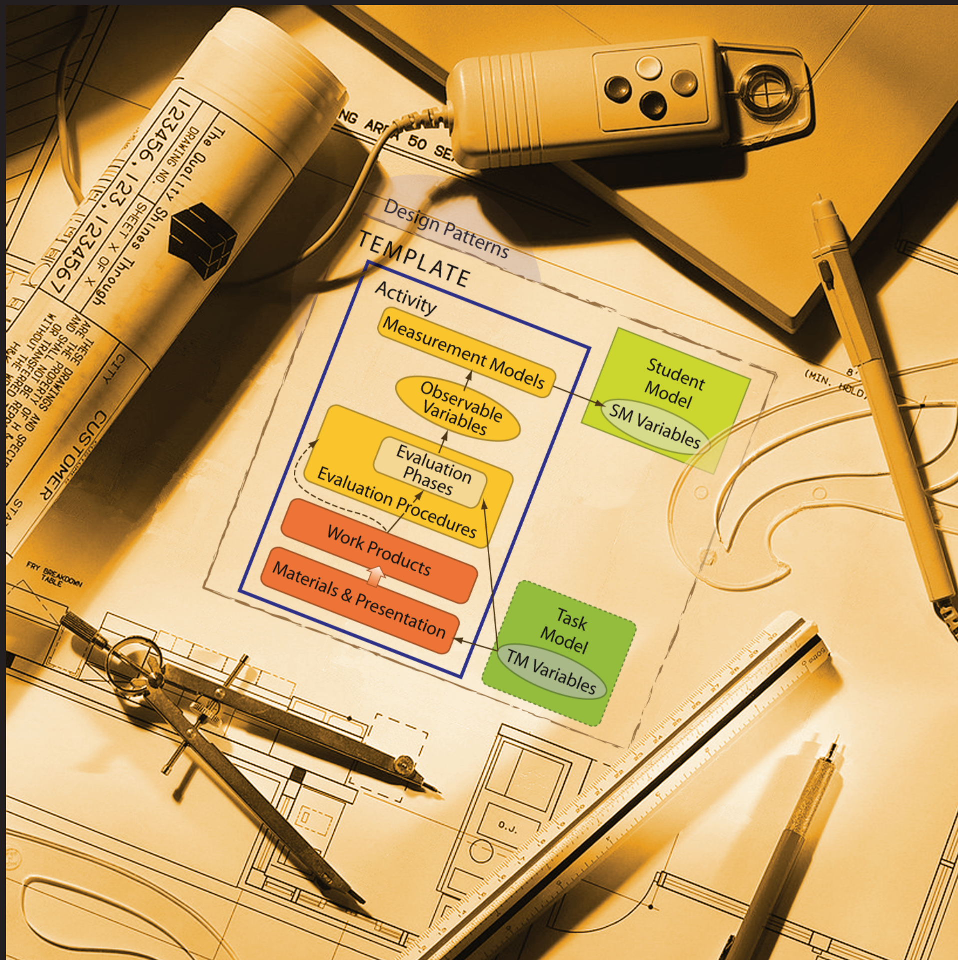


PADI Technical Report 14 | September 2006

Assessing Local Item Dependence in Building Explanation Tasks

PADI | Principled Assessment Designs for Inquiry

Han Bao, University of Maryland

Amelia Wenk Gotwals, Michigan State University

Robert J. Mislevy, University of Maryland

Report Series Published by SRI International

SRI International
Center for Technology in Learning
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
http://padi.sri.com

PADI Technical Report Series Editors
Alexis Mitman Colker, Ph.D., *Project Consultant*
Geneva D. Haertel, Ph.D., *Co-Principal Investigator*
Robert Mislevy, Ph.D., *Co-Principal Investigator*
Klaus Krause, *Technical Writer/Editor*
Lynne Peck Theis, *Documentation Designer*

P R I N C I P L E D   A S S E S S M E N T   D E S I G N S   F O R   I N Q U I R Y

T E C H N I C A L   R E P O R T   1 4

Assessing Local Item Dependence in Building Explanation Tasks

DRAFT

Prepared by:
Han Bao, University of Maryland
Amelia Wenk Gotwals, University of Michigan
Robert Mislevy, University of Maryland

Acknowledgments
This material is based on work supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant). We would like to thank Nancy Butler Songer, Principal Investigator of the BioKIDS project, for permission to use the data in the examples and for her time, patience, and insights in helping us understand the substantive aspects of the modeling described here. We are grateful to Lawrence Hamel, Kathleen Haynie, and Cathleen Kennedy, members of the PADI project team, for lots of help along the way.

C ONTENTS

Figures

TABLES

Abstract

A common practice in the assessment of science education is to have a shared stimulus followed by a number of questions. In these cases, one might doubt the usual assumption of standard Item Response Theory of local item independence among items that are supposed to measure the same latent proficiency. As an anticipated violation of conditional independence within these item bundles or testlets, such a violation might contribute to the misfit of a unidimensional model; 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 report discusses the extension and application of the Item Bundle Model developed by Wilson and Adams (1995) that takes into account multidimensionality and item dependence simultaneously. The use of the measurement model is illustrated in the framework of one of the examples of the Principled Assessment Designs for Inquiry (PADI) Project, namely, the University of Michigan's BioKIDS 2003 Fall Assessment.

1.0 Introduction

The cornerstones of item response theory (IRT) are the assumptions of *unidimensionality* and *local independence*. Unidimensionality is the assumption that the items measure only one common trait; local independence 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 that may have been presented to the examinee (Lord, 1980). More formally, it is asserted that responses to test items are conditionally independent, given item parameters and . Local independence can be violated even though there is unidimensionality. In some measurement situations, such as performance assessments in science, 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 chains, a special type of item bundle, are desirable in performance assessments measuring scientific inquiry skills because they reflect real life situations in which subproblems are interrelated within the same item stem 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.

Sometimes different types of 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 associated with dimensionality. A multidimensional model would be more appropriate and would reduce or eliminate conditional dependence. It is useful, therefore, to distinguish between local dependence and departures from unidimensionality. This paper proposes the extension of the Multidimensional Random Coefficients Multinomial Logit (MRCML) model (Adams, Wilson, & Wang, 1997) to item bundles, to check for item dependence after taking dimensionality into consideration.

Connecting the study of psychometric theory with practical application reflects the purposes of the Principled Assessment Designs for Inquiry (PADI) project. Recent research and science education reform documents (e.g., National Research Council, 1996) have given strong endorsements for inquiry-based science teaching and learning. However, science education researchers are still in the process of articulating the best methods for measuring both content knowledge and inquiry-reasoning abilities. To keep pace with inquiry-oriented views on teaching and learning, the assessment of science education via large- and small- scale tests has utilized alternative formats such as constructed response items and performance tasks. Some challenges to introducing more complex item types to science assessments are (1) how to measure complex reasoning involving both inquiry skills and content knowledge within assessment tasks when the unidimensionality assumption (that each task measures a single ability) is violated, and (2) how to deal with cases of assessing complex reasoning using bundled items/testlets where the local dependence assumption is violated. The PADI project, supported by the National Science Foundation to improve the assessment of science inquiry, is developing a task design framework with a particular focus on tasks that stress concepts and problem solving, building and using models, or cycles of investigation. The purpose of this report is to provide appropriate psychometric techniques to make more valid inferences about students' complex reasoning skills by application of one of the well-developed experimental examples of PADI project, namely, the BioKIDS 2003 Fall Assessment.

2.0 Theoretical Framework

Much previous work on local independence exists in the educational and psychological testing literature. Rosenbaum (1988) invented the concept of "item bundle" to denote subsets of items sharing a common test stimulus; he also described the idea of "bundle independence" for cases in which bundle response patterns, rather than individual items, are conditionally independent given latent student abilities. Wainer and Kiely (1987) used the notion of a "testlet" and suggested treating each testlet as an ordered polytomous item and using an ordered item response model to analyze the bundled scores. Wilson (1988) used a partial credit model and a rating scale model in a similar way. In order to allow customized models for particular test situations, Wilson and Adams (1995) described an alternative approach—treating the bundle itself as source of data and using the random coefficients multinomial logit model (RCMLM; Adams & Wilson, 1992) to investigate whether the conditional independence assumption is violated. Within nonparametric and factor analytic perspectives on IRT, a significant amount of research (e.g., Stout, 1987; Zhang & Stout , 1999) involves determining when the unidimensionality of examinee proficiency is violated, estimating the degree of dependence, and assessing the number of latent factors. Recently, using a Bayesian parametric perspective, Bradlow, Wainer and Wang (1999) accounted for shared variation of items within testlets by modifying standard IRT models to include an additional random effect for items nested within the same testlet.

When local dependence is detected in a set of test items that are assumed to measure one latent ability, the violation of local item independence might suggest the existence of an additional latent trait because of misfit of a unidimensional model. A set of interdependent items may reflect highly correlated latent traits, and thus dimensionality tests could provide the misleading result that only one dimension needs to be modeled. Determining the presence of local dependence or multidimensionality can be difficult in an interdependent set of items.

We will apply the concept of item bundles to multidimensional analysis. An item bundle model nested in a MRCML model (Adams, Wilson & Wang, 1997) provides a way to handle multiple dimensions and local dependence simultaneously. This is done by carefully modeling the expected patterns of dependence based on our substantive knowledge about the structures and the demands of the tasks. Because local independence always can be achieved simply by increasing dimensionality as needed, the distinction between local independence and multidimensionality 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 forms of tasks.

In this present report, we will review briefly the item bundle MRCML model and then work through an example of a bundle and dimensionality framework using data from the project *BioKIDS: Kids Inquiry of Diverse Species*. The ideas are illustrated in some detail with two complex tasks that require answering questions about a substantive situation and then providing a scientific explanation of the situation. We will focus on two item bundles, analyzing them one at a time. Focusing on each bundle in turn, we first assume unidimensionality, and we fit (1) the items bundled to model conditional dependence and

(2) the items not bundled to model conditional independence, keeping the same model for all the other items and item bundles in the test. This approach allows for detection of dependence in each item bundle, while assuming unidimensionality. Next, we consider the analysis of a multidimensional model suggested by the content demands of the items and determine whether dependence still exists after we take the issue of dimensionality into consideration.

3.0 The Model

3.1 The Item Bundle Multidimensional Random Coefficient Multinomial Logit Model

The multidimensional random coefficients multinomial logit model (MRCMLM) is a multidimensional extension of the random coefficients multinomial logit model (RCMLM), which can be applied to multidimensional polytomous test items (Adams & Wilson, 1996). Here we will present the model in terms of 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* (where *j* is the index over possible response patterns in a bundle) of bundle *c* can be modeled as

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

where $X_C = j$ is the person's response in category *j* of item bundle c, and

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

2. The scoring vector $\overrightarrow{b_{cj}} = (b_{cj1}, b_{cj2}, \cdots, b_{cjD})'$ specifies the "performance level" for bundle response pattern *j* of bundle *c*. It can be collected into a bundle scoring sub-matrix $B_C = (\overrightarrow{b'_{c1}}, \overrightarrow{b'_{c2}}, \cdots, \overrightarrow{b'_{ck_c}})'$, for bundle *c*, and further into a test scoring matrix $B = (B'_1, B'_2, \cdots, B'_C)$. 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. The vector $\vec{\xi} = (\xi_1, \xi_2, \cdots, \xi_P)'$ is used to model the *p* bundle response patterns, which can be characterized as bundle difficulty, bundle step difficulty, etc;

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

3.2 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:

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

Following Rosenbaum (1984, 1988), bundle independence means that examinee *n*'s response pattern U_{nc} on a set of items in a bundle are independent of the response patterns in other bundles given latent proficiency parameters $\overrightarrow{\theta_n}$ and item bundle parameters $\vec{\xi}$.

4.0 Example

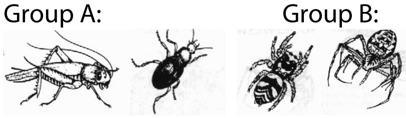
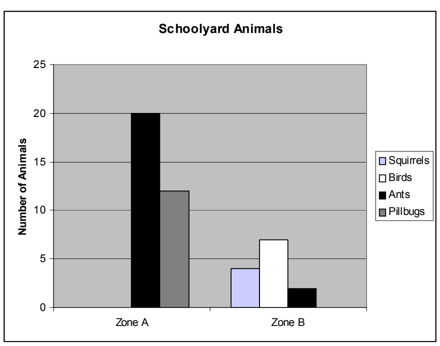
4.1 Data

The data set explored in this paper is from the BioKIDS Fall 2003 Pretest Assessment. A total of 220 students took the test. The *BioKIDS: Kids' Inquiry of Diverse Species* project at the University of Michigan (for more information, see <http://www.biokids.umich.edu>) is in the process of developing, testing, and organizing inquiry-focused, technology-rich science programs. The content of these programs is Biodiversity, Weather, Motion, and other topics, spanning grades five through eight. The assessment system includes: (1) tasks formulated around three different inquiry skills at different levels of complexity and (2) scoring rubrics used to capture students' knowledge of both content knowledge and inquiry skills. In the current BioKIDS Fall 2003 Pretest Assessment, there are 19 total items with three types of item formats: multiple choice, fill-in-the-blank, and open-ended. For this set of items, 16 cover Biodiversity content and 3 cover Simple Machines content. Because a central aim of the BioKIDS project is the development of sensitive assessment instruments for science inquiry, the accompanying assessment tasks are designed to evaluate students' understandings of content knowledge and inquiry skills. There are four kinds of inquiry skills addressed in this assessment: hypothesis/predictions, explanations, interpreting data, and re-expressing data. Each assessment task can be associated with content knowledge, inquiry skills, or both.

4.2 Analysis Design

In our analysis, we focus on 16 Biodiversity items that comprise 6 item bundles. In order to examine the dependence and dimensionality issues, we identified two item bundles (BioKIDS04 and BioKIDS05) with similar structures, each having a claim item and an evidence item (see Figure 1). These represent a special type of item bundle. For these bundles, students are prompted to make a claim using particular Biodiversity content knowledge; in the subsequent part, students are asked to explain their reasoning in producing the previous claim. Here, Biodiversity content knowledge and a particular inquiry skill are required. For both item bundles, the claim item is coded with "0" representing an incorrect answer and "1" representing a correct answer; the evidence item is coded (0, 1, 2) with "0" representing an incomplete answer, "1" as a partially correct answer and "2" as a complete answer. The ConQuest computer program (Wu, Adams, & Wilson, 1997) was used to analyze the item bundle using the MRCML model. In our analysis, we addressed issues of dimensionality and dependency as described below.

Figure 1. Two Items of BioKIDS 2003 Fall Assessment: BioKIDS04 and BioKIDS05

		▪	
BioKIDS04. 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 **Group A /Group B** Circle one Name two physical characteristics that you used when you decided to place the fly in this group: (a) (b)	1 2 Total = 3	(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	Explanations – Step 2 Moderate
BioKIDS05. 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…	1 2 Total = 3	(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"	Interpreting Data – Step 2 Moderate

Dimensionality. Three kinds of models were fit to the BioKIDS Fall 2003 Pretest Assessment data set: a unidimensional model, a two-dimensional model and a five-dimensional model. In the case of a unidimensional model, we regard content combined with inquiry skill as one latent trait. In the case of a two-dimensional model, we consider all inquiry skills as one dimension and content knowledge as a second dimension. In the case of a five-dimensional model, we associate each item with one of four inquiry skills [hypothesis/predictions, explanations, interpreting data, and re-expressing data] and one content dimension; here, some items depend mainly on content knowledge, some depend mainly on inquiry skills, and others require both. Content-experts' judgments of the content knowledge and inquiry skills required to respond to the items determined the *a* vectors for each item in the MRCMLM scoring matrix (e.g., see Formula 1). For a given conjecture about local dependence within bundles, the Design Matrices of these three models are all the same; the only differences are in the Scoring Matrices, which relate the observed response or responses patterns to the latent trait(s). The Scoring Matrix of the one-dimensional model is a n*1 matrix with each item/step (on each row) having an entry under the column representing the one dimension. The Scoring Matrix of the two-dimensional model is a n*2 matrix with each item/step (on each row) having two entries under the columns representing content knowledge and the combined inquiry skill, respectively. The Scoring Matrix of the five-dimensional model is a n*5 matrix with each item/step (on each row) having two entries, with one value under the column representing content knowledge and another value under one of four columns representing one of the four inquiry skills, while the remaining three columns would have values of 0.

Dependence. In order to address the dependence issue carefully, we compared two models of items within claim/evidence tasks: a Not Bundled model and a Bundled model. These were considered for each of the (one-, two-, and five-) dimensionality situations separately. Recall that the Bundled model assumes local *de*pendence and the Not Bundled model assumes local *in*dependence. For the Not Bundled model, we treat the claim and evidence items within the focal item bundle as two separate items, and bundle all the others that are proposed as item bundles in the test. The Design and Scoring Matrices of the BioKIDS04 claim item and evidence item are shown in Table 1 and Table 2, respectively. For the dichotomous claim item, there is only one difficulty parameter $b_{c=1}$ associated with response category 1; for the three-category evidence item, there are two step difficulty parameters $b_{e=1}$ and $b_{e=2}$ associated with response categories 1 and 2, respectively. Note that in the two- and five-dimensional models, the claim item is posited to depend only on content knowledge while the evidence item is posited to depend on both content knowledge and inquiry skill.

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	
		θ	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 the correct answer to the claim Item; Θ is the overall latent trait; θc is the content and θI is the overall inquiry skill in two-dimension model; θc is the content and θe is the specific inquiry skill of explanation of this item in five-dimension model.*

Table 2. Parameterization of Not Bundled Model of BioKIDS04 (Evidence)

Claim response category	(Sub) Design matrix (B) $b_{c=1}$	(Sub)Scoring matrix (A) One-dimension model θ	Two-dimension model θc	θI	Five-dimension model θc	θe
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 the correct answer to the claim Item; Θ is the overall latent trait; θc is the content and θI is the overall inquiry skill in two-dimension model; θc is the content and θe is the specific inquiry skill of explanation of this item in five-dimension model.*

The Design and Scoring Matrices of the bundled BioKIDS04 item are shown in Table 3. (For the Bundled Model under the unidimensional condition, we bundle all the items that are conjectured to be bundled, *including* the BioKIDS04 bundle.) For the BioKIDS04 bundle, we bundle the claim item and the evidence item, producing six possible response patterns with (0, 0) as a reference pattern. We recode the patterns as 0–5.

Table 3. Parameterization of Bundled Model of BioKIDS04

Claim response category	(Sub) Design matrix (B) $b_{c=1}$	(Sub)Scoring matrix (A) One-dimension model θ	Two-dimension model θc	θI	Five-dimension model θc	θe
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 the correct answer to the claim Item; Θ is the overall latent trait; θc is the content and θI is the overall inquiry skill in two-dimension model; θc is the content and θe is the specific inquiry skill of explanation of this item in five-dimension model.*

Table 3 also identifies the parameterization of the Bundled Model for BioKIDS04. The Bundled Model is more complex than the Not Bundled Model in that there are more item bundle parameters (3 parameters in Not Bundled, 6 parameters in Bundled). The Bundled Model and the Not Bundled Model differ qualitatively. For 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 $b_{e2|c=1}$. For the Not Bundled Model, we are asserting that the difficulty of increasing the evidence score an increment does not depend on what the response to the claim item was. These difficulty parameters are $b_{e=1}$ and $b_{e=2}$. This is the essence of the conditional dependence we are testing in this pair of items.

The next step in our analysis is to estimate item parameters of the two customized models in ConQuest. For the Bundled Model, we can substitute the results into equation (1) to calculate the probability of each response pattern. For the Not Bundled Model, assuming the responses to the claim item and the evidence item are independent, the probability that an examinee gets a response *j* for the claim and a response *k* for the evidence is the product of the two probabilities, or

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

In order to examine whether the models fit the data—that is, 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 are constructed over the same six cells: (0,0) (0,1) (0,2) (1,0) (1,1) (1,2). The first element is associated with the response to the claim, and the second is associated with the response to the evidence. The observed proportions of examinees with correct answers in each cell are reported in the ConQuest output. The predictions for each cell are calculated in the following two formulas by using trait and item parameter estimates obtained from ConQuest output; and the calculations were carried out using the computer program *Mathematica* (Wolfram, 2003). All the item parameters of the different models used are provided in Table 8 through Table 10. Three relevant summaries of ConQuest output are presented in the Appendix.

Assuming the latent trait (theta) is normally distributed, the predicted proportion in cell *r* under the Not Bundled Model is obtained from:

$$E_r = \int p(x_C = k)p(x_E = j)p(\vec{\theta})d(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int \frac{\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) + \overrightarrow{b_C}\vec{\theta} + \overrightarrow{a_C}\vec{\xi} + \overrightarrow{b_E}\vec{\theta} + \overrightarrow{a_E}\vec{\xi})}{\sum(\overrightarrow{b_C}\vec{\theta} + \overrightarrow{a_C}\vec{\xi}) * \sum(\overrightarrow{b_E}\vec{\theta} + \overrightarrow{a_E}\vec{\xi})} d(\vec{\theta}) \tag{4}$$

For the Bundled Model, the prediction is obtained from:

$$E_r = \int p(x = j|\vec{\theta})p(\vec{\theta})d(\vec{\theta}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int \frac{\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) + \vec{b}\vec{\theta} + \vec{a}\vec{\xi})}{\sum_{k=1}^{K_C}\exp(\vec{b}\vec{\theta} + \vec{a}\vec{\zeta})} d(\vec{\theta}) \tag{5}$$

For the one-dimensional solution, theta is a scalar representing one overall proficiency underlying task performance. For the two-dimensional solution, two thetas are bivariate normally distributed, and the expected proportion of each cell in the contingency table is integrated over the joint range of both dimensions. Under the five-dimensional solution, each item requires content knowledge and a specific inquiry skill. For example, BioKIDS04 is associated with Biodiversity content knowledge and the inquiry skill of Building Explanation from Evidence. Therefore, the expected proportion is dually integrated over the bivariate normal distribution of just these two thetas. The resulting contingency tables for BioKIDS04 and BioKIDS05 are listed in Table 11 through Table 16. The expression used below to compute a goodness of fit Chi-square statistic for each model is as follows:

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

Let 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 will first investigate whether conditional dependence exists between the claim and evidence parts under the unidimensional situation. We will begin by individually considering each of the two targeted items. We will then consider what happens when we add more latent traits, as suggested by our substantive knowledge about the items. 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.

4.3 Results

The Chi-square results and the ConQuest deviance (-2 log likelihoods) statistics of the BioKIDS04 and BioKIDS05 with Bundled and Not Bundled Models across one-, two-, and five-dimensional situations are discussed below. The AIC[1] (Akaike Information Criterion) and BIC[2] (Bayesian Information Criterion) indices for the one-, two- and five-dimensional Bundled Models are also provided.

The Chi-square values for BioKIDS04 and BioKIDS05 are listed in Table 4 and Table 5, respectively. The critical Chi-square value with 5 degrees of freedom and significance level of .010 is 15.086. The results for BioKIDS04 for the unidimensional solution, both the Bundled and Not Bundled Models, can be considered to fit the data adequately because the Chi-square values are all less than the critical value of 15.086. For the two-dimensional solution, we can see that the Chi-square of the Bundled Model is much less than the value of the Not Bundled Model, which suggests that there exists conditional dependence between the claim and evidence items in this bundle. Comparison with the one-dimensional result indicates a slightly better fit of the two-dimensional over the unidimensional model. The Chi-square value of the five- dimensional model is not much less than that of the two-dimensional model, suggesting that the evidence part of this bundle only requires a general inquiry skill in the model as opposed to a set of specific inquiry skills. In sum, for the BioKIDS04 task, the simplest model considered, unidimensional with no bundling proves satisfactory.

1. AIC is a measurement of model fit based on concepts derived from information theory.
2. BIC is a measurement of model fit based on framework of Bayesian argument.

Table 4. Chi-Square Results of BioKIDS04

Evidence response category	(Sub) Design matrix (B) $b_{e=1}$	$b_{e=2}$	(Sub) Scoring matrix (A) One-dimension model	Two-dimension model		Five-dimension model	
			θ	θc	θl	θc	θe
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;

For BioKIDS05, Table 5 shows that the Bundled Model fits the data much better than the Not Bundled Model for the one-dimensional, two-dimensional and especially five-dimensional 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

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: the item bundle parameter $\vec{\xi}=(b_{e1|c=0}, b_{e2|c=0}, b_{c=1}, b_{ei|c=1}, b_{e2|c=1})$, $b_{e1|c=0}$, for example, represents the step difficulty parameter associated with the evidence response category 1, conditioned on a claim response in category 0.

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 dimensionality situations

	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

From Table 6, noting that the Not Bundled Model is nested within the Bundled Model under the same dimensionality situation, we can see that the increases in deviances of the Not Bundled Model vs. Bundled Model across three dimensionality situations for BioKIDS04 are less than those of BioKIDS05, which suggests a similar conclusion as the Chi-square analysis. The dependencies do seem to exist for the task BioKIDS05.

In order to make the comparison of the relative fit of the unidimensional and multidimensional models, the AIC and BIC indices are appropriate for comparing model fit whether or not the models are nested. Due to this particular pattern of within-item multidimensionality[3] as expressed in the MRCMLM, the one-dimensional Bundled Model is not nested within the two-dimensional Bundled Model[4], while the two-dimensional Bundled Model is nested within the five-dimensional Bundled Model[5]. We may compare the AIC and BIC indices of all three models within a test occasion and additionally use the deviance values to compare the two- and five-dimensional solutions. The deviance, AIC, and BIC results are listed in Table 7. Having the same number of item parameters to estimate for a model with a particular bundling configuration, the degree of freedom of our example is based on the differences caused by increasing of the number of dimensions. Because the means of the thetas in each dimension are set to zero to identify the models, the increase in the number of parameters is the number of additional variance and covariance elements. Going from one to two dimensions adds two parameters; going from two to five dimensions adds twelve. (The Appendix provides the variance and covariance information about the Bundled Model and Not Bundled Model of BioKIDS04 and BioKIDS05.)

---

3. Within-item multidimensionality refers to an assessment where both content and inquiry skills are measured by a single item. This is different than between-item Multidimensionality, in which items in a given assessment only measure a single ability.
4. In going from the 2-d to the 1-d model (noting the same data and Design Matrix are used in both), proper nesting would hold if the row-wise sum of the two columns in the Scoring Matrix were proportional to the single column scoring matrix in the 1-d model. The idea is that if the two □s in the 2-d model were identical, the MRCMLM expression for the probability of all item responses in the 1-d and 2-d models would be the same. In this instance items loading on both dimensions in the 2-d model have sums twice that of the items loading on content only.
5. In going from 5-d to the 2-d model (again noting the same data and Design Matrix are used in both), proper nesting holds if the content columns of their Scoring Matrices are the same and if row-wise sum of the four columns in the 5-d model Scoring Matrix for inquiry θs are equal to the inquiry column of the Scoring Matrix in the 2-d model. If the four inquiry θs in the 5-d model were identical, the MRCMLM expression for the probability of all item responses in the 2-d and 5-d models would be the same. In this instance the required relationship holds because each item loads on exactly one inquiry skill.

Table 7. The Deviance, AIC and BIC of One-, Two- and Five- Dimensional Bundled Models

	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

The AIC and BIC indices of the two-dimensional Bundled Model are the smallest, This means that the two-dimensional Bundled Model fits the data better than the one-dimensional and five-dimensional Bundled Models. Additionally, as to the two-dimensional Bundled Model vs. five-dimensional Bundled Model, the difference in the deviance is 5.72, with 12 degrees of freedom. This is not statistically significant at the 0.01 level compared to the critical value of $\chi^2_{(12,0.01)} = 26.22$. Because a simpler model is preferred, the two-dimensional Bundled Model is more satisfactory, which is consistent with the AIC and BIC results.

All of these results tell us that there is more than one dimension in the assessment. In our example, the two-dimensional Bundled Model, with content knowledge as one dimension and an overall inquiry skill as a second dimension, fits the data satisfactorily.

Also, because the Bundled Model fits better than the Not Bundled Model under the different dimensionality situations, the multidimensional analysis does not reduce the interdependence of the items.

Table 8. Overall Deviance Results of Bundled Model and Not Bundled Model Under Three Dimensionality Situations

	Bundled model	Not bundled model for BioKIDS04		Not bundled model for BioKIDS05	
	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)	5694.573 (+99.51)	+9.681	5778.502 (+93.646)	+93.610
Five-dimension (Deviation decrease compared to two-dimension)	5679.172 (+5.72)	5689.051 (+5.522)	+9.879	5764.793 (+13.709)	+85.621

Note: "+" represents for increase and "-" represent for decrease.

Table 9. The Deviances, AIC and BIC of One, Two and Five Dimensional Bundled Models

Dimensions and number of parameters estimates	Deviance	AIC	BIC
One-dimension(46)	5790.543	5882.543	6038.650
Two-dimension(48)	5684.892	5780.892	5943.786
Five-dimension(60)	5679.172	5799.172	6002.790

Table 10. Parameter Estimates of Bundled Model

One-dimension	Two-dimension	Five-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

Table 11. Parameter Estimates of Not Bundled Model of BiKIDS04

One-dimension	Two-dimension	Five-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

Table 12. Parameter Estimates of Not Bundled Model of BioKIDS05

One-dimension	Two-dimension	Five-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

Table 13. Contingency Table of BioKIDS04: Bundled One-Dimensional 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 14. Contingency Table of BioKIDS04: Bundled Two-Dimensional Model vs. Not Bundled Two-Dimensional 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 15. Contingency Table of BioKIDS04: Bundled Five-Dimensional Model vs. Not Bundled Five-Dimensional 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

Table 16. Contingency Table of BioKIDS05: Bundled One-Dimensional Model vs. Not Bundled One-Dimensional 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 17. Contingency Table of BioKIDS05: Bundled Two-Dimensional Model vs. Not Bundled Two-Dimensional 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 18. Contingency Table of BioKIDS05: Bundled Five-Dimensional Model vs. Not Bundled Five-Dimensional 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

5.0 Discussion

5.1 Sources of Dependence in Science Assessments

Rosenbaum (1988) discussed the importance of distinguishing between the violation of local independence and the departure from unidimenisonality. Misfit of a unidimensional measurement model could be a reason for local dependence. However, our analysis suggests that there exists local dependence in set of tasks in the BioKIDS assessment across one-dimensional, two-dimensional, and five-dimensional situations. This indicates that the task format of bundling—in our case, item chaining—contributes mainly to the interdependence of items within the bundle.

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 is related to the idiosyncratic features of specific tasks. For some items, it may be necessary to provide a correct answer to the claim in order to provide a good explanation; whereas, for other items, a good explanation might be provided without providing the specifics of the claim. Also, if the examinees have general science knowledge, they may provide correct answers for the claim part but not evidence part. In these latter cases, the two subitems may be less dependent.

For our example, these reasons are plausible. With BioKIDS04 (see Figure 1), students may choose the group of insects based on surface characteristics because a fly looks more like the insects than the spiders. Without having much knowledge of Biodiversity (e.g., a fly fits with other insects because all have wings, six legs, and antennae), students may give the correct answer to the claim part of the item, but not the evidence part of the item because the latter needs more specialized inquiry skills as well as more sophisticated content knowledge. For BioKIDS05, students may not correctly answer the claim item unless they know that trees are the habitat of squirrels and birds. Therefore, more content knowledge is demanded in the claim part of BioKIDS05 than in that of BioKIDS04. Because the claim and the evidence parts of BioKIDS05 are more interrelated, there is a greater degree of local dependence.

5.2 The Multidimensional Nature of Science Knowledge

Scientific knowledge is inherently multidimensional; in order to fully comprehend and explore scientific situations, one must not only hold firm conceptual understandings but also be able to apply these understandings in performing scientific investigations, analyzing data, and proposing explanations. Thus, when measuring scientific knowledge, it is important to recognize the different aspects of students' understandings. If we do not assess multiple aspects of scientific knowledge, then there will be a large void in our knowledge of what students know and can do in science. The BioKIDS assessments were specifically designed to address both content knowledge as well as four inquiry reasoning skills. By examining a number of multidimensional models, we were able to confirm that the two-dimensional model did indeed fit the data satisfactorily, indicating that complex inquiry skills were utilized by students when interacting with the items. This finding confirms our belief that scientific knowledge has many dimensions and that when assessing science knowledge, it is important to have models that account for this phenomenon.

6.0 Conclusion

Both small- and large-scale assessments have become ubiquitous in today's science learning environment, and many of the assessments are analyzed using Item Response Theory models. When IRT models are applied to scientific assessment data with multi-part responses and mixtures of content demands, the typical assumptions of IRT should be investigated for accuracy. This paper focused on one of the main assumptions of IRT: local independence. 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.

Our multidimensional item bundle analysis suggests that at least one item cluster within the BioKIDS Fall 2003 Pretest Assessment exhibits local dependence based on a common stimulus situation. By taking dimensionality and idiosyncratic features of the item bundles into account when expressing and testing various hypotheses about the nature of item dependence, our analysis of local item independence was rendered more accurate and meaningful (Xie, 2001). Modeling such data using the MRCML item bundle model allows us to deal with item dependence and dimensionality simultaneously. Through our analyses, sources of shared variation were expressed and disentangled, thereby reducing distortions in item parameter estimates as well as proficiency estimates (e.g., caused by ignoring these important features of patterns in data).

References

Adams, R. A., & Wilson, M. (1992). *A random coefficients multinomial logit: Generalising Rasch models*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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.

Bradlow,E. T., Wainer, H. & Wang, X. H. (1999). 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 Associates.

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

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

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.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

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* (5th Edition). Champaign, IL: Wolfram Research, Inc.

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

Xie, Y. Y. (2001). *Dimensionality, dependence, or both? An application of item bundle model to multidimensional data*. Unpublished manuscript,.

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

Yen, W. M. (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

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

Appendix A: Summaries of ConQuest Output for Bundled Model under Three Dimensionality Situations

One Dimension Bundled	**Two Dimensions Bundled**	**Five Dimensions Bundled**
Deviance = 5790.641	Deviance = 5686.884	Deviance = 5679.172

Parameters estimated = ?

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)

APPENDIX B

Appendix B: Summaries of ConQuest Output for Not Bundled Model of BioKIDS04 under Three Dimensionality Situations

One Dimension Bundled	Two Dimensions Bundled	Five Dimensions 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)

Appendix C

Appendix C: Summaries of ConQuest Output for Not Bundled Models of BioKIDS05 under Three Dimensionality 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)

Sponsor
The National Science Foundation, Grant REC-0129331

Prime Grantee
SRI International. *Center for Technology in Learning*

Subgrantees
University of Maryland
University of California, Berkeley. *Berkeley Evaluation & Assessment Research (BEAR) Center* and *The Full Option Science System (FOSS)*
University of Michigan. *BioKIDs*