Assessing Local Item Dependence in Building Explanation Tasks

An Application of the Multidimensional Random Coefficients Multinomial Logit Item Bundle Model

Han Bao

University of Maryland, College Park

Robert J. Mislevy

University of Maryland, College Park

Correspondence:

Department of Measurement, Statistics & Evaluation

1230 Benjamin Building

University of Maryland, College Park, 20742-1115

Han Bao

Telephone: (301)405-4196

Fax: (301)314-9245

E-mail: hanbao@umd.edu

ABSTRACT

A common practice in assessment is to have a shared stimulus followed by a number of questions. One might doubt the usual assumption of standard item response theory of local item independence between items in these cases. Such a violation might contribute to the misfit of a unidimensional model, as an anticipated violation of conditional independence within these item bundles or testlets; one might consider a unidimensional model that incorporates local dependence. On the other hand, violations of local independence in a unidimensional model might in some cases be more satisfactorily solved with a multidimensional model with local independence. Even a multidimensional model with local dependence might be entertained. This paper discusses the application of Item Bundle Model developed by Wilson and Adams that can take account of multidimensionality and item dependence simultaneously. The use of the model is illustrated in the framework of the University of Michigan's BioKids 2002-2003 Fall Assessment.

Key words: item bundles, local item independence, multidimensional random coefficients multinomial logit (MRCML) model

Assessing Local Item Dependence in Building Explanation Tasks

An Application of the Multidimensional Random Coefficients Multinomial Logit Item Bundle Model

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant). We would like to thank Nancy Butler Songer and Amelia Wenk Gotwals of the BioKIDS project for permission to use the data in the example, and their time, patience, and insights in helping us understand the substantive aspects of the modeling described here. We are grateful to Larry Hamel and Cathy Kennedy for lots of help along the way.

INTRODUCTION

The cornerstone of item response theory (IRT) is the assumption of *local independence*, which posits that an examinee's response to a given test item depends on an unobservable examinee parameter _ but not on the identity of or responses to other items the examinee may have been presented (Lord, 1980). More formally, responses to test items are conditionally independent, given item parameters and _. This is a strong assumption. In some measurement situations, especially performance assessment, the items are grouped into bundles marked by shared common stimulus materials, common item stems, or common item structures. These situations call into question the assumption of local independence (Wilson & Adams, 1995). For example, item bundles are desirable in mathematics and science performance assessments because they reflect real life situations in which subproblems are interrelated and work is organized in steps. Solving one item in a bundle might increase the chances of solving the next. Even if a unidimensional model is appropriate for modeling responses, these interrelationships within clusters of related tasks constitute conditional dependence.

Moreover, sometimes additional and different content knowledge or skills are required in each problem step. In such cases the resulting local dependence in a unidimensional model may be better thought of as model misfit, as a multidimensional model would be a more appropriate and reduce or eliminate conditional dependence. It is useful, therefore, to distinguish between local dependence and departures from unidimensionality. This paper proposes the application of Item Bundle Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Adams, Wilson, & Wang,

1997) to check whether there is still dependence after taking dimensionality into consideration.

THEORETICAL FRAMEWORK

Much previous work on local independence exist in educational and psychological testing literature. Rosenbaum (1988) christened the concept of 'item bundle' to denote item subsets sharing common test stimulus and the idea of 'bundle independence', which is that bundle response patterns rather than individual items are conditionally independent given latent student variables. Wainer and Kiely (1987) used their notion of a "testlet", and suggested taking each testlet as an ordered polytomous item and using an ordered item response model to analyze the bundled scores. Wilson (1988) used the partial credit model and the rating scale model in a similar way. In order to provide the flexibility of customizing models for particular test situations, Wilson and Adams (1995) described an alternative approach to look at the bundle itself as source of data, and used the random coefficients multinomial logit model (RCML; Adam & Wilson, 1992) to investigate the violation of the conditional independence assumption. From nonparametric and factor analytic perspectives on IRT, a significant amount of research (e.g. Stout, 1987; Zhang,1996; and Zhang & Stout,1999) has been done to determine when proficiency unidimensionality is violated, estimate the degree of dependence, and assess the number of latent factors. Recently, from a Bayesian parametric point of view, Bradlow, Wainer and Wang (1999) modified standard IRT models to include an additional random effect for items nested within the same testlet to account for shared variation of items within testlets.

When a local dependence problem occurs in a set of test items that measures more than one latent ability, on the one hand, the violation of local item independence might suggest the existence of a new latent trait because of misfit of unidimensional model. On the other hand, the interdependence of the items might cause the dimensions to appear highly correlated, and thus the multidimensional analysis will provide the misleading results that only one dimension needs to be modeled. Is there dependence, dimensionality, or both?

By applying the concept of item bundles to multidimensional analysis, an item bundle model nested in a multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wang & Wilson, 1997) provides us opportunity to take account of multiple dimensions and item dependence simultaneously, by carefully modeling the expected patterns of dependence based on substantive knowledge about the structures and the demands of the tasks. Since local independence can always be achieved simply by increasing dimensionality as needed, the distinction is not mathematical. As our examples show, the investigation is instead an interplay between alternative mathematical models and what is known substantively about the contents and the forms of tasks.

The present paper briefly reviews the item bundle MRCML model, and then works through an example of a bundle and dimensionality framework using data from the project *BioKids: Kids Inquiry of Diverse Species*. Ideas are illustrated in some detail with two complex tasks, which required answering questions about a substantive situation then providing a scientific explanation of the situation. We focused on two item bundles, analyzing one at a time. For a given bundle, first under the unidimensional model, we fit the items bundled to model conditional dependence (i.e., bundle subitems in the task that is in focus) and again not bundled (maintaining the assumption of conditional independence among the subitems in that task), while keeping the same model for all the other items or item bundles in the test. This comparison allows for checking whether dependence exists in the bundle while positing unidimensional proficiency. Next, an analogous analysis was pursued with a multidimensional model suggested by the content demands of the items, to see whether dependence still existed after we take the issue of dimensionality into consideration. Structurally similar item bundles are built and analyzed in the same way as in the unidimensional comparison.

THE MODEL

The Item Bundle Multidimensional Random Coefficient Multinomial Logit Model

The MRCML model is a multidimensional extension of the random coefficients multinomial logit model, which can be applied to multidimensional polytomous test items. This paper will focus on item bundles. Suppose there is a set of C bundles; I_c is the number of items in each bundle and K_c is the total number of distinct response patterns in item bundle c, the number of all combinations of responses across all items in the bundle. For example, if a bundle is composed of one dichotomous item and one three-category item, there can be six distinct response patterns: (0,0), (0,1), (0,2), (1,0), (1,1) and (1,2). The probability of one particular response pattern *j* (e.g., *j* is the index over possible response patterns in a bundle) of bundle *c* can be modeled as

$$P(X_C = j \mid \vec{\theta}, A_C, B_C, \vec{\xi}) = \frac{\exp(\overrightarrow{b'_{cj}}\vec{\theta} + \overrightarrow{a'_{cj}}\vec{\xi})}{\sum_{k=1}^{K_C} \exp(\overrightarrow{b'_{ck}}\vec{\theta} + \overrightarrow{a'_{ck}}\vec{\xi})} \quad (1)$$

Where: $X_C = j$ is the person's response in category *j* of item bundle c;

1. Vector $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)$' defines a D-dimensional latent trait (if we take one-dimensional model as a special case of MRCML model, θ is a scalar in the equation);

2. The scoring vector $\overrightarrow{b_{cj}} = (b_{cj1}, b_{cj2}, \ldots, b_{cjD})'$ specifies the "performance level" to bundle response pattern *j* of bundle *c*. It can be collected into a bundle scoring sub-matrix $B_C = (\overrightarrow{b'_{c1}}, \overrightarrow{b'_{c2}}, \ldots, \overrightarrow{b'_{ck_c}})$, for bundle *c*, and further into a test scoring matrix $B = (B'_1, B'_2, \ldots, B'_C)$. The bundle scoring sub-matrix allows different numbers of categories for different item bundles and thus affords the possibility for the model to calibrate both dichotomous and polytomous item bundles simultaneously;

3. Vector $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_P)'$ is used to model the *p* bundle response patterns, which can be characterized as bundle difficulty, bundle step difficulty, etc;

4. The design vector $\vec{a}_{cj} = (a_{cj1}, a_{cj2}, \ldots, a_{cjp})'$ relates each observed response pattern to item bundle parameter vector $\vec{\xi}$, which can be gathered into bundle sub design matrix $A_C = (\overrightarrow{a'_{c1}}, \overrightarrow{a'_{c2}}, \ldots, \overrightarrow{a'_{cK_c}})$, with p columns, and which in turn can be gathered into the test design matrix $A = (A_1, A_2, \ldots, A_C)'$; the design matrix determines how the

model is specified for a collection of items as a whole.

As will be shown, appropriate choices of design matrices *B* and scoring matrices *A* allow us to express, and then compare, different conjectures about the nature of dependence and dimensionality affecting responses to items within and between bundles.

Bundle Independence

Suppose there is a set of C bundles and U_{nC} is person *n*'s response pattern for bundle *c*. Bundle independence can be defined in the following way:

$$P(U_{n1}, U_{n2}, \ldots, U_{nC} \mid \overrightarrow{\theta_n}, \vec{\xi}) = \prod_{c=1}^{C} P(U_{nc} \mid \overrightarrow{\theta_n}, \vec{\xi}), \tag{2}$$

Following Rosenbaum's idea, bundle independence means that examinee *n*'s response pattern U_{nc} on a set of items in a bundle are independent with those in other bundles given latent proficiency parameters $\overrightarrow{\theta_n}$ and item bundle parameters $\vec{\xi}$.

EXAMPLE

Data

The data set explored in this paper is from BioKids Fall 2003 Pretest Assessment. Totally, 220 students take the test. The *BioKids: Kids' Inquiry of Diverse Species* project at the University of Michigan (Prof. Nancy Butler Songer, Principal Investigator) is developing, testing, and organizing inquiry-focused, technology-rich science programs in biodiversity, weather, motion and other content areas to span from 5-8th grades. The assessment system includes tasks formulated around three different inquiry skills at different levels of complexity and scoring rubrics used to capture students' knowledge of both content and inquiry skills. In the current BioKids Fall 2003 Assessment, there are totally 19 items with three types of format: Multiple choices, Fill-in-blank, and Open-ended. Within these items, 16 focus on Biodiversity and the other 3 cover Simple Machines content. Since the development of sensitive inquiry assessment instruments is the central focus of the project, the assessment tasks are designed to evaluate student's understanding of inquiry thinking beyond content knowledge. There are four kinds of inquiry skills addressed in this assessment, including Hypothesis/Predictions, Explanations, Interpreting Data, and Re-express data. Each task may be associated with content, inquiry skills, or both. The two tasks addressed in the data analysis are shown in Appendix A.

Analysis Design

In the current analysis, we focus on the first 16 Biodiversity items and address 6 bundles in total. In order to examine the dependence and dimensionality issues, we found two item bundles (BioKids04 and BioKids05) with similar structures, each of them having a claim item and an evidence item. They can be thought of a special type of item bundle. At first, student is asked to make a claim using certain Biodiversity content knowledge; in the subsequent evidence part, the student is asked to explain the reasoning used to produce the previous answer where Biodiversity content knowledge and specified inquiry skill are both required. For both examples, the claim item is dummy coded with "0" representing for "incorrect" and "1" as "correct" and the evidence item is coded as (0,1,2) with "0" representing for "incomplete" answer, "1" as "partial correct" answer and "2" as "complete" answer. The Conquest computer program (Adams, Wilson &

Wang, 1997) was used to analyze the item bundle MRCML model. The dimensionality and dependency problems are addressed as follows:

Dimensionality. Three kinds of models are fitted to the data set: unidimensional model, two-dimension model and five-dimension model. First, if we regard content combined with inquiry skill as one latent trait, a one-dimensional model is being applied. Then, taking all kinds of inquiry skills as one dimension and the content knowledge as another dimension, we use two-dimension model. Some items have been identified by the BioKIDS content experts as depending mainly on content knowledge, others as depending mainly on inquiry skill, and others as requiring both. These judgments determine the *a* vectors for each item in the MRCMLM scoring matrix. Finally, if we associate each item with its specified inquiry skill, a five-dimension model with one content and four inquiry skills is fit to the data. For a given conjecture about local dependence within bundles, the Design Matrices of these three dimensionality models are all the same; the only differences are the Scoring Matrices, which relate the observed response or responses patterns to the latent trait(s).

Dependence. In order to address the dependence issue carefully, we compare two models concerning items within claim/evidence tasks: Not Bundled model vs. Bundled model, under three kinds of dimensionality situations separately. (Recall Bundled = locally *de*pendent; not bundled = *in*dependent.) As to the Not Bundled model: focusing on one item bundle, for example Biokids04, we only do not bundle this item. This means we take the claim item and evidence item of this bundle as two separate items, and bundle all the others in the test that have been proposed as candidates to be bundled. Table 1 and Table 2 show the design matrix and scoring matrix of a claim item and evidence items under conditions of the three dimensions. Note that in the 2- and 5-dimensional models, the claim item is posited to depend only on the content skill, while the evidence item is posited to depend on both the content and inquiry dimensions.

[[Table 1: Parameterization of Not Bundled Model of Biokids04 (claim)]]

[[Table 2: Parameterization of Not bundled Model of Biokids04 (Evidence)]]

Whereas, for the bundled model under the same unidimensional condition, we bundle in the same way all the other items that conjectured to be bundled but now *including* the one we are interested in. For the same example Biokids04, if we bundle the claim item and evidence item, there will be six possible response patterns with (0,0) as a reference pattern. We recode them using 0~ 5. The design matrix and scoring matrix of this bundle are shown in Table 3.

[[Table 3: Parameterization of Bundled Model of Biokids04]]

Obviously, the bundled model is more complex than the not bundled model because there are more item bundle parameters. Intuitively, the difference between the models is this: Under the not-bundled model, the difficulty of going from 0 to 1, then 1 to 2 on the evidence item is the same no matter what the response to the claim item was. These difficulty parameters are $b_{e=1}$ and $b_{e=2}$. Under the bundled model, the increments in the evidence steps can be different if claim=0 or claim=1, as indicated by $b_{e1|c=0}$ and $b_{e2|c=0}$, versus $b_{e1|c=1}$ and $b_{e2|c=1}$.

After we finished defining the whole test design matrix and test scoring matrix, item parameters of those two customized models can be estimated in Conquest. Therefore we can substitute the results into equation (1) for the bundled model to calculate the probability of each response pattern. As to the not bundled model, assuming the response to the claim item and that of evidence item are independent, the probability that an examinee get a response *j* for claim and a response *k* for the evidence is just the product of the two probabilities, or

$$p(X_e = k)p(X_c = j) = \frac{\exp(\overrightarrow{b_C}\vec{\theta} + \overrightarrow{a_C}\vec{\xi} + \overrightarrow{b_E}\vec{\theta} + \overrightarrow{a_E}\vec{\xi})}{\sum_{c=0}^{1}\exp(\vec{b}_C\vec{\theta} + \overrightarrow{a_C}\vec{\xi}) * \sum_{e=0}^{2}\exp(\vec{b}_E\vec{\theta} + \overrightarrow{a_E}\vec{\xi})} \qquad (3)$$

Then, in order to exam whether not-bundled model or bundled model fit the data, or in other words, to gauge the degree of local dependence, a Chi-square test can be used to compare observed and predicted response pattern counts. Two contingency tables corresponding to the bundled and not bundled models have been constructed over the same six cells: (0,0) (0,1) (0,2) (1,0) (1,1) (1,2), with the first element is associated with the response to the claim and the second with the response to the evidence. The observed proportions of examinees with correct answers in each cell are reported in the ConQuest initial analysis results. The predictions for each cell are obtained by using trait and item parameter estimates from ConQuest output as shown below (calculations were carried out using the computer program *Mathematica,* Wolfram, 2003). All item and person parameters of different models used are provided in Appendix B~D.

Assuming theta is normally distributed, the predicted proportion in cell *r* under the bundled model is obtained as:

$$E_r = \int p(x = j|\vec{\theta})p(\vec{\theta})d(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int\frac{\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) + \vec{b}\vec{\theta} + \vec{a}\vec{\xi})}{\sum_{k=1}^{K_C}\exp(\vec{b}\vec{\theta} + \vec{a}\vec{\zeta})}d(\vec{\theta}) \qquad (4)$$

For not bundled model, the prediction is obtained as:

$$E_r = \int p(x_C = k)p(x_E = j)p(\vec{\theta})d(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int \frac{\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) + \overrightarrow{b_C\theta} + \overrightarrow{a_C\xi} + \overrightarrow{b_E\theta} + \overrightarrow{a_E\xi})}{\sum(\overrightarrow{b_C\theta} + \overrightarrow{a_C\xi}) * \sum(\overrightarrow{b_E\theta} + \overrightarrow{a_E\xi})} d(\vec{\theta})$$

(5)

Under the one-dimensional solution, theta is a scalar representing an overall proficiency in the collection of tasks. Under the two-dimensional situation, two thetas are bivariate normally distributed, and the expected proportions are integrated over the joint range of both dimensions. Under the five-dimensional solution, each item requires content knowledge and specific inquiry skill(s). For example, BioKids04 is associated with Biodiversity content knowledge and the inquiry skill of Building Explanation from Evidence. Therefore, the expected proportions are dually integrated over the bivariate normal distribution of just these two thetas. The resulting contingency tables for Item 4 and Item 5 are listed in Appendix E. The expression used below to compute a goodness of fit Chi-square for each model is as follows:

$$\chi^2 = \sum_r \frac{(n_r - \hat{n}_r)^2}{\hat{n}_r} \tag{6}$$

Letting n_r be the observed frequency for each response category r; based on a total of N observations, the expected frequency for the rth response category is

$$\hat{n}_s = N * E_r \tag{7}$$

Where E_r is the expected proportion for each response category, which is computed in equation 4 or 5 as appropriate.

We were first interested in investigating whether conditional dependence exists between claim and evidence parts under unidimensional situation, looking at each of the two targeted items one at a time? Then what happens when we add more latent traits that are suggested by what we know about the items substantively? Items that are locally dependent in a lower-dimensional model may or may not be locally dependent in a higher dimensional model that is suggested by the substantive analysis of the items' demands.

Results

The Chi-square values for BioKids04 and BioKids05 are listed in Table 4 and Table 5 separately. For evaluations, the critical Chi-square value is 15.086 with degrees of freedom of 5 and significant level of .010. The results for Item 4 are shown in Table 4.

[[Table 4: Chi-Square results of Biokids04]]

For BioKids Item 4, under the one-dimension solution, since the Chi-square values are all less than the critical value of 15.086, both the Bundled and not Bundled models can be considered to fit the data adequately. Shifting attention to the two-dimensional solution, we can see that Chi-square values decreased especially for Bundled model, which suggests there does exist conditional dependence between the claim and evidence items in this bundle. Comparison with the one-dimensional result slightly indicates a better fit of the two-dimensional over the unidimensional model. When we move from two-dimension model to five-dimension model, the Chi-square value does decrease but not a lot, suggesting that the evidence part of this bundle only requires a general inquiry skill in the model as opposed to specific ones such as the Building Explanation from Evidence dimension associated with this item. In all, the simplest model considered, namely unidimensional with conditional independence, proves satisfactory for this item.

The results for BioKids Item 5 are shown in Table 5. We can reach conclusion that the bundled model fits the data much better than the not bundled model for the one-dimension, two-dimension and especially five-dimension solutions. Obviously, dependence does exist in this case even after we take a detailed dimensionality issue into consideration.

[[Table 5: Chi-Square results of Biokids05]]

How could we get different results for these two items, even though they have similar structures? A number of possibilities exist. The dependence issue also related to the idiosyncratic features of the tasks. For example, for some items it may be virtually necessary to provide a correct answer to the claim, or more content related item, in order to provide a good explanation, whereas for other items one might be able to provide a good explanation despite having stumbled on the specifics of the claim. Therefore, the two subitems seem less dependent. Or if the examinees have general science knowledge, they may be able to get correct answers for the claim part but not evidence part.

For our example, the first reason seems more plausible. With Biokids04, students will choose insects group on their first glance because a fly looks more like insects than spiders. Thus, even without having a certain amount of Biodiversity knowledge (such as flies and insects both have wings, six legs and antennae, etc.), they may give the correct answer to claim part but not evidence part since the latter needs more specialized inquiry skills as well as deeper content knowledge. However, as to Biokids05, students won't answer the claim item correctly if they don't know that trees are habitat of squirrels and birds. In other words, more content knowledge is demanded in the claim part of

Biokids05 than that of Biokids04; the two parts are more interrelated in the bundle Biokids05, thus there is more evidence of local dependence.

Beyond the Chi-Square analysis, the Deviance results provided by ConQuest as a model-fit index are also of interest. Deviance is approximately -2*log likelihood and a formal statistical test of relative fit of nested models can be undertaken by comparing the deviance of the two models. Table 6 lists the deviance information for all the models used in the analysis.

[[Table 6: Overall deviance results of bundled model and not bundled model under three dimensional situations]]

From the above table, on the one hand, we can see that the increases of deviances of the not bundled model vs. bundled model across three dimension situations for Item 4 are less than those of Item 5, which suggests a similar conclusion as the Chi-square analysis. The dependencies do seem to exist, especially for bundle BioKids Item 5.

On the other hand, using the deviance statistics for the three models can make the comparison of the relative fit of the unidimensional and multidimensional models. Since the one dimensional model is nested within the two and the two-dimensional model is nested within the five, the difference in the deviance between the two is distributed as approximately chi-square, with degree of freedom equal to the number of additional parameters in the more complex model. Having the same number of item parameters to estimate for a model with particular bundling configuration, the degree of freedom of our example is based on the differences caused by increasing of the number of dimensions. Since the thetas in each dimension are set to zero to identify the models, the increase in number of parameters is the number of additional variance and covariance elements.

Going from one to two dimensions adds two parameters; from two to five dimensions adds twelve.

For the one-dimension bundled model and the two-dimension bundled model, the difference in the deviance is 105.651, with two degrees of freedom. This is extreme statistically significant at the 0.01 level comparing to the critical value of $\chi^2_{(2,0.01)} = 9.21$, which means that two-dimension bundled model fits the data much better than one-dimension bundled model. However, as to the two-dimension bundled model vs. five-dimension bundled model, the difference in the deviance is 5.72, with 12 degrees of freedom. It is not statistically significant at the 0.01 level compared to the critical value of $\chi^2_{(12,0.01)} = 26.22$. The results tell us there are more than one dimensions in the assessment, and since the results are consistent under the different bundling options, the interdependence of the items does not cause misleading results of multidimensional analysis. In our example, the two-dimension bundled model with content knowledge as one dimension and an overall inquiry skill as another dimension fits the data satisfactorily.

DISCUSSION

When item response theory models are applied to test data with multi-part responses and mixtures of content demands, one has cause to investigate whether the assumptions on which the analysis is based are completely justified. This paper focuses on one of the main assumptions: the local independence. If the local independence assumption is not met, there is local dependence. In particular, we investigated situations that could involve both conditional dependence caused by a common problem situation and multidimensionality caused by different configurations of knowledge demands.

A multidimensional item bundle analysis suggest that the BioKids Fall 2003 Pretest exhibits item local dependence in at least one cluster of items based on a common stimulus situation. Taking dimensionality and idiosyncratic features of test format into consideration when expressing and testing various hypotheses about the nature of item dependence makes the analysis of local item independence more accurate and meaningful. Modeling such data using the MRCML item bundle model can deal with item dependence and dimensionality simultaneously, expressing and disentangling these sources of shared variation, thereby reduce distortions in item parameter estimates as well as proficiency estimates caused by ignoring these important features of patterns in data.

REFERENCES

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

Adams, R.J., Wilson, M. & Wang, W.C. (1997). The Multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* 21(1), 1-12.

Eric, T. Bradlow, Howard Wainer & Xiaohui, Wang. A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153-168.

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

Robert, J. Mislevy & Huahua, Chang. (1998) Does adaptive testing violate local independence? *CSE Technical report 476.*

Rosenbaum, P.R. (1988). Item bundles. *Psychometrika*, 53, 349-359.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49, 425-435.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52,589-617

Wendy, M. Yen (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.

Wendy, M. Yen (1984). Effects of local item dependence on the fit and equation performance of the three-parameter logistic model. *Applied Psychological measurement,* 8(2), 125-145

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, 12,353-364.

Wilson, M. & Adams, R.J. (1995). Rasch models for item bundles. *Psychometrika*, 60(2), 181-198.

Wolfram, S. (2003). *The Mathematica Book.* Stephen Wolfram, Fourth Edition.

Wu, M. L., Adams, R. J. & Wilson, M. (1997). Conquest: Generalized item response modeling software, Draft Release2. Hawthorn, Australia: ACER.

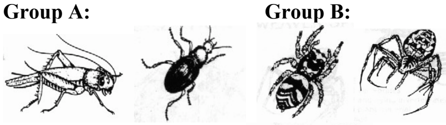
Zhang, J. & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64,213-249.

Appendix A

Two items of BioKids 2003 Fall Assessment

<table>
<tr>
<td>

4. Shan and Niki collected four animals from their schoolyard. They divided the animals into Group A and Group B based on their appearance as shown below:
Group A: Group B:



They want to place this fly in either Group A or Group B. Where should this fly be placed?

A fly should be in <u>Group A /Group B</u>
Circle one

Name two physical characteristics that you used when you decided to place the fly in this group:
(a)
(b)

</td>
<td>1

2

Total =
3</td>
<td>

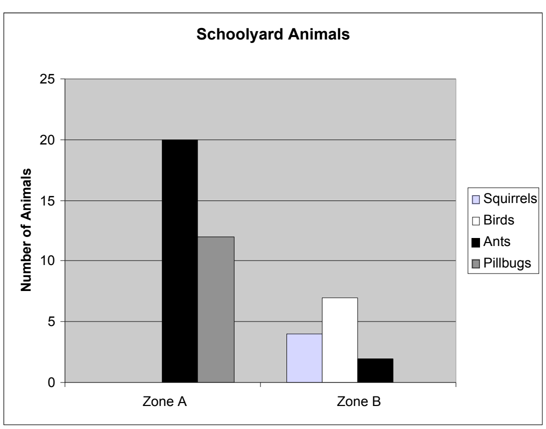
(Claim)
correct (1) – Group A
incorrect (0) – Group B, multiple circles or no response

(Data/Evidence)
complete (2) – two correct responses with no incorrect responses
partial (1) – one correct response; or two correct responses *with additional incorrect responses*
incomplete (0) – other responses or no response
Correct Responses include:

- **having six legs/how many legs**
- **having wings**
- **having three body parts**
- **being insects**
- **not being spiders**
- **having antennae**
- **spiders and insects are not in the same group**

</td>
<td>**Explanations – Step 2 Moderate**</td>
</tr>
<tr>
<td>

5. 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…

</td>
<td>**1**

2

total =
3</td>
<td>

(Claim)
Correct: Zone B
Incorrect: Zone A or blank

(Data/Evidence)
complete: (2) – Mentions both Animal (Bird OR squirrel) AND habitat function (live/found in trees, get food from trees, hide from predators in trees) with no incorrect responses
partial: (1) – Mentions EITHER Animal (Bird OR squirrel) OR habitat function (live in trees, get food from trees, hide from predators in trees); or mentions both animal and habitat function but with additional incorrect responses.
incomplete: (0) – other responses or no response

Disregard irrelevant responses related to “ants”

</td>
<td>**Interpreting Data – Step 2 Moderate**</td>
</tr>
</table>

Appendix B

Table 1

Parameterization of Not Bundled Model of Biokids04 (claim)

	(Sub) Design Matrix (B)	(Sub)Scoring Matrix (A)				
Claim Response Category	$b_{c=1}$	One-Dimension Model	Two-Dimension Model		Five-Dimension Model	
		_	_c	_I	_c	_e
0	0	0	0	0	0	0
1	1	1	1	0	1	0

Note: b_{c-1} is the difficulty parameter associated with response category 1, which is correct answer to claim Item; _ is the overall latent trait; _c is the content and _I is the overall inquiry skill in two-dimension model; _c is the content and _e is the specific inquiry skill of explanation of this item in five-dimension model.

Table 2

Parameterization of Not bundled Model of Biokids04 (Evidence)

	(Sub) Design Matrix (B)		(Sub) Scoring Matrix (A)				
Evidence Response Category	$b_{e=1}$	$b_{e=2}$	One-Dimension model	Two-Dimension model		Five-dimension model	
			_	_c	_l	_c	_e
0	0	0	0	0	0	0	0
1	1	0	1	1	1	1	1
2	1	1	2	2	2	2	2

Where: $b_{e=1}$ is the step difficulty parameter associated with response category 1, which is giving "partial correct" answer to evidence item; $b_{e=2}$ is the step difficulty parameter associated with response category 2, which is giving "complete correct" answer to evidence item;

Table 3

Parameterization of Bundled Model of Biokids04

Bundle Response Category		(Sub) Design Matrix (B)					(Sub)Scoring Matrix A)				
j	Response patterns	$b_{e1\|c=0}$	$b_{e2\|c=0}$	$b_{c=1}$	$b_{e1\|c=1}$	$b_{e2\|c=1}$	One-dimension model	Two-dimension model		Five-dimension model	
							_	_c	_l	_c	_e
0	00	0	0	0	0	0	0	0	0	0	0
1	01	1	0	0	0	0	1	1	1	1	1
2	02	1	1	0	0	0	2	2	2	2	2
3	10	0	0	1	0	0	1	1	0	1	0
4	11	0	0	1	1	0	2	2	1	2	1
5	12	0	0	1	1	1	3	3	2	3	2

Note: item bundle parameter $\vec{\xi}$ =($b_{e1|c=0}$, $b_{e2|c=0}$, $b_{c=1}$, $b_{ei|c=1}$, $b_{e2|c=1}$)', for example $b_{e1|c=0}$ represent for step difficulty parameter associated with response category of evidence equals 1 conditioning on claim equals 0, etc.

Table 4

Chi-Square results of Biokids04

	Observed Chi-Square value		
	One-dimension	Two-dimension	Five-dimension
Not bundled model	8.140	7.392	5.365
Bundled model	6.141	0.029	0.014

Table 5

Chi-Square results of Biokids05

	Observed Chi-Square value		
	One-dimension	Two-dimension	Five-dimension
Not bundled model	81.532	74.360	55.905
Bundled model	8.868	5.676	0.060

Table 6

Overall deviance results of bundled model and not bundled model under three dimensionality situations

	Bundled model Number of parameter: 46	Not bundled model for BioKids04 Number of parameter: 45		Not bundled model for BioKids05 Number of parameter:44	
	Deviance	Deviance	Compared to Bundled Model	Deviance	Compared to Bundled Model
One-dimension	5790.543	5794.083	+3.540	5872.148	+81.605
Two-dimension (Deviation decrease compared to one-dimension)	5684.892 (+105.651)	5674.573 (+119.51)	+10.319	5778.502 (+93.646)	+93.610
Five-dimension (Deviation decrease compared to two-dimension)	5679.172 (+5.72)	5689.051 (-14.478)	+9.879	5764.793 (+13.709)	+85.621

Note: “+” represents for increase and “-” represent for decrease.

Appendix C

Conquest output for Bundled model under three dimensional situations

one dimension Bundled		two dimension Bundled		Five dimension Bundled	
Deviance =	5790.641	Deviance =	5686.884	Deviance =	5679.172

Covariance for 1-dimension

Content and inquiry
0.48047 (0.04581)

Covariance for 2-dimension

Content	overall inquiry
0.42751 (0.04076)	0.19347 (0.04762)
0.19347 (0.04762)	0.58360 (0.05564)

Covariance for 5-dimewnsion

Content	hypothesis	explanation	interpreting data	reexpressing data
0.67492 (0.06435)	0.15761 (0.04709)	0.21863 (0.04884)	0.28128 (0.05053)	-0.04724 (0.06536)
0.15761 (0.04709)	0.36145 (0.03446)	0.25434 (0.03574)	0.15654 (0.03697)	0.20807 (0.04783)
0.21863 (0.04884)	0.25434 (0.03574)	0.38884 (0.03707)	0.16094 (0.03835)	0.18236 (0.04961)
0.28128 (0.05053)	0.15654 (0.03697)	0.16094 (0.03835)	0.41606 (0.03967)	0.01180 (0.05132)
-0.04724 (0.06536)	0.20807 (0.04783)	0.18236 (0.04961)	0.01180 (0.05132)	0.69624 (0.06638)

Regression Coefficients for 1-dimension

Content and inquiry
0.00000 (0.04673)

Regression Coefficients for 2-dimension

Content	overall inquiry
0.00000 (0.04408)	0.00000 (0.05150)

Regression Coefficients for 5-dimension

Content	hypothesis	explanation	interpreting data	reexpressing data
0.00000 (0.05539)	0.00000 (0.04053)	0.00000 (0.04204)	0.00000 (0.04349)	0.00000 (0.05626)

Table 7

Parameter estimates of bundled model

1-dimension	**2-dimension**	**5-dimension**
1 -1.90407	1 -2.18562	1 -2.17575
2 -1.50047	2 -1.72504	2 -2.03343
3 -0.50008	3 -0.56282	3 -0.58111
4 1.32805	4 1.52884	4 1.38513
5 -2.08820	5 -1.57776	5 -1.69133
6 0.41541	6 0.38344	6 0.37278
7 1.16510	7 1.63383	7 1.71286
8 3.52958	8 3.40131	8 3.37554
9 -1.47288	9 -0.96670	9 -1.09714
10 0.04798	10 0.76406	10 0.65355
11 -0.60458	11 -0.99410	11 -1.00746
12 -0.74532	12 -0.60158	12 -0.56441
13 0.05914	13 0.18771	13 0.24085
14 -0.91026	14 -1.06013	14 -0.67245
15 1.58958	15 2.26307	15 2.67936
16 -2.30890	16 -1.15813	16 -0.88845
17 -1.29099	17 -1.47512	17 -1.29660
18 1.63530	18 2.21090	18 2.40020
19 -0.52003	19 -0.54743	19 -0.21567
20 -1.64522	20 -1.85353	20 -2.05836
21 2.72203	21 3.05791	21 3.11325
22 0.60953	22 0.72645	22 0.70638
23 -0.84905	23 -1.18005	23 -1.18631
24 0.34344	24 0.57518	24 0.54919
25 1.19408	25 1.37626	25 1.34387
26 3.40875	26 4.07930	26 4.08991
27 0.79895	27 0.98743	27 0.90773
28 0.14235	28 0.30537	28 0.30519
29 2.33578	29 2.48952	29 2.55135
30 -0.10618	30 0.11131	30 0.02962
31 -1.18302	31 -1.36818	31 -1.37703
32 -0.24004	32 -0.83322	32 -1.02550
33 0.92285	33 0.51393	33 0.63782
34 -0.29741	34 -0.59332	34 -0.65791
35 -0.35421	35 -0.39235	35 -0.46921
36 1.58720	36 1.46080	36 1.58971
37 -0.76385	37 -0.77099	37 -0.93417
38 -0.37692	38 -0.19285	38 -0.43333
39 2.01567	39 1.72154	39 2.18346
40 3.06655	40 3.08824	40 3.38111
41 2.84262	41 2.93256	41 2.97868
42 -0.44232	42 0.10222	42 0.15976
43 0.86912	43 1.55436	43 1.72539
44 0.27090	44 0.33783	44 0.34303
45 2.45077	45 2.64392	45 2.63294
46 0.21930	46 0.84504	46 0.83820

Appendix D

Conquest output for Not Bundled model of BioKids04 under three dimensional situations

one dimension Bundled		**two dimension Bundled**		**Five dimension Bundled**	
Deviance =	5794.083	Deviance =	5694.573	Deviance =	5689.051

Covariance for 1-dimension

Content and Inquiry
0.53231 (0.05075)

Covariance for 2-dimension

Content	**overall inquiry**
0.51229 (0.04884)	0.21491 (0.04537)
0.21491 (0.04537)	0.44204 (0.04215)

Covariance for 5-dimewnsion

Content	**hypothesis**	**explanation**	**interpreting data**	**reexpressing data**
0.83732 (0.07984)	0.08671 (0.05481)	0.20030 (0.05345)	0.25570 (0.05454)	-0.09048 (0.07117)
0.08671 (0.05481)	0.39467 (0.03763)	0.22791 (0.03670)	0.12282 (0.03744)	0.19811 (0.04886)
0.20030 (0.05345)	0.22791 (0.03670)	0.37535 (0.03579)	0.12961 (0.03652)	0.14600 (0.04765)
0.25570 (0.05454)	0.12282 (0.03744)	0.12961 (0.03652)	0.39078 (0.03726)	-0.01259 (0.04862)
-0.09048 (0.07117)	0.19811 (0.04886)	0.14600 (0.04765)	-0.01259 (0.04862)	0.66535 (0.06344)

Regression Coefficients for 1-dimension

Content and Inquiry
0.00000 (0.04919)

Regression Coefficients for 2-dimension

Content	**overall inquiry**
0.00000 (0.04826)	0.00000 (0.04482)

Regression Coefficients for 5-dimension

Content	**hypothesis**	**explanation**	**interpreting data**	**reexpressing data**
0.00000 (0.06169)	0.00000 (0.04236)	0.00000 (0.04131)	0.00000 (0.04215)	0.00000 (0.05499)

Table 8

Parameter estimates of not bundled model of Biokids04

1-dimension	2-dimension	5-dimension
1 -1.91530	1 -2.21304	1 -2.18099
2 -1.50793	2 -1.75451	2 -1.98903
3 -0.49746	3 -0.60019	3 -0.55696
4 -1.95406	4 -1.98486	4 -2.02855
5 0.52774	5 0.47805	5 0.52876
6 1.21507	6 1.60558	6 1.80584
7 3.50326	7 3.37880	7 3.31299
8 -1.48113	8 -1.05464	8 -1.02737
9 0.01698	9 0.70239	9 0.59750
10 -0.59156	10 -0.98660	10 -0.97757
11 -0.72081	11 -0.64978	11 -0.49966
12 -0.00775	12 0.29291	12 0.25515
13 -0.95182	13 -1.03698	13 -0.64858
14 1.56723	14 2.26116	14 2.78927
15 -2.38561	15 -1.07265	15 -0.83566
16 -1.29042	16 -1.51187	16 -1.34422
17 1.63957	17 2.17706	17 2.48510
18 -0.50440	18 -0.50420	18 -0.27644
19 -1.62242	19 -1.95733	19 -2.03168
20 2.75973	20 2.99930	20 3.15009
21 0.62215	21 0.67798	21 0.75311
22 -0.86327	22 -1.20105	22 -1.19516
23 0.36447	23 0.52054	23 0.60851
24 1.21013	24 1.32484	24 1.38701
25 3.45508	25 4.04333	25 4.20736
26 0.78972	26 0.95172	26 0.90211
27 0.15083	27 0.28353	27 0.31705
28 2.33619	28 2.50625	28 2.42390
29 -0.06521	29 0.09421	29 0.21752
30 -1.16421	30 -1.42893	30 -1.35890
31 -0.14885	31 -0.92106	31 -0.86906
32 0.96082	32 0.52755	32 0.67203
33 -0.25226	33 -0.66850	33 -0.68438
34 -0.35015	34 -0.43119	34 -0.42145
35 1.56486	35 1.44237	35 1.61035
36 -0.78078	36 -0.79280	36 -0.88314
37 -0.37064	37 -0.27237	37 -0.33706
38 2.06071	38 1.64193	38 2.27140
39 3.16486	39 2.96775	39 3.51793
40 2.85409	40 2.89592	40 3.01191
41 -0.40731	41 -0.00715	41 0.26528
42 0.91832	42 1.51372	42 1.82579
43 0.28088	43 0.29225	43 0.38688
44 2.46706	44 2.58854	44 2.67920
45 0.26227	45 0.79095	45 0.94973

Appendix E

Conquest output for Not Bundled models of BioKids05 under three dimensional situations

one dimension Bundled	**two dimension Bundled**	**Five dimension Bundled**
Deviance = 5872.148	Deviance = 5778.502	Deviance = 5764.793

Covariance for 1-dimension

Content and Inquiry
0.62844 (0.05992)

Covariance for 2-dimension

Content	**Overall Inquiry**
0.71922 (0.06857)	0.22324 (0.04633)
0.22324 (0.04633)	0.32822 (0.03130)

Covariance for 5-dimewnsion

Content	**hypothesis**	**explanation**	**interpreting data**	**reexpressing data**
1.07843 (0.10282)	-0.00755 (0.06249)	0.18934 (0.05835)	0.29300 (0.05944)	-0.22108 (0.08577)
-0.00755 (0.06249)	0.39834 (0.03798)	0.18148 (0.03547)	0.08512 (0.03612)	0.23697 (0.05213)
0.18934 (0.05835)	0.18148 (0.03547)	0.34734 (0.03312)	0.10053 (0.03373)	0.12920 (0.04868)
0.29300 (0.05944)	0.08512 (0.03612)	0.10053 (0.03373)	0.36034 (0.03436)	-0.06248 (0.04958)
-0.22108 (0.08577)	0.23697 (0.05213)	0.12920 (0.04868)	-0.06248 (0.04958)	0.75040 (0.07155)

Regression Coefficients for 1-dimension

Content and Inquiry
0.00000 (0.05345)

Regression Coefficients for 2-dimension

Content	**Overall Inquiry**
0.00000 (0.05718)	0.00000 (0.03863)

Regression Coefficients for 5-dimension

Content	**hypothesis**	**explanation**	**interpreting data**	**reexpressing data**
0.00000 (0.07001)	0.00000 (0.04255)	0.00000 (0.03973)	0.00000 (0.04047)	0.00000 (0.05840)

Table 9

Parameter estimates of not bundled model of Biokids05

1-dimension	2-dimension	5-dimension
1 -1.94873	1 -2.22215	1 -2.25003
2 -1.53506	2 -1.75696	2 -2.10838
3 -0.50636	3 -0.58509	3 -0.61872
4 1.17722	4 1.37460	4 1.17979
5 -2.29295	5 -1.72889	5 -1.91666
6 0.44358	6 0.38851	6 0.36933
7 1.25467	7 1.63701	7 1.78221
8 -0.96382	8 -0.99098	8 -1.06569
9 0.52015	9 0.27539	9 0.12754
10 -0.82902	10 -0.67150	10 -0.57553
11 -0.12932	11 0.43214	11 0.28059
12 -1.05261	12 -0.98730	12 -0.84793
13 1.48529	13 2.37986	13 2.78410
14 -2.63009	14 -0.98531	14 -0.89808
15 -1.32786	15 -1.48843	15 -1.45245
16 1.62354	16 2.26686	16 2.47059
17 -0.55034	17 -0.46229	17 -0.33904
18 -1.58937	18 -1.97888	18 -2.11862
19 2.81911	19 3.04054	19 3.15016
20 0.63280	20 0.71055	20 0.69718
21 -0.90423	21 -1.21166	21 -1.26179
22 0.39039	22 0.56322	22 0.55372
23 1.22766	23 1.36318	23 1.35852
24 3.52789	24 4.12014	24 4.19932
25 0.74976	25 0.91528	25 0.82116
26 0.14960	26 0.30805	26 0.29253
27 2.31823	27 2.47870	27 2.36472
28 -0.03163	28 0.13618	28 0.20244
29 -1.11756	29 -1.39333	29 -1.36922
30 -0.01493	30 -0.92631	30 -0.87674
31 1.02168	31 0.60603	31 0.60913
32 -0.18293	32 -0.70600	32 -0.72125
33 -0.35631	33 -0.41354	33 -0.50953
34 1.50264	34 1.45076	34 1.49291
35 -0.85227	35 -0.78685	35 -1.02675
36 -0.41069	36 -0.22413	36 -0.49469
37 2.08172	37 1.68962	37 2.06142
38 3.26863	38 3.08525	38 3.30585
39 2.86465	39 2.92878	39 3.00921
40 -0.35184	40 0.05661	40 0.31220
41 0.99607	41 1.58084	41 1.89585
42 0.28645	42 0.32027	42 0.34872
43 2.48599	43 2.62899	43 2.63654
44 0.33021	44 0.85906	44 0.91728

Appendix F

Table 10

Contingency table of BioKids04: Bundled one-dimension model vs. Not bundled one-dimension model

Probabilities of each response pattern	Evidence / Claim	0	1	2
Observed	0	.1273	.0182	0
	1	.4227	.2955	.1364
Bundled Model	0	.0796	.0453	0
	1	.4699	.2676	.1177
Not bundled Model	0	.1273	.0182	.0199
	1	.4227	.2955	.1364

Table 11

Contingency table of BioKids04: Bundled two-dimension model vs. Not bundled two-dimension model

Probabilities of each response pattern	Evidence / Claim	0	1	2
Observed	0	.1273	.0182	0
	1	.4227	.2955	.1364
Bundled Model	0	.1280	.0183	0
	1	.4254	.2903	.1379
Not bundled Model	0	.0993	.0364	.0106
	1	.4579	.2718	.1240

Table 12

Contingency table of BioKids04: Bundled five-dimension model vs. Not bundled five-dimension model

Probabilities of each response pattern	Evidence / Claim	0	1	2
Observed	0	.1273	.0182	0
	1	.4227	.2955	.1364
Bundled Model	0	.1412	.0244	0
	1	.5320	.3657	.1456
Not bundled Model	0	.1192	.0467	.0113
	1	.5792	.3557	.1321

Table13

Contingency table of BioKids05: Bundled one-dimension model vs. Not bundled one-dimension model

Probabilities of each response pattern	Evidence / Claim	0	1	2
Observed	0	.2818	.0045	.0136
	1	.0864	.1636	.45
Bundled Model	0	.2070	.0038	.0140
	1	.1202	.1877	.4674
Not bundled Model	0	.1107	.0503	.1384
	1	.2592	.1177	.3239

Table 14

Contingency table of BioKids05: Bundled two-dimension model vs. Not bundled two-dimension model

Probabilities of each response pattern	Evidence / Claim	0	1	2
Observed	0	.2818	.0045	.0136
	1	.0864	.1636	.45
Bundled Model	0	.2988	.0052	.0147
	1	.0941	.1129	.4742
Not bundled Model	0	.1540	.0514	.0952
	1	.2194	.1191	.3609

Table 15

Contingency table of BioKids05: Bundled five-dimension model vs. Not bundled five-dimension model

Probabilities of each response pattern	Evidence / Claim	0	1	2
Observed	0	.2818	.0045	.0136
	1	.0864	.1636	.45
Bundled Model	0	.2831	.0047	.0136
	1	.0894	.1667	.4425
Not bundled Model	0	.1645	.0509	.0835
	1	.2095	.1205	.3712