Graduate
Management
Admission
Council®

Bundle Models for Computerized Adaptive Testing in E-Learning Assessment

Kathleen Scalise
University of Oregon
and
Mark Wilson
University of California, Berkeley

Presented at the New CAT Models Paper Session, June 8, 2007

2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

A multifacet bundle model, herein called the “iota model” was used to estimate "pathway" parameters through partially hierarchical testlets. The model is useful for computerized adaptive assessment in e-learning contexts, when students are receiving individualized, or personalized, delivery of content based on embedded assessments. Testlets in this case are small bundles of items that act as questions and follow-up probes to interactively measure and assign scores to students. Research considered whether testlets can serve as a valid and reliable design to collect data and implement interactions for personalized delivery of content, whether path scores through the testlet modeled to a cognitive framework can be considered equivalent, and how three testlet designs compared in the quality and consistency of data collected. An example is shown that is multistage CAT, in which sequential or preplanned pathways are adaptively presented to students within the testlets based on student responses, and updating of θ and standard CAT algorithms can be used between the testlets.

Acknowledgments

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®. This paper is based on work supported by the National Science Foundation under Grant No. DUE: 0125651. The authors thank Sheryl Mebane, Nathaniel Brown, Jennifer Claesgens, and Angelica Stacy for their assistance and inspiration with instrument development, data collection, scoring, and interpretation of student learning patterns.

Citation

Scalise, K. & Wilson, M. (2007). Bundle models for computerized adaptive testing in e-learning assessment. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Kathleen Scalise, 170L College of Education, 5267 University of Oregon, Eugene, OR 97403-5267. Email: kscalise@uoregon.edu

Bundle Models for Computerized Adaptive Testing in E-Learning Assessment

Personalized Content in E-Learning

In e-learning contexts, it is becoming increasingly common to adapt the flow of materials so that each student receives content that is tailored, or personalized, to meet particular needs (Scalise & Claesgens, 2005; Taylor, 2002; Trivantis, 2005; Turker, Görgün, & Conlan, 2006). This is sometimes called dynamically delivered content or data-driven content, both of which use the acronym DDC. The motivation for differentiated instruction (Tomlinson & McTighe, 2006), whether differentiated through teacher intervention or use of other strategies such as computer-adaptive technology, includes that traditional curricular materials and assessments can lead to the production of inert learning activities, sometimes marginally responsive to where the student is in the knowledge acquisition cycle (Gifford, 1999; Hopkins, 2004). By comparison, differentiated instruction approaches are seen as moving teaching and learning activities toward the needs of the student. Technology can help teachers lower the resource barrier for differentiated instruction and also combine potentially powerful assessment tools with new information technologies to capture and analyze student data, rapidly deploy new media, facilitate collaboration, and provide other e-learning amenities such as asynchronous learning (Gifford, 2001; Parshall, Davey, & Pashley, 2000).

Technology to deliver differentiated instruction is now readily available, with back-end databases and a variety of multimedia-rich streaming techniques for which the flow of content to students can be adjusted in near real-time (Turker, Görgün, & Conlan, 2006) . However, the inferential machinery necessary to decide who should get what, and thus the measurement approaches and assessment techniques by which such inferences will be made, are mainly lacking or show limited development (Scalise et al., 2006; Timms, 2000). The usual measurement concerns of high quality data and inferences can quickly derail efforts to make such inferences in an accurate and speedy fashion (Osterlind, 1998; Wilson & Scalise, 2003), threatening to undermine the usefulness of dynamically personalized learning objects and products .

Many approaches have been taken to develop effective approaches to data-driven content, traditionally as in intelligent tutoring systems based on trying to capture how an expert instructor would assess students and assign material and attempting to code or program this "expert" knowledge into different types of expert systems. A variety of approaches have been taken such as information processing search space techniques, rule-based methods, and bug-based approaches (Russell & Norvig, 1995; Timms, 2000). However, major limitations include whether such systems can successfully capture the knowledge of experts and which expert's approaches, among many possibly competing models, might be preferable. Without effective ways to model, validate, compare, and test inferences about students, it is difficult to decide optimal directions for adaptivity. The proposed iota testlet model helps address concerns for a more robust evidence path.

A number of challenges to the quality of the evidence are apparent in e-learning adaptivity. First, since e-learning content is intended to be an effective learning experience, the measures for adapting the flow often need to take place “bundled,” or grouped, within a particular learning context around which lessons, or lesson components, are designed (Scalise & Gifford, 2006).

This will usually violate the local independence assumption most commonly used in assessment modeling (Wainer & Kiely, 1987; Wilson & Adams, 1995). Secondly, as the content flow changes for different students and students move into different areas of learning, it often makes sense for students to receive different item sets. This introduces computerized adaptive testing (CAT), and such concerns as equating, alternate forms, and comparing students based on a variety of dynamic subtests (Eignor, 1993). Finally, as the intent in e-learning is to take advantage of the rich potential of the computer to deliver a variety of new media content, including audio, video, animation, and simulation, e-learning developers often want to use new media formats in item designs that reflect the nature of the learning materials (Parshall, Davey, & Pashley, 2000; Parshall, Spray, Kalohn, & Davey, 2002). These item formats can be complicated to model and compare (Scalise & Wilson, 2006). In this paper we explore the use of testlets combined with a new multi-facet bundle measurement model called the iota model to address these challenges. An example is shown that is multistage CAT, in which sequential or preplanned pathways are adaptively presented to students within the testlets based on student responses, and updating of θ and standard CAT algorithms can be used between the testlets.

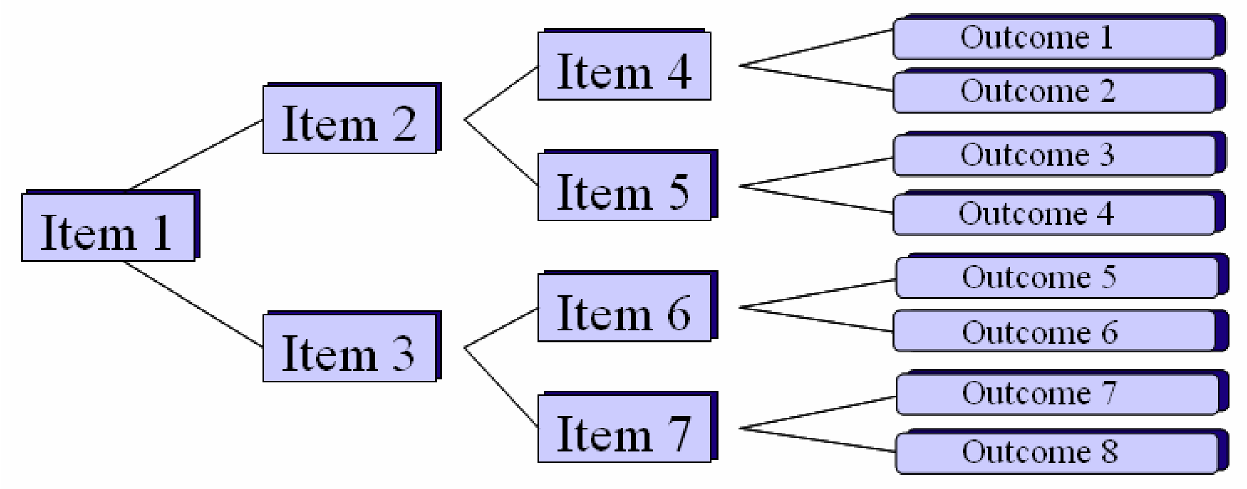
Testlets or Item Bundles

The item format we explore in this study is a CAT item type based on partially hierarchical item bundles, or testlets (Rosenbaum, 1988; Wainer & Kiely, 1987). Testlets have been described as useful for testing a variety of types of complex thinking (Haladyna, 1994). An introductory stimulus can be presented and then a variety of questions can be asked about the same or similar material, to probe the reasoning process. The stimulus for any item set can also vary, so that the item writer or content creator has tremendous opportunity for creative instructional designs, especially in technology settings, where the stimulus can be video, audio, animation, or simulation, as well as diverse learning objects such as works of art, written passages, cartoons, or references to an event, person or object, as Haladyna describes (1994). Although in 1994 Haladyna called the testlet an interesting format but futuristic, today the technology readily exists to implement the format in e-learning.

A testlet or an item bundle is a small bundle or group of assessment tasks and questions that share some common stimulus material. The term “item bundle” is often used in the measurement literature and most commonly refers in large-scale testing to a linear format called a “composite item”. These items include some common stimulus material, such as a reading passage with a sequence of subsequent questions based on the passage. All examinees receive the same set of questions, in the same order. The term “testlet,” by comparison, is the term more often used by technology developers, and usually refers to CAT versions of item bundles, in which not all students receive the same item bundle but rather parts of the bundle are delivered adaptively. We will call this kind of adaptive item bundle a “testlet” throughout this paper.

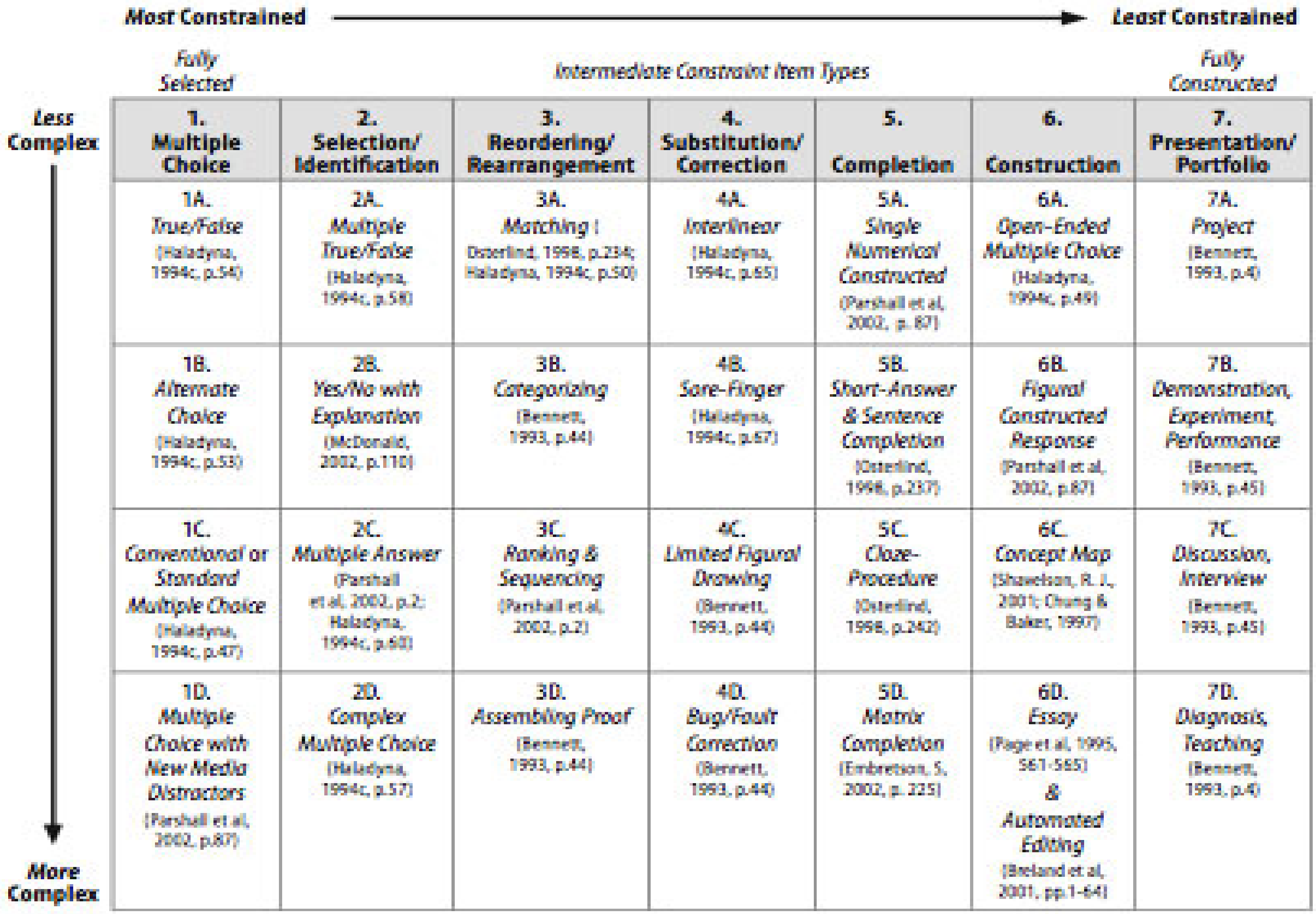
Testlets come in two formats: hierarchical and partially hierarchical (Wainer & Kiely, 1987). A diagram explaining the two formats is shown in Figure 1. All testlets begin with some common stimulus material, such as information and a question, and then students receive subsequent questions, or probes, adaptively depending on their responses. However, in hierarchical testlets, each path through the bundle of items has a unique score. In a fully hierarchical testlet, only one unique path through the bundle of items reaches each possible score outcome. In partially hierarchical testlets, different paths of questions and answers can achieve the same score. A partially hierarchical testlet would allow the same outcome or outcomes to be reached by multiple paths through the bundle of items, for instance if Items 4 and 5 both led to outcome 3.The partially hierarchical design is more flexible as reasoning facets can be tracked and then valued as similar or described as different, depending on expert advice and the results of data analysis.

Figure 1. Diagram of a Fully Hierarchical Testlet



A wide variety of item types can be used within a single testlet. In prior work we have described a “Taxonomy of Item Types," see Table 1 (Scalise & Gifford, 2006), which shows 28 item examples organized into a taxonomy based on the level of constraint in the item/task response format. The taxonomy describes 16 classes of item types with responses that fall somewhere *between* fully constrained responses (at left in Column 1) and fully constructed responses (at right in Column 7). We call these “intermediate constraint” items (Scalise & Gifford, 2006), which are organized with decreasing degrees of constraint from left to right. Fully constrained items include the traditional multiple-choice question, which is sometimes far too limiting to tap much of the potential of new information technologies, and fully constructed designs include the traditional essay, which remains a formidable challenge for computers to meaningfully analyze, even with sophisticated tools such as latent semantic analysis (Scalise & Wilson, 2006). The Intermediate Constraint Taxonomy describes and gives examples of 16 iconic intermediate constraint item types that feature a variety of innovations in the stimulus and/or in the response of the observation and might be useful, for instance, for automated scoring in computer-based testing. There is additional ordering in Table 1 that can be seen by what we call “within-type” when progressing down each column, with a general trend for the innovations to become increasingly complex from top to bottom. Any of these item types potentially could be combined into testlets to create a wide range of possibilities for instructional design of interactive assessments.

Table 1. Intermediate Constraint Taxonomy for E-Learning Assessments and Tasks



Most Constrained → **Least Constrained**

Fully Selected | *Intermediate Constraint Item Types* | *Fully Constructed*

	1. Multiple Choice	**2. Selection/ Identification**	**3. Reordering/ Rearrangement**	**4. Substitution/ Correction**	**5. Completion**	**6. Construction**	**7. Presentation/ Portfolio**
Less Complex	1A. *True/False* (Haladyna, 1994c, p.54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed* (Parshall et al, 2002, p. 87)	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
	1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
	1C. *Conventional or Standard Multiple Choice* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
More Complex	1D. *Multiple Choice with New Media Distractors* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) *&* *Automated Editing* (Breland et al, 2001, pp.1-64)	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

The Iota Model

Psychometrically, context effects and inter-item dependence are a threat to testlets, and need to be modeled by correct statistical models. Important sources for formal modeling options for testlets include Li, Bolt & Fu (2006), Wilson & Adams (1995) and Wainer & Kiely (1987). In testlet structures, clustered items usually are linked by attributes such as common stimulus material and common item stem, structure, and content (Wilson & Adams, 1995). This suggests that the usual assumption of conditional or local independence between items necessary for item response modeling is not met within the testlet. Local independence in item response models "means that the response to any item is unrelated to any other item when trait level [or student performance level] is controlled" (Embretson & Reise, 2000). The local independence assumption is commonly stated as

$$P(X_{is}=1\,|\,X_{js},\xi_i,\theta_s)=P(X_{is}=1\,|\,\xi_i,\theta_s) \qquad (1)$$

where X_{is} represents the score of student *s* on item i, X_{js} represents the score of the same student on another item j, ξ_I represents a vector of item parameters for item *i*, and θ_s represents the performance ability of student s. One approach to addressing such within-bundle dependence is to treat each bundle of dependent items as a single item, awarding degrees of partial credit over the testlet depending on level of overall performance indicated by the series of responses, which can be called the bundle response vector (Wilson & Adams, 1995).

Testlets previously have been psychometrically modeled in a variety of ways, most usually with some version of a partial credit model. The partial credit model is the more general of two polytomous Rasch models (Wright & Masters, 1982) commonly expressed according to Equation 2:

$$P(X_{is}=x\,|\,\theta_s)=\frac{\exp\sum_{j=0}^{x}\left(\theta_s-\delta_{ij}\right)}{\sum_{r=0}^{m_i}\exp\sum_{j=0}^{r}\left(\theta_s-\delta_{ij}\right)} \qquad (2)$$

for item *i* scored *x* = 0, …, m_i, where X_{is} is the score of student *s* on item *i*, *x* represents a given score level, θ_s represents the parameter associated with the person performance ability of student s, *r* in the denominator represents a summation of terms over the steps, and δ_{Ij} represents the parameter associated with the difficulty of the *j*th step of item *i*.

Another modeling approach is testlet response theory (Wainer, et al., 2006), with a 3-parameter logistic multi-faceted model “testlet effect," which is a special student ability that applies to all the items in a given testlet for that student. As Wainer describes, because the partial credit model collapses all response patterns over the testlet with the same number correct into the same category it can potentially lose information, whereas the testlet effect model preserves the individual item structure and retains the information. Alternative models for testlets (Li, Bolt, & Fu, 2006) treat the testlet effect as if it were another ability dimension in a multidimensional item response theory (IRT) model, with three different approaches to the general model, varying constraints on slope parameters and item discrimination parameters.

None of these models, however, directly address a critical question for CAT testlets in e-learning—is it enough to consider the final score achieved in a testlet or is the “path,” or series of adaptive items, by which the score is achieved also important to consider? The iota model we discuss here is of importance in partially hierarchical testlets, when more than one path through a bundle of items achieves the same score. The iota model tests the question of how significant the pathways through adaptive testlets are. It does so with the addition of an iota pathway parameter, ι_{ijp} over pathway p, where the summation of all ι_{ijp} for the n_i paths in a given item equals zero. The difficulty of a score level achieved according to a given path becomes $\delta_{ijp,}$ where $\delta_{ijp} = \delta_I + \iota_{ijp}$. This generates the iota model shown in Equation, which was used to model the testlet data we describe here:

$$P(X_{is} = x \mid \theta_s) = \frac{\exp \sum_{j=0}^{x} \sum_{p=0}^{n_i} \left(\theta_s - \delta_{ijp}\right)}{\sum_{r=0}^{m_i} \exp \sum_{j=0}^{r} \sum_{p=0}^{n_i} \left(\theta_s - \delta_{ijp}\right)} \tag{3}$$

The likelihood function for a standard IRT model is different from an item bundle model. In a standard IRT model the likelihood function is the product of the probabilities of scores achieved on the *items,* whereas for the bundle models such as the iota model it is the product of the probabilities of scores achieved on the *bundles.*

To give an example of this second difference, take for instance a test with four dichotomous items calibrated under the Rasch model. For each of the items 1 through 4, students would receive either a 0 or 1 score on each item. The likelihood function the item score vector achieved by a student would be

$$L = P_1(x_{1s} = y \mid \theta_s) P_2(x_{2s} = y \mid \theta_s) P_3(x_{3s} = y \mid \theta_s) P_4(x_{is} = y \mid \theta_s) \tag{4}$$

where P_n is the probability of achieving the score y that was actually achieved on item n by student s. Here the probabilities of achieving the score of either 0 or 1 on the items are multiplied together.

Now compare this to the situation in which the same four items are arranged in two bundles and calibrated under the iota model. Instead of receiving four item scores, each student would receive two bundle scores. If the first bundle consisted of items 1 and 2, students could achieve these possible patterns over the two items of the bundle as:

00 — student misses both items, coded as a score of "0" on the bundle
01 — student misses item 1 and achieves item 2, coded as a score of "1" on the bundle
10 — student achieves item 1 and misses item 2, coded as a score of "2" on the bundle
11 — student achieves both items, coded as a score of "3" on the bundle

The same situation would hold for the second bundle of two items, and the student would receive another score of 0-3 on this bundle, depending on their score pattern over the two items of the second bundle.

The likelihood function for a bundle model such as the iota model then for these data becomes

$$L = P_1(x_{1s} = y \mid \theta_s) P_2(x_{2s} = y \mid \theta_s) \tag{5}$$

where P_n is the probability of achieving the score *y* that was actually achieved on *bundle n* by student *s*. Here the probabilities of achieving the scores 0 - 3 that were actually achieved on each *bundle* are multiplied together.

Ideally, the ι_{ijp} and $\iota_{ijp'}$ components of the item difficulty should have relatively small differences over all paths for a given ι_{ij}, as equivalently scored item paths though the bundle should have near equal difficulties if the construct modeling assumptions hold. To model this assumption, the iota model, which is an ordered partition model (Wilson, 1992), is used. In this application, the various pathways to a single score within an item bundle are unique and, therefore, are given individual parameterizations even though they are scored the same. This model can be compared to the aggregate partial credit model for the data, in which the pathways are aggregated and treated as a single score, in order to gauge statistical significance of considering the iota pathway parameters individually as compared to the hierarchically more parsimonious partial credit model. The iota model is hierarchically nested within the partial credit model, and thus can be compared by a likelihood ratio chi-square test, with the difference in estimated parameters between the two models equal to degrees of freedom.

Method

In this paper we model testlet data from the University of California at Berkeley "Smart Homework" implementation of ChemQuery, an NSF-funded project in which one component consisted of data-driven content to individually tailor "smart" homework sets in high school and university level chemistry. The adaptivity approach of the homework sets is called BEAR CAT, since it draws on a CAT approach to the BEAR Assessment System (Wilson, 2005) for the specification of properties, or variables, of interest to measure. The BEAR CAT approach is a multistage CAT, in which sequential or preplanned pathways are adaptively presented to students within the testlets based on student responses, and updating of θ and standard CAT algorithms can be used between the testlets.

The variables that the Smart Homework example measures are the *Perspectives of Chemists* (Claesgens, Scalise, Draney, Wilson, & Stacy, 2002) developed to measure understanding in chemistry by experts in the UC Berkeley Chemistry Department and School of Education (Wilson & Scalise, 2003). An overview of the first variable in this construct, Matter, is shown in Figure 2.

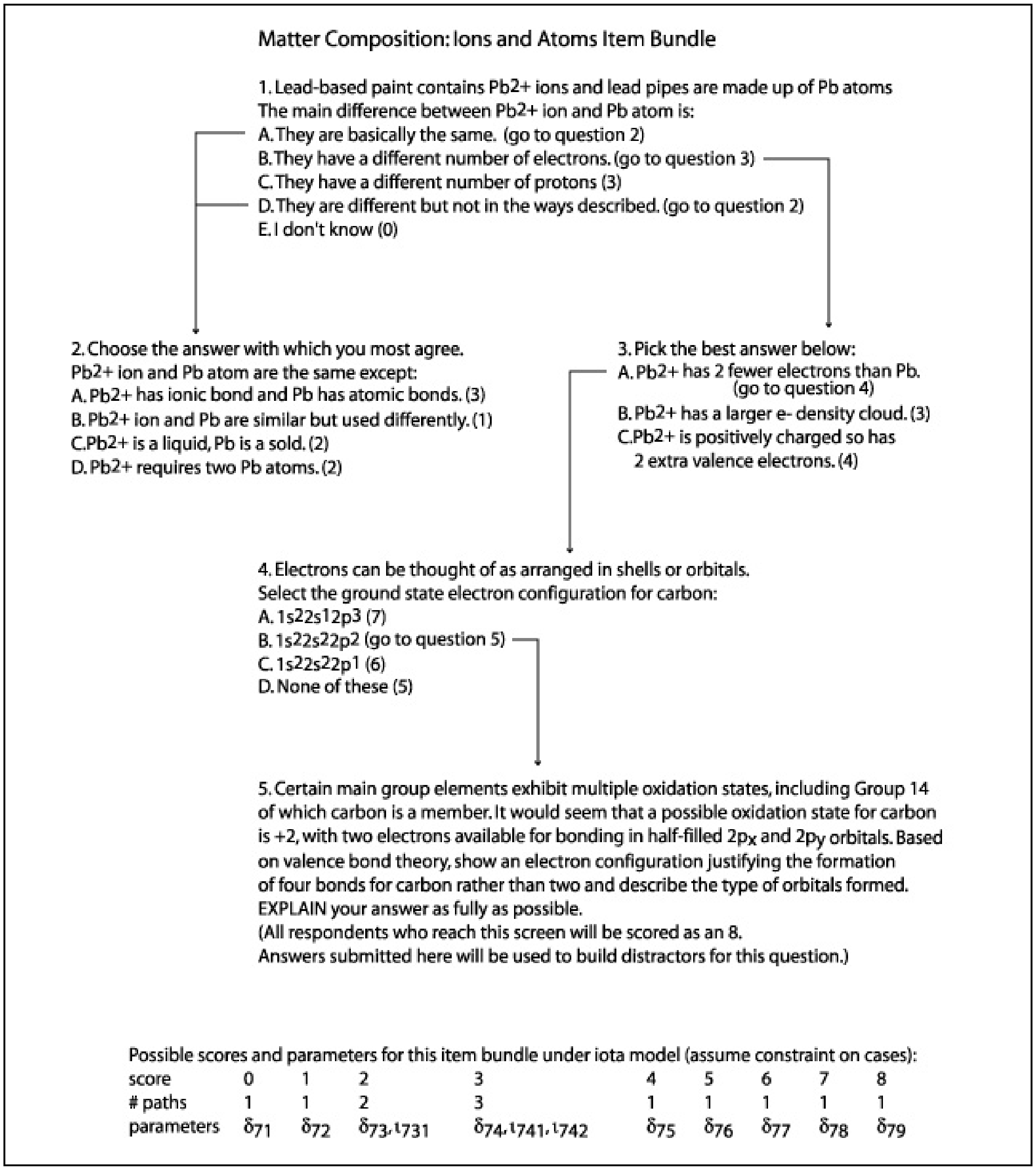
Homework sets with 15 adaptive testlets per instrument were developed from a paper-and-pencil open-ended item bank with construct modeling. Testlets of "learning facets" were developed using the large amount of data available showing actual student responses at a variety of levels. Item paneling and sensitivity reviews were also conducted. A storyboard example of a BEAR CAT testlet is shown in Figure 3. The scores to be assigned or the next item to be delivered are shown to the right of each distractor. A table showing the number of paths to each score appears at the bottom.

In order to investigate a variety of instructional design approaches within the BEAR CAT bundles, three testlet designs were invented and used across content areas. The three bundle designs differed in the target level of the opening question, the number of allowed paths to the same score, and the range of item formats employed within the testlet. Note that the testlet designs were not intended to systematically explore all possible designs, or even a complete sequence of designs, but were selected to represent what might be some useful designs for

Figure 2. Perspectives of Chemists Framework, Matter Variable

Level of Success	Big Ideas	Descriptions of Level	Item Exemplars
Generation 13-15	Bonding models are used as a foundation for the generation of new knowledge (e.g., about living systems, the environment, and materials).	Students are becoming experts as they gain proficiency in generating new understanding of complex systems through the development of new instruments and new experiments.	a) Composition: What is the composition of complex systems? (e.g., cells, composites, computer microchips) b) Structure: What gives rise to the structure of complex systems? (e.g., skin, bones, plastics, fabrics, paints, food,) c) Properties: What is the nature of the interactions in complex systems that accounts for their properties? (e.g., between drug molecules and receptor sites, in ecosytems, between device components) d) Quantities: How can we determine the composition of complex systems? (e.g., biomolecules, nanocomposites)
Construction 10-12	The composition, structure, and properties of matter are explained by varying strengths of interactions between particles (electrons, nuclei, atoms, ions, molecules) and by the motions of these particles.	Students are able to reason using normative models of chemistry, and use these models to explain and analyze the phase, composition, and properties of matter. They are using accurate and appropriate chemistry models in their explanations, and understand the assumptions used to construct the models.	a) Composition: How can we account for composition? b) Structure: How can we account for 3-D structure? (e.g., crystal structure, formation of drops,) c) Properties: How can we account for variations in the properties of matter? (e.g., boiling point, viscosity, solubility, hardness, pH, etc.) d) Amount: What assumptions do we make when we measure the amount of matter? (e.g., non-ideal gas law, average mass)
Formulation 7-9	The composition, structure, and properties, of matter are related to how electrons are distributed among atoms.	Students are developing a more coherent understanding that matter is made of particles and the arrangements of these particles relate to the properties of matter. Their definitions are accurate, but understanding is not fully developed so that student reasoning is limited to causal instead of explanatory mechanisms. In their interpretations of new situations students may over-generalize as they try to relate multiple ideas and construct formulas.	a) Composition: Why is the periodic table a roadmap for chemists? (Why is it a “periodic” table?) How can we think about the arrangements of electrons in atoms? (e.g., shells, orbitals) How do the numbers of valence electrons relate to composition? (e.g., transfer/share) b) Structure: How can simple ideas about connections between atoms (bonds) and motions of atoms be used to explain the 3-D structure of matter? (e.g., diamond is rigid, water flows, air is invisible) c) Properties: How can matter be classified according to the types of bonds? (e.g., ionic solids dissolve in water, covalent solids are hard, molecules tend to exist as liquids and gases) d) Amount: How can one quantity of matter be related to another? (e.g., mass/mole/number, ideal gas law, Beer's law)
Recognition 4-6	Matter is categorized and described by various types of subatomic particles, atoms, and molecules.	Students begin to explore the language and specific symbols used by chemists to describe matter. They relate numbers of electrons, protons, and neutrons to elements and mass, and the arrangements and motions of atoms to composition and phase. The ways of thinking about and classifying matter are limited to relating one idea to another at a simplistic level of understanding.	a) Composition: How is the periodic table used to understand atoms and elements? How can elements, compounds, and mixtures be classified by the letters and symbols used by chemists? (e.g., $CuCl_2$ (s) is a blue solid, $CuCl_2$(aq) is a clear, blue solution) b) Structure: How do the arrangements and motions of atoms differ in solids, liquids, and gases? c) Properties: How can the periodic table be used to predict properties? d) Amount: How do chemists keep track of quantities of particles? (e.g., number, mass, volume, pressure, mole)
Notions 1-3	Matter has mass and takes up space.	Students articulate their ideas about matter, and use prior experiences, observations, logical reasoning, and knowledge to provide evidence for their ideas.	a) Composition: How is matter distinct from energy, thoughts, and feelings? b) Structure: How do solids, liquids, and gases differ from one another? c) Properties: How can you use properties to classify matter? d) Amount: How can you measure the amount of matter?

Figure 3. Storyboard Showing Item Design for an Item Bundle on Ions and Atoms



Matter Composition: Ions and Atoms Item Bundle

1. Lead-based paint contains Pb^{2+} ions and lead pipes are made up of Pb atoms
The main difference between Pb^{2+} ion and Pb atom is:
A. They are basically the same. (go to question 2)
B. They have a different number of electrons. (go to question 3)
C. They have a different number of protons (3)
D. They are different but not in the ways described. (go to question 2)
E. I don't know (0)

2. Choose the answer with which you most agree.
Pb^{2+} ion and Pb atom are the same except:
A. Pb^{2+} has ionic bond and Pb has atomic bonds. (3)
B. Pb^{2+} ion and Pb are similar but used differently. (1)
C. Pb^{2+} is a liquid, Pb is a sold. (2)
D. Pb^{2+} requires two Pb atoms. (2)

3. Pick the best answer below:
A. Pb^{2+} has 2 fewer electrons than Pb.
(go to question 4)
B. Pb^{2+} has a larger e- density cloud. (3)
C. Pb^{2+} is positively charged so has
2 extra valence electrons. (4)

4. Electrons can be thought of as arranged in shells or orbitals.
Select the ground state electron configuration for carbon:
A. $1s^22s^12p^3$ (7)
B. $1s^22s^22p^2$ (go to question 5)
C. $1s^22s^22p^1$ (6)
D. None of these (5)

5. Certain main group elements exhibit multiple oxidation states, including Group 14 of which carbon is a member. It would seem that a possible oxidation state for carbon is +2, with two electrons available for bonding in half-filled $2p_x$ and $2p_y$ orbitals. Based on valence bond theory, show an electron configuration justifying the formation of four bonds for carbon rather than two and describe the type of orbitals formed.
EXPLAIN your answer as fully as possible.
(All respondents who reach this screen will be scored as an 8.
Answers submitted here will be used to build distractors for this question.)

Possible scores and parameters for this item bundle under iota model (assume constraint on cases):

score	0	1	2	3	4	5	6	7	8
# paths	1	1	2	3	1	1	1	1	1
parameters	δ_{71}	δ_{72}	δ_{73}, ι_{731}	$\delta_{74}, \iota_{741}, \iota_{742}$	δ_{75}	δ_{76}	δ_{77}	δ_{78}	δ_{79}

content developers, ranging in the complexity of item representation that could be used. The designs are:

> *Design 1: Simple Linear*. The first question in the item bundle targeted the first level in the construct framework. All items used within the bundle were standard multiple choice, with a single correct answer. Students achieving the correct answer at each level received a subsequent probe at the next level of the framework. The design is considered simple in item format (all standard multiple choice) and linear, in that levels are targeted beginning at 1 and extending up.

Design 2: Complex Split. The first question in the item bundle targeted the transition between two levels of a framework, with students "split" or filtered by the question into higher and lower levels for the next probe. Items used within the bundle drew on a variety of complex item types available through the Distributed Learning Workshop Homework Tool, including multiple answer, modify option, and open-ended, as well as multiple choice. The design is considered complex in item format and split in the filtering mechanism between levels.

Design 3: Complex Linear or Split, With Permutations. Opening questions and probes in this testlet design could target levels consecutively or split students by filtering, and included all the complex designs of Design 2, with an additional design of "permutations." Permutations allowed the decision on whether to advance a student to the next level probe to be based on not a single student answer but a series of student answers, with the pattern of answers interpreted as meaningful against the construct.

Note that although the intent of data driven content includes providing feedback and customized learning interventions attuned to the embedded assessments, in this study no feedback or interventions were included in the calibration stage, and students were instructed not to refer to outside resources in making their responses, in order to model a static rather than a possibly more dynamic, or changing, θ that might occur over the course of homework sets where active learning was taking place.

Data were collected from 521 students involved in the BEAR CAT study: 399 students from UC Berkeley's second-year Chemistry 3B course, 67 students on completion of first semester chemistry at another California university in the medical pathway, most training to become nurses and 55 students completing first semester high school chemistry at a high school in the San Francisco Bay Area.

To administer the adaptive testlets, the BEAR CAT smart homework sets were deployed through the Homework Tool capabilities of the Distributed Learning Workshop Learning Management System (Gifford, 2001), with the Homework Tool modified to accommodate the adaptive instructional flow. The modified Homework Tool was successfully alpha tested in Fall 2002 with an small pilot trial of about 70 students, and beta-tested in the full BEAR CAT study described here in Fall 2003.

Results

The testlets for the BEAR CAT sample group were first fit with a partial credit model, with data aggregated over pathways. Generalized item thresholds for the aggregated pathways (all the paths to the same score for a particular item) were compared to see if the same scores had similar difficulties across items. Criterion zones—standards for achieving each of the possible scores in the construct—were set based on the “criterion zone” rule, developed in conjunction with new approaches to standard setting, to maintain meaningful measures over time (Wilson & Draney, 2002), in which each new criterion zone is specified at the first appearance of a generalized item threshold for that zone. Thresholds are considered to "misfit" if they cross more than one adjacent criterion zone.

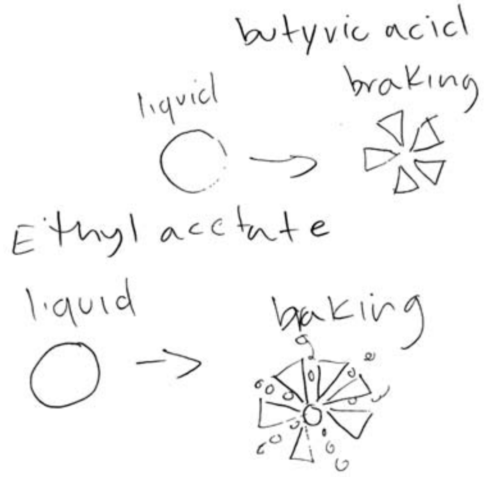
The reliability of the instrument calibrated under this model, in which the paths were not taken into consideration, was determined. Then the iota model was fit, taking the paths into consideration. The iota model is hierarchically nested within the partial credit model, and thus can be compared by a likelihood ratio chi-square test, with the difference in estimated parameters between the two models equal to the degrees of freedom. Deviance [defined as –2 × log(likelihood)] between the two hierarchical models was compared to see if the iota model, which includes the path parameters, resulted in a significantly better goodness of fit, thus shedding light on whether, in this example case, it was necessary to model the path parameters through the item bundles, or whether paths could simply be aggregated by the framework-based level and treated with a standard partial credit model.

Results from the path-aggregated partial credit analysis of the BEAR CAT testlets showed an expected a posteriori over persons variance (EAP/PV) reliabilities (Wu, Adams, Wilson, & Haldane, 2007) for the 15-bundle BEAR CAT instrument of .82. (EAP/PV reliability is explained variance according to the estimated model divided by total persons variance.) This compares to a reliability of .80 for an open-ended constructed instrument developed for the same construct; an example of an item of this type is shown in Figure 4.

Figure 4. Example of a Constructed-Response Task and Associated Scoring

You are given two liquids. One of the solutions is butyric acid with a molecular formula of $C_4H_8O_2$. The other solution is ethyl acetate with the molecular formula $C_4H_8O_2$. Both of the solutions have the same molecular formulas, but butyric acid smells bad and putrid while ethyl acetate smells good and sweet. Explain why you think these two solutions smell differently.

Notions (1-3)	1	Response: If they have the same formula, how can they be different? Analysis: Student makes one macroscopic observation by noting that the molecular formulas in the problem setup are the same.
	2	Response: I think there could be a lot of different reasons as to why the two solutions smell differently. One could be that they're different ages, and one has gone bad or is older which changed the smell. Another reason could be that one is cold and one is hot. Response: Using chemistry theories, I don't have the faintest idea, but using common knowledge I will say that the producers of the ethyl products add smell to them so that you can tell them apart. Response: Just because they have the same molecular formula doesn't mean they are the same substance. Like different races of people: black people, white people. Maybe made of the same stuff but look different. Analysis: These students use ideas about phenomena they are familiar with from their experience combined with logic/comparative skills to generate a reasonable answer, but do not employ molecular chemistry concepts.
	3	Response: "Maybe the structure is the same but when it breaks into different little pieces and changes from liquid into gas they have a different structure in the center and have a different reaction with the air. (Shows drawing:)

		 butyvic acicl liquid braking Ethyl acetate liquid baking Analysis: This answer acknowledges that chemical principles or concepts can be used to explain phenomena. Attempts are made to employ chemical concepts based on a "perceived" but incorrect understanding.

Recognition (4-6)	4	Response: "I think these two solutions smell different is because one chemical is an acid and most acids smell bad and putrid while the ethyl acetate smells good and sweet because its solution name ends with "ate" and that usually has a good sweet smell." Analysis: This response correctly cites evidence for the difference in smells between the two chemicals, appropriately using smell combinatorial patterns taught in class and chemical naming conventions, but does not explain the root cause as the difference in molecular structure between the two chemicals.
	5	Response: "They smell differently b/c even though they have the same molecular formula, they have different structural formulas with different arrangements and patterns." Response: "Butyric acid smell bad. It's an acid and even though they have the same molecular formula but they structure differently." Both responses appropriately cite the principle that molecules with the same formula can have different structures, or arrangements of atoms within the structure described by the formula. However the first answer shows no attempt and the second answer shows an incomplete attempt to use such principles to describe the simple molecules given in the problem setup.
	6	Response: (Begins with problem setup below, showing molecular formula of labeled butyric acid and same formula labeled ethyl acetate**.)** $C_4H_8O_2$ - butyric acid $C_4H_8O_2$ - ethyl acetat "The two molecules smell differently because the have different

		molecular structures. The butyric acid contains a carboxylic acid structure (which smells bad) and the ethyl acetate contains an ester (which smells good). We can tell which molecule will smell bad and which will smell good by studying the molecular structure and by looking at the names. Any 'ACID' ending name will smell bad and any '-ATE' ending name will smell good." Analysis: Response cites and appropriately uses the principle that molecules with the same formula can have different structures. Student correctly cites rule learned in class pertaining to smell patterns in relation to functional groups identified by chemical name, and uses this information to begin to explore simple molecules. However, student stops short of a Level Three response, which could be made by examining structure-property relationships through, for instance, presenting possible structural formulas for the two chemicals and explaining the bonding involved.

A standard error plot for the BEAR CAT instrument, shown in Figure 5a, indicates that the standard error for the BEAR CAT instruments is fairly flat, at about .1 logits over the student performance interval measured, and ranging from .09 to about .16 logits. The relative "flatness" of the adaptive BEAR CAT instrument as compared to the SEM plot shown in Figure 5b for the linear open-ended instrument reflects the computerized adaptive nature of BEAR CAT instruments. The adaptive approach can be viewed as a custom "series" of instruments intended to adjust the flow of items to have lower standard errors at the point of each student's individual measurement, thus as typical with CAT tending to flatten the usual u-shaped SEM plot shown by linear instruments.

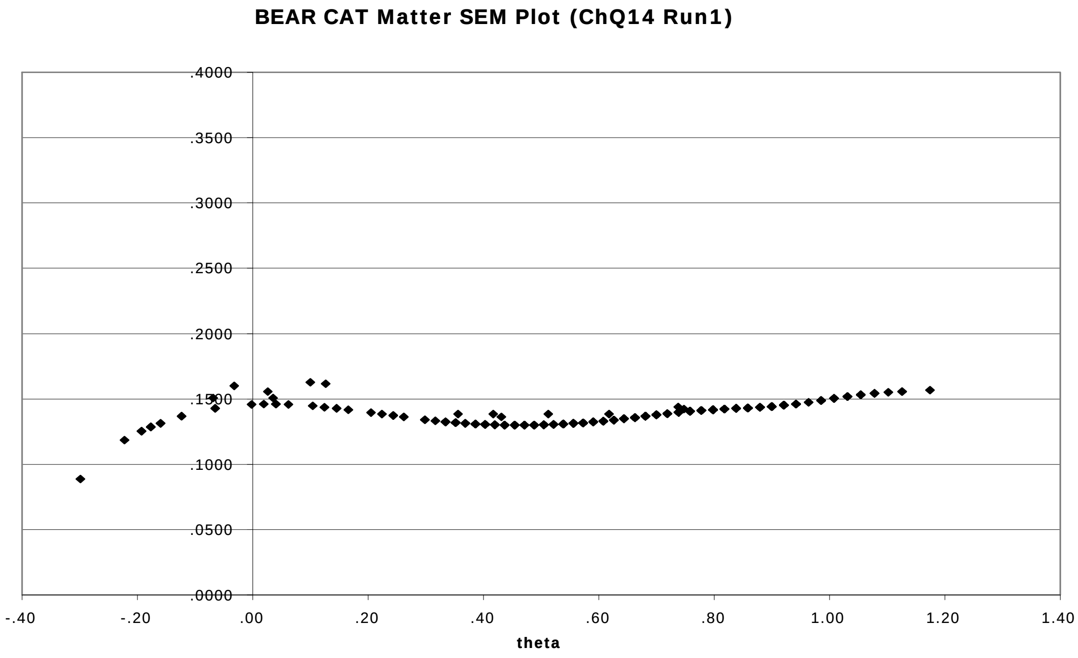
The iota model was fit to the same data as the partial credit model, but this time with the path scores not aggregated but modeled separately. The iota model is an ordered partition model, where partitions compare the equivalence of paths to the same score within testlets.

Note that a first challenge for the iota model was sparseness of data for some paths, those which did not prove to be popular paths for students. To address this, only paths with all paths-to-score including at least 5 responses in the 500-person dataset were included in this iota model calibration. About half of the total set of paths in the instrument met this "data-present" criteria. The remaining more sparse path groups were left aggregated for the iota measurement model runs, just as in the partial credit analysis model.

The iota model did show significantly better fit (deviance 18956, parameters 119) as compared to the partial credit model (deviance 19207, parameters 104), yielding a difference in deviance of 250 on 15 degrees of freedom, which is statistically significant ($p<.001$). Most of the sister paths did fall within a single criterion zone according to this model, as shown in Figure 6. Only a single path in one testlet, labeled as testlet B1, showed quite a substantially different difficulty across criterion zones from the sibling paths in its score level, while one other path pair, in testlet B4, additionally split across a criterion zone because the path groups fell close to the zone cut score.

Figure 5. Standard Error Plots Item Bundles

a. BEAR CAT



b. Open-Ended Comparison Instrument

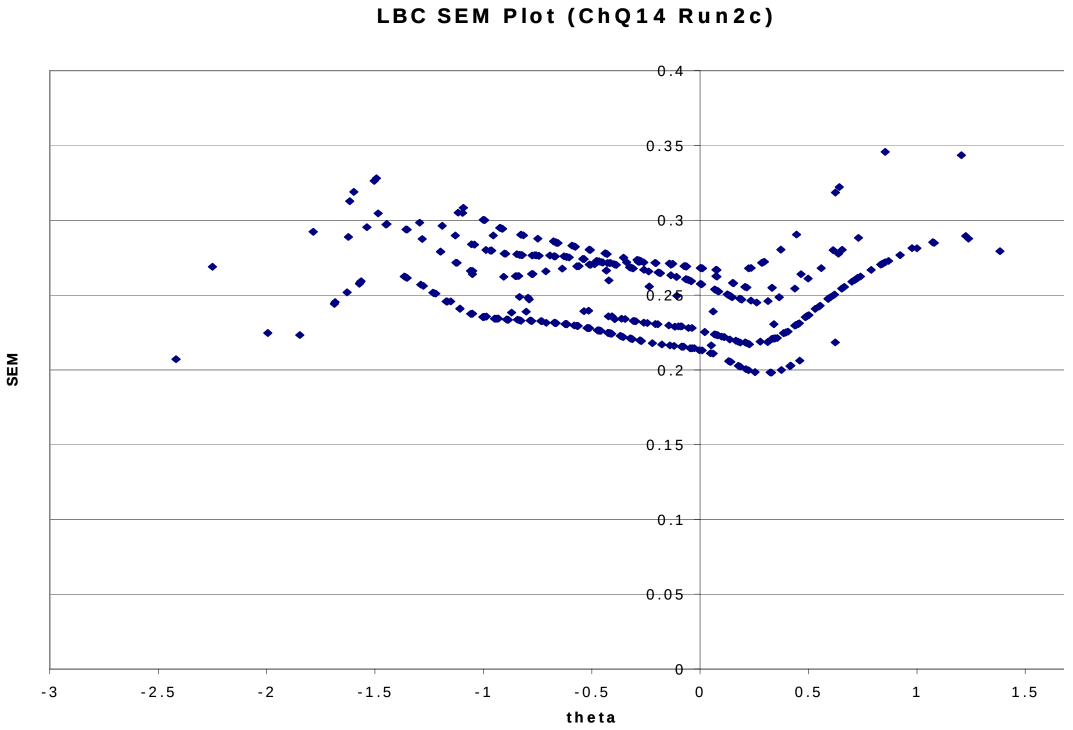


Figure 6. Locations of Sister Iota Paths in the Partially Hierarchical BEAR CAT Testlets

Students	B1	B3	B4	B6	B7	B11	B12	B14	l
Construction (8-10)									
Formulation (5-7)									
X									
X									
X									
XX									
XX									1
XXXX									
XXXX									
XXXX									
XXXXXX									
XXXXXX									
XXXXXX									
XXXXXXXX			4.3						
XXXXXXXXX		3.6							
XXXXXXXXX		3.7							
XXXXXXXXXX									
XXXXXXX									
XXXXXXXX									
XXXXXXXX			4.2						
XXXXXX									
XXXXXXXX									
XXXXXXXX									
XXXXX					7.8				
XXXXXX					7.7				
XX					7.6				
XXX					7.5		12.5 12.6		
XXX	1.3 1.4				7.4				
XX							12.4		
Recognition (2-4)									
X				6.3		11.6			
XX				6.2		11.5		14.4	
XX						11.4	12.3		
X					7.2			14.3	0
X									
X									
X									
						11.3			
Notions (1)									
X									
	1.2								

When these two iota path parameters were added to the partial credit model, reducing the degrees of freedom between the partial credit model and the iota model from 15 to 13, the deviance difference between models dropped substantially, from a delta of 250 to 142. While a delta of 142 on 13 degrees of freedom is still significant, if perceived from an effect size viewpoint the effect of taking into consideration the remaining paths is arguably small. This is because the sisters paths for the remaining path parameters all fell within the same criterion zones, or score bands of the construct, and were distinct from the other possible score levels. While one path might be slightly higher or lower in the score band, all the sister paths were shown as appropriate to receive the same score. Furthermore, even the effect size of the two modeled paths could probably be considered relatively small since these were the sparsest paths modeled in the dataset and affected the scores of only a few people (less than 15 in each case in a dataset of almost 500) on a single bundle in which the instrument consisted of 15 total bundles.

Using the partial credit model, all BEAR CAT item difficulty and step parameters fit within a standard range of .75 to 1.33 (3/4 to 4/3) mean square weighted fit (Wu, Adams, & Wilson, 1998), for parameters in which the weighted fit *T* was greater than 2, as shown in Table 2. Under the iota model, while most of the estimated parameters also fit within this tolerance, a few step parameters misfit somewhat beyond the criteria above, probably because of sparser data in these partitions, allowing outliers to have more effect.

Note that, not to be discussed fully here, a comparison of instruments was undertaken to compare the testlet results against assessment results for comparison of open-ended constructed and multiple-choice instruments. About two-thirds of students scored at the same level on a 10-point scale across the testlet and paper-and-pencil instruments, and no student differed by more than one level between instruments.

In regard to classical item discrimination across instruments, all items on the BEAR CAT and open-ended instruments showed good item discrimination, with the open-ended discrimination on average somewhat higher (mean .64, SD .10) compared to BEAR CAT (mean .53, SD .06). The multiple-choice instrument item discrimination was the lowest (mean .21, SD .15). All item discriminations were positive and non-zero except for the last items on the multiple-choice instrument, which probably were too difficult for the high school population to which they were administered and therefore non-discriminating for this group, with item discriminations near or at zero. The trend of highest discrimination for fully constructed items to lowest discrimination for the multiple-choice items reflects a trend sometimes seen for constructed and selected response, with the usual time-versus-information trade-off in the item format.

Overall, the BEAR CAT testlet instrument did achieve a higher reliability than the constructed response instrument (.82 as compared to .80) in less time, approximately 35 minutes as compared to 50 minutes for the informant group. However, rigorous comparison of BEAR CAT time across the entire sample was not possible because students completed the Smart Homework off-site, and were allowed to stay "logged-in" to their instruments during any breaks they wanted to take, making it impossible to know whether lag time was on task or on break. This was deemed appropriate for the instructional homework setting. Informant time data was based on observations of the informant group of students.

Table 2. Goodness of Fit Within Acceptable Ranges for All Item Bundles

```
================================================================================
ChQ14 Run 1 (2004 1D, Bear Cat PC, item + item*step)       Tue May 18 11:09:02
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
================================================================================
TERM 1: item
------------------------------------------------------------------------------------------
   VARIABLES                              UNWEIGHTED FIT          WEIGHTED FIT
---------------                        -----------------------   -----------------------
     item           ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
------------------------------------------------------------------------------------------
 1   PC4              -0.242   0.018    1.12 ( 0.86, 1.14)  1.7   1.09 ( 0.89, 1.11)  1.6
 2   PC7              -0.667   0.022    0.67 ( 0.87, 1.13) -5.5   0.79 ( 0.79, 1.21) -2.2
 3   PC11              0.411   0.017    1.05 ( 0.86, 1.14)  0.8   1.05 ( 0.88, 1.12)  0.9
 4   PC12              0.613   0.018    0.98 ( 0.87, 1.13) -0.3   1.03 ( 0.84, 1.16)  0.3
 5   PC16             -1.710   0.024    0.91 ( 0.87, 1.13) -1.3   0.93 ( 0.90, 1.10) -1.5
 6   PC25              0.416   0.018    1.08 ( 0.86, 1.14)  1.1   1.05 ( 0.87, 1.13)  0.8
 7   PC30a             0.106   0.018    1.19 ( 0.87, 1.13)  2.6   1.06 ( 0.88, 1.12)  0.9
 8   PC34b             0.513   0.018    1.04 ( 0.87, 1.13)  0.5   0.97 ( 0.87, 1.13) -0.4
 9   PC39ab           -0.708   0.023    0.98 ( 0.87, 1.13) -0.3   0.96 ( 0.90, 1.10) -0.7
 10  PC44             -0.116   0.018    1.05 ( 0.87, 1.13)  0.7   1.03 ( 0.90, 1.10)  0.6
 11  PC52a             0.557   0.020    0.87 ( 0.86, 1.14) -1.9   0.86 ( 0.84, 1.16) -1.8
 12  PC81             -0.043   0.019    0.99 ( 0.86, 1.14) -0.1   0.95 ( 0.86, 1.14) -0.7
 13  PC82             -0.171   0.018    1.09 ( 0.87, 1.13)  1.3   1.09 ( 0.91, 1.09)  1.8
 14  PC83              0.299   0.017    1.25 ( 0.87, 1.13)  3.4   1.21 ( 0.88, 1.12)  3.3
 15  PC84              0.741*  0.072    1.05 ( 0.87, 1.13)  0.7   1.01 ( 0.87, 1.13)  0.2
--------------------------------------------------------------------------------
*Parameter estimate was constrained.
Chi-square test of parameter equality = 11460.525,  df = 14,  Sig Level = 0.000

TERM 2: item × step
----------------------------------------------------------------------------------------
     VARIABLES                           UNWEIGHTED FIT          WEIGHTED FIT
-----------------                     -----------------------   -----------------------
     item   step   ESTIMATE  ERROR^   MNSQ       CI        T    MNSQ       CI        T
----------------------------------------------------------------------------------------
 1   PC4      0                       0.23 ( 0.86, 1.14)-17.0   1.02 ( 0.00, 2.39)  0.3
 1   PC4      1     -2.796   0.107    1.02 ( 0.86, 1.14)  0.3   1.01 ( 0.80, 1.20)  0.2
 1   PC4      2      0.512   0.104    1.15 ( 0.86, 1.14)  2.2   1.06 ( 0.83, 1.17)  0.7
 1   PC4      3      1.158   0.122    1.16 ( 0.86, 1.14)  2.3   1.04 ( 0.75, 1.25)  0.3
 1   PC4      4      2.669   0.165    0.90 ( 0.86, 1.14) -1.5   1.01 ( 0.27, 1.73)  0.1
 1   PC4      5     -0.682   0.179    1.01 ( 0.86, 1.14)  0.2   1.00 ( 0.71, 1.29)  0.1
 1   PC4      6     -0.860*           1.04 ( 0.86, 1.14)  0.5   1.04 ( 0.93, 1.07)  1.0
 2   PC7      0                       0.17 ( 0.87, 1.13)-19.8   0.98 ( 0.04, 1.96)  0.1
 2   PC7      1     -1.413   0.125    0.76 ( 0.87, 1.13) -3.8   0.92 ( 0.75, 1.25) -0.6
 2   PC7      2      2.858   0.161    0.63 ( 0.87, 1.13) -6.4   1.00 ( 0.27, 1.73)  0.1
 2   PC7      3     -0.456   0.174    0.90 ( 0.87, 1.13) -1.5   0.99 ( 0.71, 1.29) -0.0
 2   PC7      4     -0.989*           0.71 ( 0.87, 1.13) -4.7   0.81 ( 0.85, 1.15) -2.8
 3   PC11     0                       0.28 ( 0.86, 1.14)-14.7   0.97 ( 0.23, 1.77)  0.0
 3   PC11     1     -2.084   0.174    0.93 ( 0.86, 1.14) -0.9   1.02 ( 0.76, 1.24)  0.2
 3   PC11     2      1.412   0.129    1.30 ( 0.86, 1.14)  3.9   1.05 ( 0.40, 1.60)  0.3
 3   PC11     3     -0.517   0.124    0.98 ( 0.86, 1.14) -0.2   1.03 ( 0.57, 1.43)  0.2
 3   PC11     4     -1.980   0.117    1.10 ( 0.86, 1.14)  1.5   1.07 ( 0.94, 1.06)  2.0
 3   PC11     5      2.285   0.108    1.12 ( 0.86, 1.14)  1.7   1.01 ( 0.58, 1.42)  0.1
 3   PC11     6     -0.117   0.111    1.09 ( 0.86, 1.14)  1.2   1.00 ( 0.67, 1.33)  0.1
 3   PC11     7     -0.517   0.119    0.91 ( 0.86, 1.14) -1.3   0.99 ( 0.83, 1.17) -0.1
 3   PC11     8      0.889   0.172    0.84 ( 0.86, 1.14) -2.5   0.93 ( 0.76, 1.24) -0.5
 3   PC11     9      0.627*           1.09 ( 0.86, 1.14)  1.3   1.00 ( 0.73, 1.27)  0.1
 4   PC12     0                       0.65 ( 0.87, 1.13) -5.7   1.02 ( 0.21, 1.79)  0.2
 4   PC12     1     -3.539   0.218    1.05 ( 0.87, 1.13)  0.7   1.02 ( 0.93, 1.07)  0.5
 4   PC12     2      0.818   0.100    0.98 ( 0.87, 1.13) -0.2   1.00 ( 0.83, 1.17)  0.1
 4   PC12     3     -0.087   0.105    0.97 ( 0.87, 1.13) -0.5   0.99 ( 0.85, 1.15) -0.1
 4   PC12     4      3.379   0.136    0.74 ( 0.87, 1.13) -4.2   0.98 ( 0.00, 2.11)  0.2
 4   PC12     5     -2.651   0.139    0.98 ( 0.87, 1.13) -0.3   0.99 ( 0.79, 1.21) -0.1
 4   PC12     6      2.478   0.246    0.64 ( 0.87, 1.13) -5.9   0.94 ( 0.18, 1.82) -0.0
 4   PC12     7     -0.447   0.287    1.25 ( 0.87, 1.13)  3.4   0.94 ( 0.45, 1.55) -0.1
 4   PC12     8      1.498   0.583    0.33 ( 0.87, 1.13)-13.4   0.88 ( 0.00, 2.04) -0.1
 4   PC12     9     -1.451*           0.86 ( 0.87, 1.13) -2.1   0.95 ( 0.65, 1.35) -0.2
 5   PC16     0                       0.59 ( 0.87, 1.13) -7.0   1.01 ( 0.00, 2.13)  0.2
 5   PC16     1     -1.705   0.102    0.93 ( 0.87, 1.13) -1.0   0.94 ( 0.93, 1.07) -1.8
```

```
5   PC16     2      1.705*           0.92 ( 0.87, 1.13) -1.2   0.93 ( 0.94, 1.06) -2.1
6   PC25     0                       0.84 ( 0.86, 1.14) -2.5   1.08 ( 0.16, 1.84)  0.3
6   PC25     1     -2.269   0.212    1.70 ( 0.86, 1.14)  8.4   1.16 ( 0.75, 1.25)  1.3
6   PC25     2      0.978   0.144    0.64 ( 0.86, 1.14) -5.9   1.00 ( 0.51, 1.49)  0.1
6   PC25     3      0.468   0.135    1.21 ( 0.86, 1.14)  2.8   1.03 ( 0.36, 1.64)  0.2
6   PC25     4     -1.307   0.130    0.98 ( 0.86, 1.14) -0.2   1.01 ( 0.72, 1.28)  0.1
6   PC25     5     -1.104   0.115    1.05 ( 0.86, 1.14)  0.8   1.03 ( 0.92, 1.08)  0.9
6   PC25     6      1.761   0.107    1.09 ( 0.86, 1.14)  1.3   1.01 ( 0.68, 1.32)  0.1
6   PC25     7     -0.144   0.112    0.96 ( 0.86, 1.14) -0.6   0.99 ( 0.77, 1.23) -0.0
6   PC25     8     -0.262   0.128    0.85 ( 0.86, 1.14) -2.2   0.96 ( 0.87, 1.13) -0.6
6   PC25     9      1.879*           0.78 ( 0.86, 1.14) -3.3   0.95 ( 0.61, 1.39) -0.2
7   PC30a    0                       0.79 ( 0.87, 1.13) -3.3   1.01 ( 0.62, 1.38)  0.1
7   PC30a    1     -0.665   0.105    1.71 ( 0.87, 1.13)  8.6   1.12 ( 0.78, 1.22)  1.0
7   PC30a    2      0.276   0.108    1.29 ( 0.87, 1.13)  4.0   1.03 ( 0.77, 1.23)  0.3
7   PC30a    3      1.264   0.127    0.93 ( 0.87, 1.13) -1.0   1.02 ( 0.59, 1.41)  0.2
7   PC30a    4     -0.409   0.144    0.96 ( 0.87, 1.13) -0.6   1.00 ( 0.77, 1.23)  0.1
7   PC30a    5      2.327   0.357    1.06 ( 0.87, 1.13)  0.9   1.00 ( 0.33, 1.67)  0.1
7   PC30a    6     -2.794*           0.96 ( 0.87, 1.13) -0.6   0.99 ( 0.92, 1.08) -0.1
8   PC34b    0                       0.66 ( 0.87, 1.13) -5.7   1.10 ( 0.24, 1.76)  0.4
8   PC34b    1     -1.782   0.274    0.40 ( 0.87, 1.13)-11.7   0.83 ( 0.68, 1.32) -1.1
8   PC34b    2     -1.005   0.174    1.75 ( 0.87, 1.13)  9.0   1.15 ( 0.81, 1.19)  1.5
8   PC34b    3      0.626   0.129    0.85 ( 0.87, 1.13) -2.3   1.01 ( 0.67, 1.33)  0.1
8   PC34b    4      2.622   0.121    0.79 ( 0.87, 1.13) -3.4   1.02 ( 0.00, 2.39)  0.3
8   PC34b    5     -2.650   0.120    0.95 ( 0.87, 1.13) -0.7   1.01 ( 0.68, 1.32)  0.1
8   PC34b    6     -0.843   0.112    0.98 ( 0.87, 1.13) -0.3   1.00 ( 0.85, 1.15)  0.0
8   PC34b    7     -0.684   0.104    0.93 ( 0.87, 1.13) -1.0   0.99 ( 0.94, 1.06) -0.4
8   PC34b    8      3.880   0.451    0.72 ( 0.87, 1.13) -4.5   0.96 ( 0.16, 1.84)  0.0
8   PC34b    9     -0.164*           0.83 ( 0.87, 1.13) -2.6   0.94 ( 0.36, 1.64) -0.1
9   PC39ab   0                       0.38 ( 0.87, 1.13)-12.0   0.99 ( 0.21, 1.79)  0.1
9   PC39ab   1     -1.752   0.101    1.08 ( 0.87, 1.13)  1.1   1.05 ( 0.89, 1.11)  0.8
9   PC39ab   2      1.758   0.135    0.97 ( 0.87, 1.13) -0.4   0.99 ( 0.82, 1.18) -0.0
9   PC39ab   3     -0.006*           0.93 ( 0.87, 1.13) -1.0   0.94 ( 0.94, 1.06) -2.1
10  PC44     0                       0.80 ( 0.87, 1.13) -3.2   1.04 ( 0.00, 2.12)  0.3
10  PC44     1     -3.142   0.106    1.23 ( 0.87, 1.13)  3.1   1.15 ( 0.88, 1.12)  2.3
10  PC44     2      1.482   0.100    1.00 ( 0.87, 1.13)  0.1   1.02 ( 0.75, 1.25)  0.2
10  PC44     3      2.610   0.109    0.88 ( 0.87, 1.13) -1.9   1.02 ( 0.21, 1.79)  0.2
10  PC44     4     -1.424   0.111    1.04 ( 0.87, 1.13)  0.6   1.01 ( 0.78, 1.22)  0.1
10  PC44     5      0.621   0.140    1.02 ( 0.87, 1.13)  0.4   1.00 ( 0.81, 1.19)  0.0
10  PC44     6     -0.148*           0.86 ( 0.87, 1.13) -2.2   0.91 ( 0.93, 1.07) -2.5
11  PC52a    0                       0.10 ( 0.86, 1.14)-22.7   1.05 ( 0.00, 2.39)  0.3
11  PC52a    1     -2.284   0.507    0.30 ( 0.86, 1.14)-14.1   0.78 ( 0.51, 1.49) -0.9
11  PC52a    2      0.687   0.262    1.61 ( 0.86, 1.14)  7.3   1.03 ( 0.14, 1.86)  0.2
11  PC52a    3     -2.247   0.233    1.04 ( 0.86, 1.14)  0.6   1.04 ( 0.75, 1.25)  0.4
11  PC52a    4      0.174   0.146    1.03 ( 0.86, 1.14)  0.4   1.02 ( 0.70, 1.30)  0.2
11  PC52a    5     -1.308   0.124    0.98 ( 0.86, 1.14) -0.3   0.99 ( 0.92, 1.08) -0.2
11  PC52a    6      0.237   0.104    0.99 ( 0.86, 1.14) -0.1   1.00 ( 0.90, 1.10)  0.1
11  PC52a    7      0.646   0.132    0.83 ( 0.86, 1.14) -2.5   0.93 ( 0.84, 1.16) -0.9
11  PC52a    8      3.387   0.581    1.00 ( 0.86, 1.14)  0.1   0.95 ( 0.00, 2.08)  0.1
11  PC52a    9      0.708*           1.38 ( 0.86, 1.14)  4.8   0.94 ( 0.00, 2.31)  0.1
12  PC81     0                       0.38 ( 0.86, 1.14)-11.9   1.03 ( 0.23, 1.77)  0.2
12  PC81     1     -1.512   0.103    1.13 ( 0.86, 1.14)  1.9   0.99 ( 0.74, 1.26) -0.0
12  PC81     2      0.512   0.099    1.00 ( 0.86, 1.14)  0.1   1.02 ( 0.71, 1.29)  0.2
12  PC81     3      0.341   0.100    1.03 ( 0.86, 1.14)  0.5   1.02 ( 0.74, 1.26)  0.2
12  PC81     4      1.617   0.105    0.93 ( 0.86, 1.14) -1.0   1.01 ( 0.52, 1.48)  0.1
12  PC81     5     -1.395   0.109    0.98 ( 0.86, 1.14) -0.2   1.00 ( 0.91, 1.09) -0.1
12  PC81     6      0.437*           0.98 ( 0.86, 1.14) -0.3   0.98 ( 0.94, 1.06) -0.8
13  PC82     0                       0.24 ( 0.87, 1.13)-16.8   0.96 ( 0.00, 2.92)  0.3
13  PC82     1     -4.604   0.118    1.24 ( 0.87, 1.13)  3.3   1.16 ( 0.91, 1.09)  3.3
13  PC82     2      3.590   0.104    0.81 ( 0.87, 1.13) -3.0   1.02 ( 0.40, 1.60)  0.2
13  PC82     3      0.000   0.106    0.92 ( 0.87, 1.13) -1.2   1.01 ( 0.60, 1.40)  0.1
13  PC82     4      0.766   0.112    0.89 ( 0.87, 1.13) -1.6   0.99 ( 0.62, 1.38)  0.0
13  PC82     5     -0.533   0.121    1.02 ( 0.87, 1.13)  0.3   1.02 ( 0.87, 1.13)  0.4
13  PC82     6      0.781*           0.91 ( 0.87, 1.13) -1.3   0.94 ( 0.89, 1.11) -1.2
14  PC83     0                       0.83 ( 0.87, 1.13) -2.7   1.06 ( 0.31, 1.69)  0.3
14  PC83     1     -1.823   0.132    0.89 ( 0.87, 1.13) -1.7   1.08 ( 0.76, 1.24)  0.7
14  PC83     2      1.322   0.112    2.18 ( 0.87, 1.13) 13.0   1.06 ( 0.44, 1.56)  0.3
14  PC83     3     -1.280   0.109    1.25 ( 0.87, 1.13)  3.4   1.06 ( 0.76, 1.24)  0.5
14  PC83     4      0.926   0.102    1.01 ( 0.87, 1.13)  0.1   1.03 ( 0.61, 1.39)  0.2
14  PC83     5     -1.570   0.100    1.06 ( 0.87, 1.13)  0.9   1.04 ( 0.93, 1.07)  1.1
14  PC83     6      5.262   0.121    0.85 ( 0.87, 1.13) -2.2   0.98 ( 0.00, 2.94)  0.3
14  PC83     7     -3.044   0.121    1.63 ( 0.87, 1.13)  7.8   1.02 ( 0.70, 1.30)  0.2
```

```
 14  PC83     8      -0.191   0.142    1.23 ( 0.87, 1.13)  3.2   1.02 ( 0.82, 1.18)  0.3
 14  PC83     9       0.398*           2.23 ( 0.87, 1.13) 13.4   0.94 ( 0.85, 1.15) -0.8
 15  PC84     0                        0.68 ( 0.87, 1.13) -5.4   0.97 ( 0.72, 1.28) -0.1
 15  PC84     1      -1.441   0.148    1.07 ( 0.87, 1.13)  1.0   1.02 ( 0.88, 1.12)  0.4
 15  PC84     2       0.960   0.107    0.97 ( 0.87, 1.13) -0.4   1.02 ( 0.68, 1.32)  0.2
 15  PC84     3       0.396   0.104    0.98 ( 0.87, 1.13) -0.2   1.01 ( 0.55, 1.45)  0.1
 15  PC84     4      -1.987   0.103    1.02 ( 0.87, 1.13)  0.3   1.00 ( 0.90, 1.10) -0.0
 15  PC84     5       0.064   0.112    1.03 ( 0.87, 1.13)  0.4   0.99 ( 0.89, 1.11) -0.2
 15  PC84     6       3.534   0.251    2.16 ( 0.87, 1.13) 13.0   0.97 ( 0.00, 2.09)  0.1
 15  PC84     7      -1.314   0.275    0.74 ( 0.87, 1.13) -4.2   0.94 ( 0.49, 1.51) -0.1
 15  PC84     8       1.916   0.711    2.26 ( 0.87, 1.13) 13.8   0.92 ( 0.00, 2.30)  0.1
 15  PC84     9      -2.127*           2.43 ( 0.87, 1.13) 15.2   0.98 ( 0.66, 1.34) -0.1
--------------------------------------------------------------------------------
*Parameter estimate was constrained.
```

Item fit statistics and internal consistency for the three types of bundles were compared. The issue here was to see whether item fit varied systematically by the type of testlet design. Since the three item designs went from a relatively simple bundle structure to structures of increasing complexity, the probability of introducing hidden variables and thus unmodeled multidimensionality might be expected to increase from testlet design 1, the simplest, to design 3, the most complex. As reported above for the aggregate partial credit model, goodness of fit across all items and step parameters was within the tolerance described for all three testlet design structures, as shown in Table 3.

Table 3. Item Information Function Summary for the 15 BEAR CAT Bundles

Bundle	Name	Design	Peaks	Info Peak	Range	Range Unit	Shape
1	4	1	.5	4.1	-1 to 2	3	.3 at left
2	7	2	-.3	2.2	-2 to 1.5	3.5	.3 at left
3	11	1	.7	4.5	-1.5 to 3	4.5	.2 at left
4	12	2	1.2	7.8	-.5 to 2.5	3	sharper
5	16	1	-.2	.29	-3 to 3	6	no drop left
6	25	2	.3	4.9	-1.5 to 2.5	4	
7	30a	2	.4	4.8	-2 to 1.8	3.8	sharper
8	34b	3	.4	5.5	-2.3 to 3	5.3	1 at right
9	39ab	2	.1	.91	-3 to 2.5	5.5	.25 at left
10	44	3	-.5	4.3	-1 to 2	3	
11	52a	3	-.2	3.3	-2 to 3	5	1 at right
12	81	2	.3	3.7	-2.5 to 2	4.5	.2 at left
13	82	2	.6	4.5	-.7 to 2	2.7	
14	83	1	.5	5.2	-2 to 2.3	4.3	
15	84	1	1.3	5.1	-2 to 2.5	4.5	step peak left

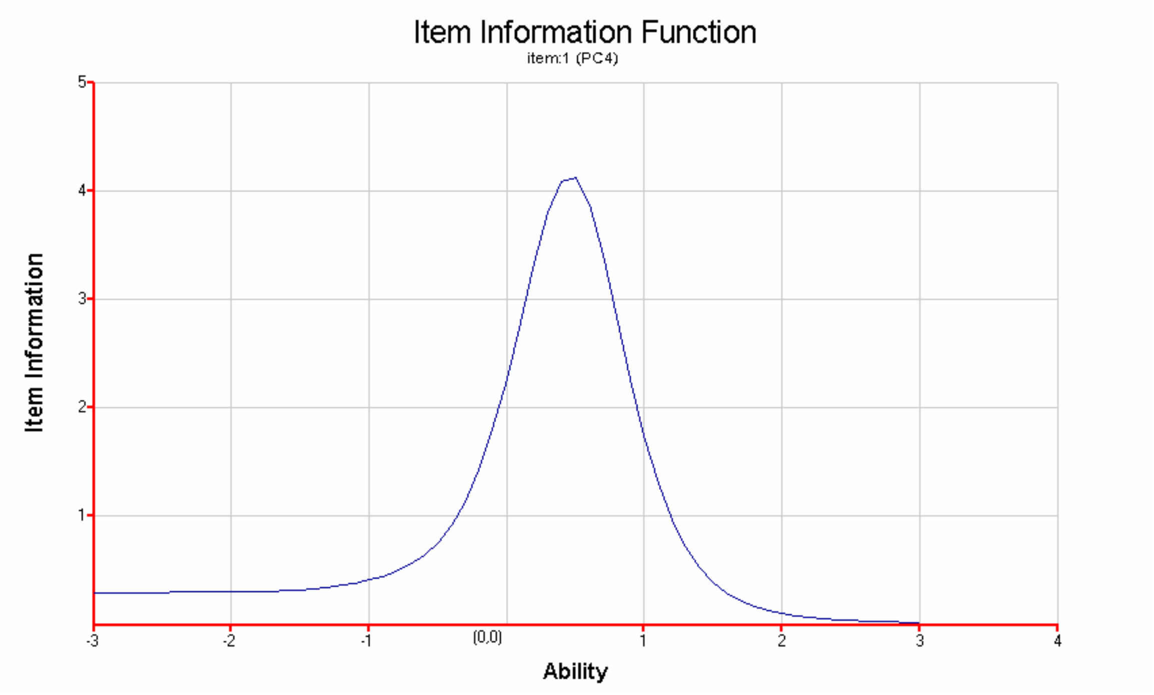
Most of the item bundles also showed good internal validity, with generalized item thresholds falling with predicted criterion zones according to the criterion zone specifications rules previously mentioned, except for nine of the 88 generalized item thresholds:

1. One level within one testlet that was determined to have inadvertently "missed" the targeted content level by including advanced relational concepts when subsequently reviewed by content experts.
2. The drift of three parameters within "new" testlets that did not have the benefit of being constructed from informant datasets from prior open-ended administrations.
3. The lowered difficulty of one testlet in the Design 3 category that was determined to offer increased opportunity for guessing because the permutations item design allowed students three opportunities to "guess" on selected response screens and assigned equal credit for guessing at all three points.

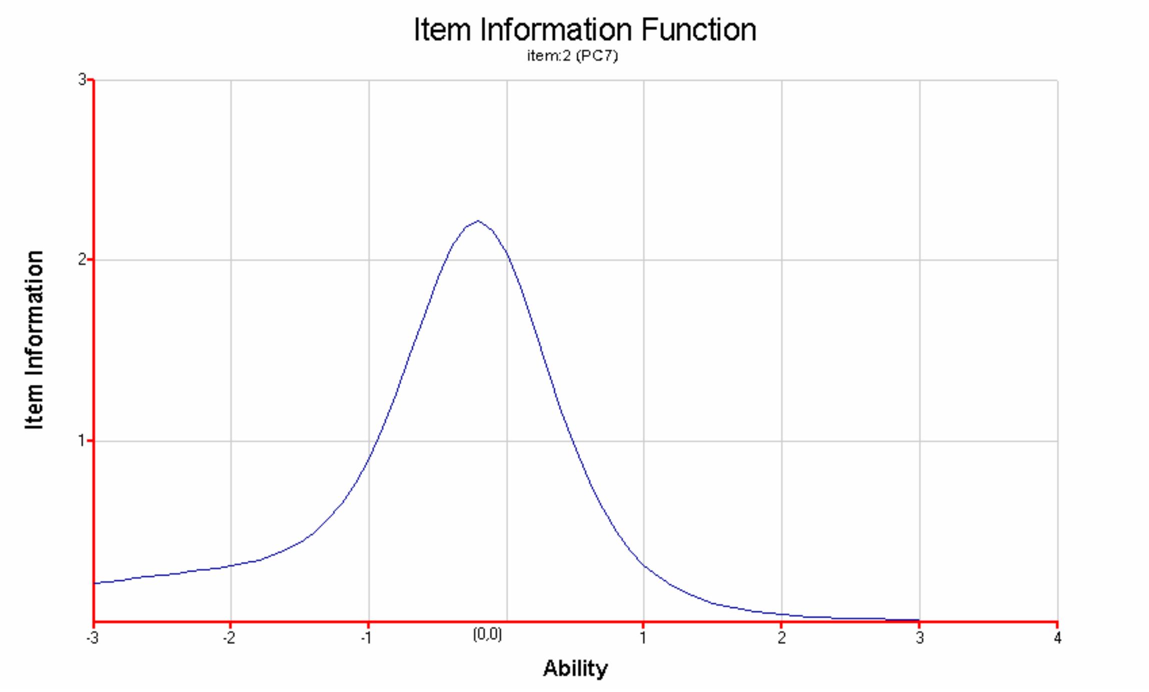
An exploration of the "informative power" of categorically different item types within the BEAR CAT instrument also was undertaken, summarized in Table 3. Based on comparisons of item information functions, examples of which are shown in Figure 7, there were no significant correlations between bundle design type and where the items information functions peaked, how much information the items offered, or the range over which they measured. Also there did not appear to be any relationship between design type and aspects of the item information function shape. Of course, bundles that included more score levels showed higher item information, but this is a previously well-known feature of partial credit items. Based on this evidence, all three types of item bundles seemed to perform reasonably similarly across the range of students measured, and no significant problems were seen except for the previously described multiple-guessing opportunities in cross-screen permutation items.

Figure 7. Examples of Item Information Functions From Items in Each Bundle Design Category

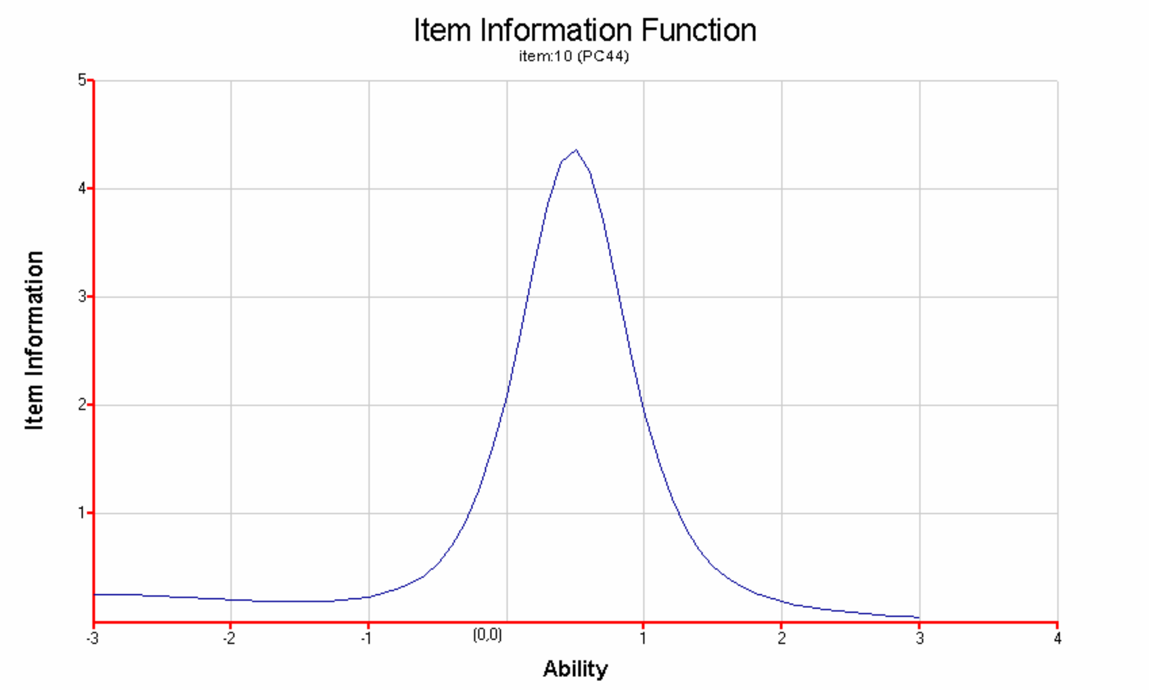
a. Bundle Design 1



b. Bundle Design 2



c. Bundle Design 3



“Within” and “Between” Testlet Adaptivity

Thus far we have discussed dynamically adapting assessment for e-learning *within* a testlet structure. Advantages here, as discussed, include controlling for local dependence with testlets, allowing a rich assembly of item formats to create complex items, identifying a procedure for accessing or equating paths to scores within an adaptive set of items, and potentially allowing feedback and learning interventions with high quality evidence within a testlet, but at a grain size appropriate to address single standards or learning objectives. We note here that adaptivity is also possible *between* testlets, or between sets of dependent items. Here, as a testlet just becomes an item score in this approach, standard CAT algorithms, pool exposure approaches, and equating techniques can apply between testlets. Combining BEAR CAT "within-testlet" adaptivity with standard CAT approaches for "between-testlet" adaptivity further increases the potential flexibility of instructional and measurement designs for e-learning products.

Summary and Conclusions

DDC suggests a means for tailoring the learning experience for students by changing or modifying materials, representations and interventions online in near real-time. Although many approaches in the past have been attempted to create adaptive streams of material for e-learning, for instance through traditional intelligent tutoring systems and other expert systems, many efforts have had limited ability to verify, compare, and test the efficacy of approaches, and to establish rigorously defensible measurement systems with verified scientific properties of validity and reliability (accuracy and precision), and some have offered limited flexibility in the complexity and contextualization of instructional design.

These results from the "Smart Homework" application suggest that adaptive testlets combined with a measurement model such as the iota model might be one basis to establish evidence-centered measurement properties in e-learning, when models fit. We found reasonably high reliability for the iota model instrument with 15 testlets—an EAP/PV reliability of .82, as compared to a slightly lower reliability of .80 for the non-adaptive paper-and-pencil comparison post-test instrument with constructed response answers. Evidence of the higher reliability showed in the somewhat flatter standard error plot for BEAR CAT as compared to the constructed response instrument, with an average BEAR CAT standard error of .1 logits. Also, students were able on average to generate this slightly more reliable score on the adaptive BEAR CAT testlet instrument in about 35 minutes in observational studies as compared to about 50 minutes as reported by teachers for the paper-and-pencil post-test instruments. The greater efficiency is not unexpected as a main advantage of CAT instruments is often considered to be time efficiency in reaching an accurate estimate of student ability.

The BEAR CAT item difficulty and step parameters showed good item fit under the partial credit model, with all fitting within a standard tolerance range for parameters in which the weighted fit *T* was greater than 2. The iota model was found to be a significantly better fit to the data than the partial credit model, with a difference in deviance of 250 with 15 degrees of freedom. It was possible to determine which paths most contributed to the effect size, with only two of 15 paths in the item set under consideration substantially contributing to the effect size.

About two-thirds of students measured at the same level on both the testlet and paper-and-pencil instruments. No students differed between the two instruments by more than one level of the ten possible levels. Instruments adapted well over a wide range of abilities, representing students drawn from first semester high school chemistry students to university-level students at the completion of three years of study in chemistry (one year in high school and two years in college).

Three different testlet designs were developed and successfully alpha and beta tested with adaptivity in the Distributed Learning Workshop *Learning Conductor* Homework Tool. All three bundle designs showed good item discrimination, with a mean of .53 (SD = .06). The intent of the three designs was to cover a reasonable range of increasing item design complexity. Most of the item bundles showed good internal validity, with no systematic problems noted except for the drift of three parameters within testlets that did not have the benefit of being constructed from informant datasets from prior administration of similar open-ended items, and the lowered difficulty of one item bundle in the Design 3 category that was determined to offer increased opportunity for guessing. No substantial difference in item information was found systematically across the three designs, except for the previously well-known result that more score categories within the testlet increased item information, provided categories were well distinguished and reasonably populated.

Though tools, content and interfaces of the testlet research project were primitive, reflecting the exploratory nature and small scale of the trial, many students reported satisfaction with such an adaptive approach to e-learning and to tailoring of materials in near-real time to student ability level, based on qualitative data results. Difficulties for the approach include sparseness of data to some adaptive paths within testlets, the challenge of identifying appropriate latent spaces for assessment with construct modeling, and the relatively large data sets needed for item response modeling of testlets.

References

Claesgens, J., Scalise, K., Draney, K., Wilson, M., & Stacy, A. (2002). *Perspectives of Chemists: A framework to promote conceptual understanding in chemistry.* Paper presented at the Validity and Value in Educational Research Session at the American Educational Research Association Annual Meeting, New Orleans, Louisiana.

Eignor, D., Stocking, M., Way, W., Steffen, M. (1993). *Case studies in computer adaptive test design through simulation: Adaptive test constraints (SAT, GRE)*. Princeton, New Jersey: Educational Testing Service.

Embretson, S. & Reise, S. P. (2000). Item response theory as model-based measurement. In *Item response theory for psychologists* (pp. 40-61). Mahwah, NJ: Lawrence Erlbaum Associates.

Gifford, B. R. (1999). *Computer-mediated instruction and learning in higher education: Bridging the gap between promise and practice.* Paper presented at the 104th Annual Meeting of the North Central Association of Colleges and Schools, Commission on Institutions of Higher Education (http://ishi.lib.berkeley.edu/sche/projects/university/april12gifford.html), Chicago, IL.

Gifford, B. R. (2001). Transformational instructional materials, settings and economics. In *The case for the Distributed Learning Workshop*. Minneapolis, MN: The Distributed Learning Workshop.

Haladyna, T. M. (1994). Writing the test item. In *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hopkins, D. (2004). *Assessment for personalised learning: The quiet revolution.* Paper presented at the Perspectives on Pupil Assessment, New Relationships: Teaching, Learning and Accountability, General Teaching Council Conference, London, England.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. Norwell, MA: Kluwer Academic Publisher.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. Van der Linden & Glas, C. A. W. (Ed.), *Computerized adaptive testing: Theory and practice* (pp. 129-148). Norwell, MA: Kluwer Academic Publisher.

Parshall, C. G., Spray, J., Kalohn, J., & Davey, T. (2002). Issues in innovative item types. In *Practical considerations in computer-based testing* (pp. 70-91). New York: Springer.

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

Russell, S. & Norvig, P. (1995). *Artificial intelligence, a modern approach*. Upper Saddle River, NJ: Prentice Hall.

Scalise, K., Bernbaum, D. J., Timms, M. J., Harrell, S. V., Burmester, K., Kennedy, C. A., et al. (2006, April 5). *Assessment for e-Learning: Case studies of an emerging field.* Paper presented at the 13th International Objective Measurement Workshop, Berkeley, CA.

Scalise, K. & Claesgens, J. (2005). *Personalization and customization in new learning technologies: Getting the right assets to the right people.* Paper presented at the Demography and Democracy in the Era of Accountability session, American Educational Research Association Conference, Montreal, Canada.

Scalise, K. & Gifford, B. R. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Teaching, Learning, and Assessment, 4*(6).

Scalise, K. & Wilson, M. (2006). Analysis and comparison of automated scoring approaches: addressing evidence-based assessment principles. In D. M. Williamson, I. J. Bejar & R. J. Mislevy (Eds.), *Automated Scoring of Complex Tasks in Computer Based Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Taylor, C. R. (2002). E-learning: The second wave.   Retrieved July 10, 2006, from http://www.learningcircuits.org/2002/oct2002/taylor.html

Timms, M. J. (2000). *Measurement issues in student modeling for intelligent tutoring systems*. Berkeley, CA: UC Berkeley POME Position Paper.

Tomlinson, C. A. & McTighe, J. (2006). *Integrating differentiated instruction + understanding by design: Connecting content and kids*. Alexandria, VA: Association for Supervision and Curriculum Development. Yes, the plus sign is correct, it is part of the title.

Trivantis. (2005). Present day custom e-learning.   Retrieved July 12, 2006, from http://www.trivantis.com/custom-elearning/custom-elearning.htm

Turker, A., Görgün, I., & Conlan, O. (2006). The challenge of content creation to facilitate personalized e-learning experiences. *International Journal on E-Learning, 5*(1), 11-17.

Wainer, H., Brown, L., Bradlow, E., Wang, X., Skorupski, W. P., Boulet, J., et al. (2006). An application of testlet response theory in the scoring of a complex certification exam. In D. M. Williamson, I. J. Bejar & R. J. Mislevy (Eds.), *Automated Scoring of Complex Tasks in Computer Based Testing*. Mahway, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*(4), 309-325.

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

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

Wilson, M. & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan & K. Kanefugi (Eds.), *Measurement and multivariate analysis (pp. 325-332). Proceedings of the international conference on measurement and multivariate analysis, Banff, Canada, May 12-14, 2000.* Tokyo, Japan: Springer-Verlag.

Wilson, M. & Scalise, K. (2003). Reporting progress to parents and others: Beyond grades. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 89-108). Arlington, VA: NSTApress.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M., Adams, R. J., & Wilson, M. (1998). The generalised Rasch model. In *User's Manual for ACER ConQuest*. Hawthorn, Australia: ACER.

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *ACER ConQuest, Version 2.0, Generalised Item Response Modelling Software*. Camberwell, Victoria: ACER Press, Australian Council for Educational Research Ltd.