence of a respondent's content knowledge and also his or her ability to formulate an explanation. One way to evaluate this type of item is to give two scores, one for the content construct and one for the explanations construct. An example of this type of item from the BioKIDS curriculum (Songer, 2000) is shown in Figure 12.

Using the graph below, predict which zone most likely has a tree in it and give one reason to support your prediction.



I think that zone _____ has a tree in it because _____

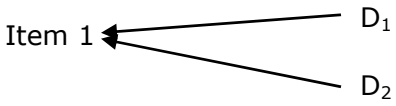
_____.

Figure 12. Example of within-item multidimensionality.

In this example, the selection of the zone is considered an indicator of content knowledge (in this case, biodiversity) and the explanation is an indicator of knowledge about building an explanation. A single item provides evidence of the respondent's location on both constructs.

Each dimension may have a different number of categories. For example, *content knowledge* may have two categories (correct and incorrect) and *building explanations* may have three categories, resulting in 6 unique combinations of responses on the item overall.

The first category of the overall item represents the situation in which the respondent had a response in the first category on the first dimension and a response in the first category on the second dimension. We construct the complete set of overall item categories by building permutations on the combinations of responses on the two dimensions. For proficiency estimation purposes we do not consider the initial item response categories again; only the overall response category information is sent to the Scoring Engine.



Final Item Scoring Matrix:

	D_1	D_2
Category 1 (0,0)	0	0
Category 2 (0,1)	0	1
Category 3 (0,2)	0	2
Category 4 (1,0)	1	0
Category 5 (1,1)	1	1
Category 6 (1,2)	1	2

Saturated Item Design Matrix:

δ_1	δ_2	δ_3	δ_4	δ_5
0	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

There are a number of options for generating design matrices for this example. The simplest is to assume the saturated model, as shown above. Another straightforward one is to assume no interaction effects between the difficulty of the task and the dimensions and to treat the new item response as a normal partial credit item with three steps.

Item design matrix:

$$\begin{array}{l} \\ \text{Category 1 (0,0)} \\ \text{Category 2 (0,1)} \\ \text{Category 3 (0,2)} \\ \text{Category 4 (1,0)} \\ \text{Category 5 (1,1)} \\ \text{Category 6 (1,2)} \end{array} \begin{array}{c} \begin{matrix} \delta_1 & \delta_2 & \delta_3 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

Another approach is to create parameters associated with the constructs. For example, a response in the second category may be associated with the difficulty of achieving a response at step 1 on the second dimension for the aggregate item, denoted $\delta_{D2,1}$ in the design matrix below. In this case, the design matrix parameters simply reflect the combined difficulty of getting the two response categories, one for each dimension. For example, to achieve an overall response in the third category, the respondent needs enough ability to achieve at the third category level on the second dimension ($\delta_{D2,1}$ + $\delta_{D2,2}$) but no incremental ability for the first dimension is required.

Item design matrix for parameters associated with dimensions:

$$\begin{array}{l} \\ \text{Category 1 (0,0)} \\ \text{Category 2 (0,1)} \\ \text{Category 3 (0,2)} \\ \text{Category 4 (1,0)} \\ \text{Category 5 (1,1)} \\ \text{Category 6 (1,2)} \end{array} \begin{array}{c} \begin{matrix} \delta_{D1,1} & \delta_{D2,1} & \delta_{D2,2} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \end{array}$$

In some cases, the design matrix may need to change to reflect a more complex conceptualization of item difficulties that includes interaction effects between multiple

dimensions. For example, the first item parameter may represent the difficulty of the first dimension, conditioned on a response in the first category on the second dimension. The second parameter may represent the difficulty of getting a response in the second category on the second dimension, conditioned on a response in the first category on the first dimension. A complete discussion of parameterization options is beyond the scope of this report.

Determining whether item responses are dependent or independent usually requires empirical analysis. An MRCML analysis can be useful in determining which model provides the best fit to the data. Clearly, the manner in which items are calibrated must be reflected in the scoring and design matrices when proficiency estimates are subsequently requested of the Scoring Engine.

Similarly, the selection of between-item or within-item multidimensionality should also be empirically confirmed. While a task designer may have a hypothesis about how various constructs work together and whether responses are conditionally dependent or independent, an analysis of alternative models may provide additional information that leads to new insights about the processes involved in performance of the task.

7. Between-item multidimensional bundle example: Bundle is multidimensional with each component item mapping to a single dimension.

In some cases, individual responses are conditionally dependent and are also indicators of different constructs. For example, in an interactive assessment of physics knowledge, students are required to select an appropriate equation for solving a distance problem (item 1), place the correct values into the equation (item 2), and then compute

the total distance traveled (item 3). An example of this type of problem is shown in Figure 13.

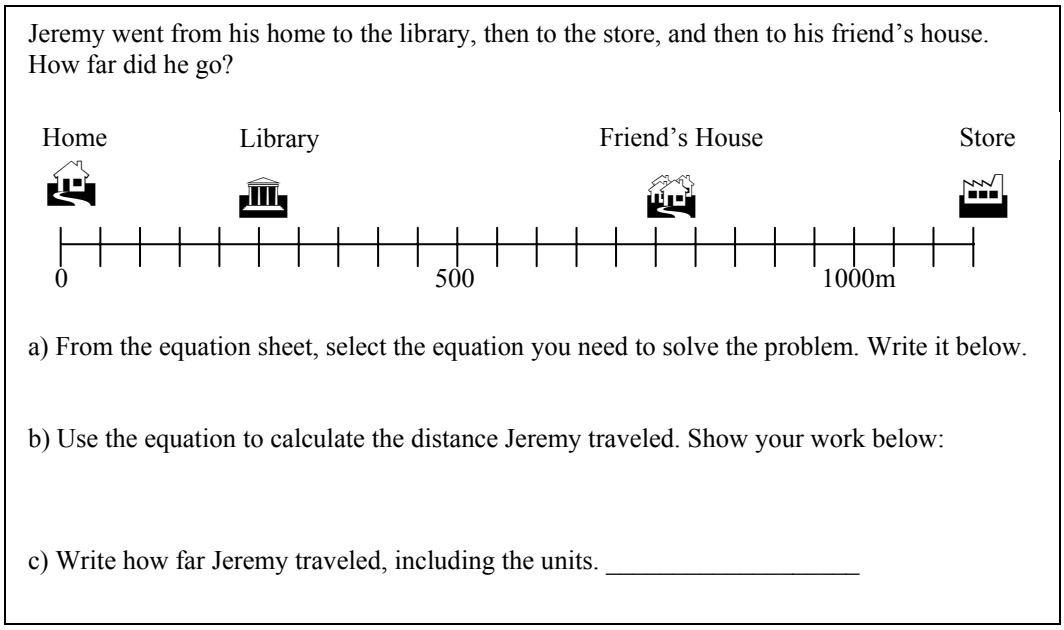
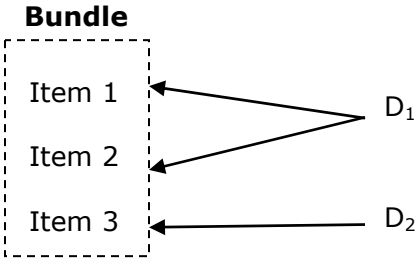


Figure 13. A between-item multidimensional bundle.

Clearly, the three item responses are conditionally dependent because selecting the wrong equation will usually lead to the wrong final answer, as will selecting the wrong values for the variables in the equation. However, selecting the equation and choosing the correct values for the variables provide evidence about the respondent's knowledge of physics while solving the equation provides evidence of mathematical ability. In this example, items 1 and 2 are indicators of the physics construct (the first dimension) and item 3 is an indicator of the mathematics construct (the second dimension). First, the three items are evaluated individually as correct or incorrect; or, as a response in the first category or a response in the second category. Then, the appropriate bundle category is determined from the pattern of responses on the three

items. Note that this example is similar to Example 5, but here the items are treated as conditionally dependent.



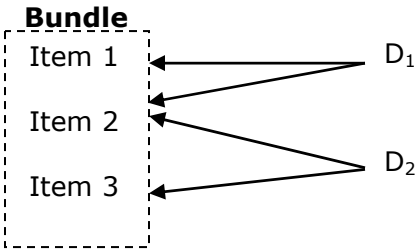
Bundle Scoring Matrix:

	D_1	D_2
Category 1 (0,0,0)	0	0
Category 2 (0,0,1)	0	1
Category 3 (0,1,0)	1	0
Category 4 (0,1,1)	1	1
Category 5 (1,0,0)	1	0
Category 6 (1,0,1)	1	1
Category 7 (1,1,0)	2	0
Category 8 (1,1,1)	2	1

The design matrix could follow any of the forms suggested in Example 5 above.

8. Within-item multidimensional bundle: Bundle is multidimensional with individual component items mapping to multiple dimensions.

If instead of associating each item to one construct we were to associate one item to multiple constructs in Example 7, we would need to construct a within-item multidimensional bundle. For example, we may believe that selecting the correct values to place into the equation (from Example 7) requires both physics knowledge and mathematical ability. In that case, item 1 is an indicator of the physics dimension, item 3 is an indicator of the mathematics dimension, and item 2 is an indicator of both physics and mathematics.



Bundle Scoring Matrix:

	D_1	D_2	
Category 1 (0,0,0)	0	0	
Category 2 (0,0,1)	0	1	from item 3 only
Category 3 (0,1,0)	1	1	from item 2 only
Category 4 (0,1,1)	1	2	from items 2 and 3
Category 5 (1,0,0)	1	0	from item 1 only
Category 6 (1,0,1)	1	1	from items 1 and 3
Category 7 (1,1,0)	2	1	from items 1 and 2
Category 8 (1,1,1)	2	2	from all items

The design matrix could follow any of the forms suggested in Example 6 above.

Conclusions

These examples show how a number of assessment tasks could be modeled for MRCML estimation. The BEAR Scoring Engine provides a mechanism for estimating respondent proficiencies from assessment data with tasks ranging from simple true-false questions to complex tasks involving a series of constructed responses that provide evidence of multiple constructs. Assessment developers can use the Scoring Engine software to develop more sophisticated interpretations of respondents' knowledge, behavior, or attitude than commonly produced using classical test theory or traditional IRT approaches.

References

Adams, R.J. & Wilson, M. R., (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard and M. Wilson (Eds.), *Objective measurement: Theory into practice (Vol. 3)*. Norwood, NJ:Ablex.

Adams, R., Wilson, M. and Wang, W. (1997) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). A four-process architecture for assessment delivery with connections to assessment design. *Journal of Technology, Learning and Assessment, 1(5).*

Hoskens, M., & De Boeck, P. (1997). A parameteric model for local dependence among test items. *Psychological methods, 2*, 261-277.

Kennedy, C.A. (2005). *The BEAR Assessment System: A Brief Summary for the Classroom Context.* BEAR Technical Report Series 2005-03-01. Berkeley, CA: University of California, BEAR Center.

Kennedy, C.A., Wilson, M. & Draney, K. (2005). *GradeMap (Version 4.1)* [computer program]. Berkeley, CA: University of California, BEAR Center.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 50*, 349-359.

Songer, N.B. (2000) BioKIDS: Kids' Inquiry of Diverse Species. Proposal funded by the Interagency Education Research Initiative (IERI).

Wang, W., Wilson, M. & Cheng, Y. (2000). *Local Dependence between Latent Traits when Common Stimuli are Used*. Paper presented at the International Objective Measurement Workshop, New Orleans, LA.

Wilson, M. (2005) *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Assoc.

Wilson, M. (1992) The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement. 16 (3)*. 309-325.

Wilson, M. & Adams R.J., (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181-198.

Wu, M., Adams, R. & Wilson, M. (2005). *ACER ConQuest*. Australian Council for Educational Research.