JOURNAL OF APPLIED MEASUREMENT, *18*(3), 268-298

Mapping a Data Modeling and Statistical Reasoning Learning Progression using Unidimensional and Multidimensional Item Response Models

Robert Schwartz
Pearson

Elizabeth Ayers
American Institutes for Research

Mark Wilson
University of California, Berkeley

Data modeling is an approach to basic concepts of data and statistics in middle school that helps students to transform their initial, and often misguided, understandings of variability and chance to forms of reasoning that coordinate chance with variability by designing learning environments that support this reasoning by allowing students to invent and revise models. The Assessing Data Modeling and Statistical Reasoning (ADMSR) project is a collaborative effort between measurement and learning specialists that has developed a curricular and embedded assessment system based on a framework of seven constructs that describe the elements of statistical learning. Taken together, the seven constructs described above form a learning progression.

There are different ways to conceive and measure learning progressions. The approach used by the ADMSR project followed the "four building blocks" approach outlined by the Berkeley Evaluation and Assessment Research (BEAR) Center and the BEAR Assessment System. The final building block of this approach involves the application of a measurement model. This paper focuses on the application of unidimensional and multidimensional item response theory (IRT) measurement models to the data from the ADMSR project. Unidimensional IRT models are applied to aid in construct development and validation to see if the proposed theory of development presented by the construct map is supported by the results from an administration of the instrument. Multidimensional IRT measurement models are applied to examine the relationships between the seven constructs in the ADMSR learning progression. When applying the multidimensional model, specific links between levels of the constructs are analyzed across constructs after the application of a technique to align the seven dimensions.

Requests for reprints should be sent to Robert Schwartz, Pearson, 4023 Greystone Dr., Austin, TX 78731, USA, e-mail: eaa1856@gmail.com.

In order to master both statistics and informal inference, one must first understand the different concepts underlying data analysis and probability, such as the nature of chance and the idea of variability (Metz, 1998). Thus, a central aspect of any statistics curriculum in primary grade-level education will be the identification of the set of basic concepts that support data based decision making, and that can serve as a basis for more advanced statistical reasoning. *Data Modeling* (Lehrer and Romberg, 1996; Horvath and Lehrer, 1998) is an approach to learning basic concepts of data and statistics. It helps students to transform initial, and often misguided, understandings of variability and chance to forms of reasoning that coordinate chance together with variability. It accomplishes this by designing learning environments that support this kind of reasoning by guiding students to invent and revise models (Lehrer and Kim, 2009).

The different components of statistical reasoning are integrated to form the *Data Modeling* approach to learning which is represented by Figure 1. As Figure 1 illustrates, *Data Modeling* arises out of an inquiry about a well-chosen real world phenomenon. The first step of the process is the selection of certain measureable attributes that have the potential to inform the inquiry. Attributes are then defined and measured. By measuring these attributes, data is generated. This data must then be structured and represented to support the purposes of the inquiry. Statistics measure characteristics of distributed data, and models of chance support inference about these statistics in light of the inherent variability in chance events (Lehrer, Kim, Ayers and Wilson, 2014).

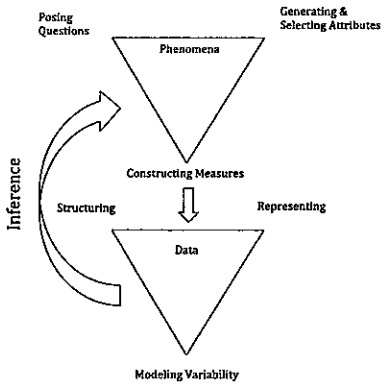

Figure 1. Data Modeling Integrates Inquiry, Data, Chance and Inference (Lehrer, Kim, Ayers and Wilson, 2014).

Assessing Data Modeling and the BEAR Assessment System

The Assessing Data Modeling and Statistical Reasoning (ADMSR) project is a collaborative effort between measurement and learning specialists to develop a curricular and embedded assessment system in the areas of statistical reasoning in a *Data Modeling* curriculum (Burmester, Zheng, Karelitz, and Wilson, 2006; Lehrer, Schauble, Wilson, Lucas, Karelitz, Kim, et al., 2007). The instruments for measuring students' ability in the *Data Modeling* domains were designed and implemented under the guidance of the Berkeley Evaluation and Assessment Research (BEAR) Center following the framework of the BEAR Assessment System (BAS; Wilson, 2005, 2009; Wilson and Sloane, 2000), which is based on the idea that good assessment addresses the need for sound measurement through four principles: (1) a developmental perspective, (2) a match between instruction and assessment, (3) the generating of quality evidence, and (4) management by instructors to allow appropriate feedback, feed forward, and follow-up. These four principles are embodied in the BAS' "four building blocks" for constructing quality assessments (Wilson, 2005):

- Construct Maps
- Items Design
- Outcome Space
- Measurement Model.

In the following paragraphs we illustrate how the Four Building Blocks have played out in the ADMSR project. The first building block, the *construct map*, is a description of a latent trait or construct and is an ordering of qualitatively different levels of performance focusing on one characteristic. A construct map is used to repre-

sent a cognitive theory of learning consistent with a developmental perspective. Figure 2 shows an example of one of the construct maps from the ADMSR project, the Conceptions of Statistics construct map.

A construct map assumes that the construct being measured can be represented by a continuum of ability punctuated by "reference points" with lower reference points of the construct at the bottom of the construct map going up toward more expert reference points at the top end of the map. These "reference points" are also often labeled as "levels." Within each of the levels, there are sub-levels (which may or may not be ordered depending on the content). The Conceptions of Statistics construct map presented here will de described in greater detail in the following paragraphs.

The ADMSR project has developed a framework of seven basic constructs that describe the elements of statistical learning. The seven constructs, or progress variables, considered in this framework were developed through a series of design experiments to explore the typical patterns of change as students learned to construct and revise models of data as a part of the *Data Modeling* curriculum. The first construct, *Theory of Measurement* (ToM), maps the degree to which students understand the mathematics of measurement and develop skills in measuring. This

Conceptions of Statistics	
CoS4 - Investigate and anticipate qualities of a sampling distribution.	
CoS4D	Predict and justify changes in a sampling distribution based on changes in properties of a sample.
CoS4C	Predict that, while the value of a statistic varies from sample-to-sample, its behavior in repeated sampling will be regular and predictable.
CoS4B	Recognize that the sample-to-sample variation in a statistic is due to chance.
CoS4A	Predict that a statistic's value will change from sample to sample.
CoS3 - Consider statistics as measures of qualities of a sample distribution.	
CoS3F	Choose/Evaluate statistic by considering qualities of one or more samples.
CoS3E	Predict the effect on a statistic of a change in the process generating the sample.
CoS3D	Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among components.
CoS3C	Generalize the use of a statistic beyond its original context of application or invention.
CoS3B	Invent a sharable (replicable) measurement process to quantify a quality of the sample.
CoS3A	Invent an idiosyncratic measurement process to quantify a quality of the sample based on tacit knowledge that others may not share.
CoS2 - Calculate statistics.	
CoS2B	Calculate statistics-indicating variability.
CoS2A	Calculate statistics indicating central tendency.
CoS1 - Describe qualities of distribution informally.	
CoS1A	Use visual qualities of the data to summarize the distribution.

Figure 2. Conceptions of Statistics (CoS) Construct Map from the ADMSR Learning Progression

construct represents the basic area of knowledge in which the rest of the constructs are played out.[1] The next construct, *Data Display* (DaD), traces a progression of learning to construct and read graphical representations of the data from an initial emphasis on cases toward reasoning based on properties of the aggregate. A closely associated construct, *Meta-Representational Competence* (MRC), proposes keystone understandings as students learn to harness representations for making claims about data and to consider trade-offs among representations in light of these claims. The fourth construct, *Conceptions of Statistics* (CoS), proposes a series of landmarks as students come to first recognize that statistics measure qualities of the distribution, such as center and spread, and then go on to develop understandings of statistics as generalizable and as subject to sample-to-sample variation. *Chance* (Cha) describes the progression of students' understanding about how chance and elementary probability operate to produce distributions of outcomes. The *Models of Variability* (MoV) construct refers to the progression of reasoning about employing chance to model a distribution of measurements. The seventh and final construct, *Informal Inference* (InI), describes a progression in the basis of students' inferences based on single or multiple samples.

The second building block of the BAS is the *items design*. In this building block, items are designed to elicit specific kinds of evidence about a respondent's ability in relation to the construct map. The prime goal of a set of items in the BAS is to generate student responses at every level of the construct map. These items can vary extensively by type, depending on the context. In the ADMSR project, the items consisted mostly of short constructed response items, but included some multiple-choice items as well. An example of the ADMSR item "Kayla's Project" is shown in Figure 3. The Kayla's Project item assesses a small part of student understanding on the Conceptions of Statistics construct. By asking students to calculate a missing value given all other values and a known mean, we are able to assess their understanding of the mean and how it is composed from component values.

After the items have been administered to the respondents, the results are interpreted using the third building block, the *outcome space*. The outcome space describes in detail how a respondent's answers to items are linked back to the different levels of the construct map. Every item in the ADMSR instruments provides evidence of a respondent's level on one or more of the seven

1 Thus, other such constructs could be, say, natural variation, leading towards topics such as evolution. ToM was chosen as an initial topic because of its transparency and accessibility for middle school students.

Kayla's Project

Kayla completes four projects for her social studies class. Each is worth 20 points.

Kayla's Projects—Points Earned
Project 1 16 points
Project 2 18 points
Project 3 15 points
Project 4 ???

The mean score Kayla received from all four projects was 17.

1. Use this information to find the number of points Kayla received on Project 4. Show your work.

Figure 3. The "Kayla's Project" Item from ADMSR

constructs. For the ADMSR project, scoring exemplars were created which explicitly scored student responses as a level on a construct map. A set of scoring exemplars for the Kayla's project item is shown in Figure 4.

The highest performing respondents to the Kayla's project item are scored at level 3 of the CoS construct. At level 3, students are able to employ more flexible strategies toward solving this problem. For example, they understand that if the mean of the four scores is 17, the scores must add to 68. Given this first step, they are able to find the missing score by subtracting the given values from 68. Students at level 2 on the CoS construct understand how to calculate the mean and use the formula as they normally would when provided a set of values. At this level, students need to use a guess and check strategy in order to solve the problem, but are able to calculate an answer. Students who gave responses judged to be relevant, but that did not provide evidence of performing at a level on the CoS construct were scored a "NL(ii)," while those who gave irrelevant responses were scored a "NL(i)." These "NL" responses are coded this way to represent responses to the item that have "no link" to the levels on the CoS construct map. Finally, respondents who saw the item but did not provide a response were scored as missing.

The final building block of the BAS is the *measurement model*. The measurement model provides a principled way to use the information about respondents and the items' responses coded in the outcome space to locate the respondents and the items on the construct map (Wilson, 2003). Different measurement models can be applied to a given instrument. Here, we apply measurement models that model the ADMSR project data all together as a single learning progression, and also individually by applying a unidimensional measurement model (the partial credit model, which is defined below) to each of the seven constructs

Level	Response Description	Example Student Responses
3D	Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among its components	1. The differences between the mean and each score are -1, 1, -2, so the last difference must be 2 and the score must be 19. 2. 16 + 18 + 15 = 49 17 X 4 = 68 68 − 49 = 19
2A	Calculate statistics indicating central tendency.	16 18 15 20 68 ÷ 4 17
NL(ii)	Student begins to carry out a strategy, but not to completion.	18 16 +15 49

Figure 4. The Scoring Exemplar for "Kayla's project"

that comprise the learning progression to see if the proposed theory of development presented by each construct map is supported by the results from the administration of the instrument to the sample of students.

A collection of construct maps taken together can comprise a learning progression (Draney, 2009; Wilson, 2009a). The seven constructs described above form a single learning progression for *Data Modeling*. A learning progression describes "successively more sophisticated ways of reasoning within a content domain that follow one another as students learn" (Smith, Wiser, Anderson, and Krajcik, 2006). Learning progressions are conjectural models of learning over time: they require empirical validation before they should be used for guiding learning and instruction. The processes of development and validation of learning progressions is accomplished through iterative cycles of empirical testing and theoretical revision and refinement (Duncan and Hmelo-Silver, 2009).

There are different ways to conceive and measure learning progressions. The BEAR Center has developed one approach to measuring learning progressions by using the assessment structure of the domain of interest. The ADMSR learning progression can be represented by the collection of the seven construct maps for the constructs described above. The ADMSR learning progression, however, hypothesizes that a student not only moves vertically up a single construct map, but can also be expected to move simultaneously across several construct maps (i.e. a student operating at a given level in one of the constructs will be operating at a specific level on one or more of the other constructs in the learning progression). The theoretical connections between the constructs are displayed in Figure 5. An arrow repre-

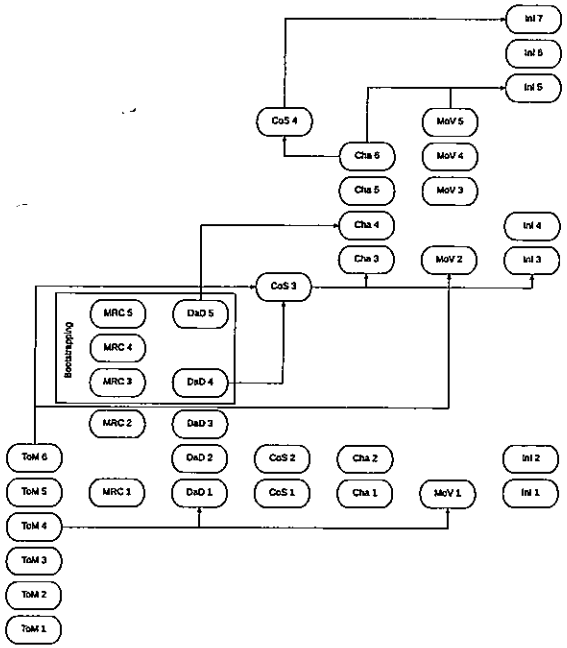

Figure 5. A Map of the Theoretical ADMSR Learning Progression

sents a specific connection between two levels of constructs – success at the level at the "point" of the arrow requires that a student has already succeeded at the level at the base of the arrow. In this paper, the relationships between the constructs are examined by modeling all seven of the constructs in the ADMSR learning progression together using a *multidimensional* measurement model. In addition, the specific links between levels of the constructs are analyzed across constructs after the application of a technique to align the seven dimensions.

The theoretical framework of the seven ADMSR constructs presented in Figure 5 imply that these constructs are closely related and should be considered both individually and as a whole. For example, the *Models of Variability* (MoV) construct indicates a progression of understanding that culminates in modeling phenomena with chance devices. As the arrows at different MoV levels in Figure 5 illustrate, this construct relies on an orchestration of the components of the other data modeling constructs. The arrows in Figure 5 represent the specific theoretical connections between two levels of different constructs – success at the level at the "point" of the arrow requires that a student has already succeeded at the level at the base of the arrow. The inclusion of these arrows in the learning progression in represents the ADMSR hypothesis that a student not only moves vertically up a single construct, but can also be expected to be moving in a coordinated way across several constructs (i.e. a student operating at a given level in one of the constructs will likely be operating at a specific level on one or more of the other constructs in the learning progression). Additionally, the shaded area labeled "Bootstrapping" between the top levels of the *Data Display* (DaD) and *Meta-Representational Competence* (MRC) constructs represents levels of the two constructs where a student's ability on one construct is increased with advancement on the other construct, and vice-versa, but where there is no theorized one-way causal connection.

Unidimensional Analysis of Individual Constructs

The ADMSR project created seven distinct constructs that compose the *Data Modeling* learning progression. Each construct is defined by a construct map, which represents a cognitive theory of learning consistent with a developmental perspective. Each latent trait is broken up into different ordered levels of performance within the construct. An instrument was created to measure the constructs, and data were collected. Following the administration of the instrument, the construct map must be re-examined in light of the data to test whether the proposed theory of development presented by the construct map is supported by the results. In addition, further investigation into the relationship between the construct map and the instrument must be undertaken in order to be facilitate the use of the student level results to evaluate progress on the latent variable. Here, two of the seven constructs, Data Display (DaD) and Conceptions of Statistics (CoS), are examined in light of the collected data.

The Data Display (DaD) Construct

The Data Display construct outlines the progression of students' perceptions of data, particularly the ways they might think about constructing or interpreting a display (e.g., a graph) as a means of better understanding the phenomenon in question. This construct describes a shift from a case-specific to an aggregate perspective of the data display. The highest level describes an integration of both perspectives. The construct map for the DaD construct is presented in Figure 6.

At level **DaD1**, students interpret displays as collections of values, but they tend not to link displays to the purposes of the display, such as the question that motivated its construction. At level **DaD2**, students interpret displays by focusing on particular cases. For example, students notice the relative value (order) of cases, their distinctiveness (e.g., outliers), or their commonalities (e.g., repeated values). Level **DaD2** is divided into

Data Display	
DaD6 – Integrate case with aggregate perspectives.	
DaD6A	Discuss how general patterns or trends are either exemplified or missing from subsets of cases.
DaD5 – Consider the data in aggregate when interpreting or creating displays.	
DaD5B	Quantify aggregate property of the display using one or more of the following: ratio, proportion or percent.
DaD5A	Recognize that a display provides information about the data as a collective.
DaD4 – Recognize or apply scale properties to the data.	
DaD4B	Recognize the effects of changing bin size on the shape of the distribution.
DaD4A	Display data in ways that use its continuous scale (when appropriate) to see holes and clumps in the data.
DaD3 – Notice or construct groups of similar values.	
DaD3A	Notice or construct groups of similar values from distinct values.
DaD2 – Focus on individual values when constructing or interpreting displays of data.	
DaD2B	Construct/interpret data by considering ordinal properties.
DaD2A	Concentrate on specific data points without relating these to any structure in the data.
DaD1 – Create displays or interpret displays without reference to goals of data creation.	
DaD1A	Create or interpret data displays without relating to the goals of the inquiry.

Figure 6. Data Display (DaD) Construct Map from the ADMSR Learning Progression

two sub-levels, **DaD2(a)** and **DaD2(b)**. At level **DaD2(a)**, students concentrate on specific data points without relating these to any structure in the data, while students at level **DaD2(b)** construct or interpret data by considering ordinal properties. These two sub-levels, like all sub-levels in the ADMSR constructs, draw a distinction between different student performances at a given level but are not ordered within that level. Thus, a student who is learning at level **DaD2(a)** is theorized to be at the same level as a student who is learning at level **DaD2(b)**.

At level **DaD3**, students begin to step toward thinking about aggregates of cases when they construct or interpret displays. Level **DaD4** marks a transition to employing a scale to thinking about aggregates of data, either by constructing displays with these characteristics (where appropriate) or by interpreting displays in light of the presence or absence of scale properties. Level **DaD4** consists of sub-levels **DaD4(a)** and **DaD4(b)**. At **DaD4(a)**, students can display data in ways that use its continuous scale to see holes and clumps in the data. At **DaD4(b)**, students begin to recognize the effects of changing bin size on the shape of the distribution. Level **DaD5** continues this shift toward the aggregate, which is assisted by quantification of aggregates. At this level, students might annotate a display to indicate the percentage of values in different classes, or they may employ statistics to quantify aggregate qualities, such as spread, and then annotate a

display accordingly. Level **DaD5** consists of two sub-levels, **DaD5(a)** and **DaD5(b)**. At **DaD5(a)**, students can recognize that a display provides information about the data as a collective. At **DaD5(b)**, students begin to quantify aggregate properties of the display by using one or more of the following: ratio, proportion or percent. Finally, at level **DaD6**, students integrate case- and density-based perspectives. They view cases as representative of regions of the data, and they begin to use aggregate data trends to evaluate individual cases. The skills shown by a student at **DaD6** are difficult to interpret in a written test, and thus no items in the current assessments tap into this level of the construct.

The Conceptions of Statistics (CoS) Construct

As its label indicates, the CoS construct describes the development of the concepts of statistics. It reflects the perspective that statistics are summary measures of data that are developed to answer research questions about distributions. It is important that students come to see the functions of statistics as ways to characterize qualities of the sample distributions (i.e., central tendency and spread) and not merely as an obligatory procedural step in working with data. Refer back to Figure 2, presented earlier, for the CoS construct map.

At level **CoS1**, students describe qualities of distribution informally by using visual qualities of data such as identifying clumps, noticing holes, or discussing the "spread" of data. At level **CoS2**, students calculate statistics, but may fail to reason about the statistic as a measure of a quality of a distribution. For example, a student may calculate the mean but neglect to relate the mean to the center of the distribution or not consider the effects of outliers on the mean. Level **CoS2** consists of two sublevels which make a distinction between calculating statistics that indicate central tendency (**CoS2(a)**) and those that indicate variability (**CoS2(b)**). At level **CoS3**, students conceive of statistics as measures of qualities of a distribution, such as its center and spread. Hence, they can reason about the effects of changes in distribution, such as the presence or absence of extreme values, on the resulting value of a statistic (**CoS3(d)**). The initial step of this level (**CoS3(a)**) starts with inventing or appropriating different ways to summarize qualities of distribution and then includes recognition that different statistics may be appropriate given particular contexts (i.e., the process generating the distribution) and forms of distribution (**CoS3(b)**). At level **CoS4**, students begin by noting and expecting sample-to-sample variability in a statistic (**CoS4(a)**) and attribute this variability to chance (**CoS4(b)**). As students investigate sampling variability, they come to understand regularities in variability that can be described by a sampling distribution. For example, students may realize that although changes in the location of the mean are expected from sample to sample, the variability of the samples' means is lower than the variability of the measurements constituting each sample (**CoS4(c)**). This culminates in predicting the effects of changes in properties of a sample on the sampling distribution (**CoS4(d)**).

Sample and Assessment Data

The ADMSR project administered a pre-test to the students prior to any of them receiving any of the *Data Modeling* curriculum, and a post-test once the lessons were completed. The students ranged in grade level from grades four through seven, and were located in Arkansas and Wisconsin public schools. Due to observations of low levels of ability from the pre-test and subsequent concerns of a floor effect that would not provide for the best possible estimates of item difficulties, the data presented here is exclusively from the post-tests, which were administered to 1002 students, who were all exposed to the *Data Modeling* curriculum. The post-test contained items that tested students' knowledge of all seven of the constructs that comprise the learning progression introduced above. A complex matrix-sampling design was used to allow for a greater number of items to be tested than each student could take in a single sitting. Using a matrix-sampling design allows gathering large amounts of data without imposing extra burden on the individual students. A total of seven different test forms were used. Each student was exposed to 20 multi-part items,

while a total of 53 multi-part items, with 110 individually scored parts, were administered.

For both the Arkansas and Wisconsin post-test scoring, multiple scorers were assigned sets of items and students to score. Depending on the number of students who saw each item, either two or three scorers were assigned to a particular item. For most items, one of the raters was an experienced scorer who was involved in the creation process of the scoring guide. For each rater-item pair, there was a minimum of 30 student scored in common. Rater analyses were performed on the data for each of the seven constructs. Differences in rating patterns were identified using Linacre's FACETS model (Linacre, 1994). This model was used to identify harsher or more lenient raters and the consistency of a rater when compared to other raters. After applying this model and analyzing the results, we concluded that there were no significant rater effects present. Upon completion of the rater analysis, in an effort to reduce parameters and simplify subsequent analyses, rater information was dropped. If two raters scored a student and their scores did not agree, then only one of the scores was chosen. If one of the scores was from a more expert rater then that score was chosen. If neither of the scorers were considered expert, then one of the scores was chosen at random. After both post-tests were scored and rater reliabilities were checked, student scores from all locations were combined for a single dataset containing 1002 students. The distribution of these items across the seven ADMSR constructs is presented in Table 1. Note that many of the items occurred in multi-part tasks.

Table 1

ADMSR Post-test Items

Construct	Multi-Part Tasks	Individually Scored Parts
Theory of Measurement (ToM)	8	14
Data Display (DaD)	12	21
Meta-Representational Competence (MRC)	7	13
Conceptions of Statistics (CoS)	10	18
Chance (Cha)	11	20
Models of Variability (MoV)	8	12
Informal Inference (Inf)	11	12

Modeling the Data using Item Response Models

To determine if each construct is being well-measured, the data is analyzed within an the Rasch framework (Rasch, 1960; Wright and Masters, 1982;) describes the relationship between the person ability and the probability of a certain response on an item. In its simplest case, it specifies a relationship between the person ability, the item difficulty, and the probability of a correct response to a dichotomous item. IRT models can also be used to handle categorical outcomes (ordinal categorical responses in this case), where the probability being modeled is that of a person responding at a certain level or higher on a polytomous item.

The IRT model that we apply to the data here is the partial credit model (PCM; Masters, 1982). The PCM was selected because the data is ordinal polytomous data, and we do not expect the differences between category step difficulties to be consistent across all items. The PCM, specified in Equation 1, models the probability of person p responding in category j of item i as a function of the person ability θ_p and step parameters δ_{ij}:

$$\Pr\left(x_{ip}=j\,|\,\theta_p\right)\frac{\exp\sum_{l=0}^{j}\left(\theta_p-\delta_{il}\right)}{\sum_{j=0}^{m_i}\exp\sum_{l=0}^{j}\left(\theta_p-\delta_{il}\right)},$$
$$j=0,1,\ldots,m_i, \qquad (1)$$

where

$$\sum_{l=0}^{0}\left(\theta_p-\delta_{il}\right)=0,\theta_p\sim N\left(0,\psi\right),$$

and m_i is the total number of steps in item i (so m_i+1 is the number of categories). The PCM can

also be specified (Equation 2) as the log ratio of the probability of person p responding in category j of item i to the probability of responding in category $j-1$ as a function of the person ability θ_p and step difficulty δ_{ij}:

$$\ln\frac{\Pr(x_{ip}=j\,|\,\theta_p)}{\Pr(x_{ip}=j-1\,|\,\theta_p)}=\theta_p-\delta_{ij},\quad j=0,1,\ldots,m_i, \tag{2}$$

where $\theta_p \sim N(0,\psi)$, and m_i is the total number of steps in item i.

A PCM was fit to the data for each of the CoS and DaD constructs individually and parameters were estimated by evaluating the marginal maximum likelihood using Gauss-Hermite quadrature with the ConQuest software (Wu, Adams, Wilson, and Haldane, 2007).

Instrument Properties: Item Fit and Reliability

After fitting the model, the first step we take is to examine whether or not items are performing in a satisfactory way. This is done by examining how well the data fit the model through the use of "fit" statistics that report how much the performance of the item differs from how we would expect it to perform in relation to the other items in the instrument. Specifically, we considered the weighted mean square fit (WMS) statistics (Wright and Masters, 1982) for the item parameter estimates. This statistic focuses attention on the question of whether the slope of the item characteristic curve is constant across the items. A WMS statistic at a value of 1.0 represents "perfect fit". Thus, values are examined as to how the statistic varies from 1.0. Mean square fit statistic values above 1.0 are indicative of situations where the item has a lower slope, and hence is behaving in a way that is less consistent with the rest of the items in the instrument than was expected. Values that are less than 1.0 indicate that an item has a higher slope than was expected, and often this is associated with local dependence issues.

As with all statistics, we pay attention to both the "effect size" and the "statistical significance" of these fit statistics. To test for statistically significant misfit, we look at the 95% confidence interval around 1.0. If a fit statistic lies outside the confidence interval, then we reject the null hypothesis that the data conforms to the model at the $p = 0.05$ level. Thus, if an item's fit statistics fall outside of the confidence interval, then the performance of the item is significantly different from what we expected based on the estimated item parameters. We will also need to look to the "effect size" of the fit statistics to determine if the misfit is large enough to deserve increased consideration. Historically, a range of 0.75 to 1.33 of the fit statistic itself is used as criterion to determine whether the items misfit (Adams and Khoo, 1993). If the fit statistic falls outside of this range, then we consider these items to be misfitting due to their effect size. We reserve our attention here for items that misfit in terms of *both* effect size and statistical significance.

In the sections below, we show example analyses form only two of the seven dimensions—we decided, on the one hand, that describing all seven, while being thorough, was more than what was needed to explain our approach. On the other hand, we felt that two examples would better portray the variations that we saw in the results.

Table 2

Data Display Weighted Fit Statistics

Item	Weighted MNSQ	95% Confidence Interval Around 1.0
GottaGo1	0.92	(0.89, 1.11)
GottaGo2	0.93	(0.91, 1.09)
GottaGo3	0.88	(0.86, 1.14)
App1	1.14*	(0.90, 1.10)
Bowl1	0.98	(0.90, 1.10)
Bowl2	1.02	(0.91, 1.09)
Crab1MC	1.02	(0.88, 1.12)
Crab2	1.17*	(0.84, 1.16)
Crab3	0.99	(0.86, 1.14)
EQ1	0.91	(0.89, 1.11)
Head2	1.03	(0.91, 1.09)
LtCherry	1.21*	(0.87, 1.13)
Max5	0.99	(0.93, 1.07)
Rocket1	0.76*	(0.83, 1.17)
Rocket2	1.14	(0.75, 1.25)
Candle1	0.93	(0.89, 1.11)
Candle2	0.98	(0.88, 1.12)
Statue4	0.98	(0.88, 1.12)
Statue6	1.01	(0.86, 1.14)
Statue8	0.93	(0.88, 1.12)
StateCap2	1.39*	(0.87, 1.13)

Note: * indicates that the value is outside the 95% confidence interval.

The weighted mean square fit statistics for the items on the DaD construct are displayed in Table 2. Five of the twenty-one items on the DaD construct had a weighted mean square fit statistic that was outside the range of the 95% confidence interval. Out of these five items that had statistically significant misfit, only one item ("StateCap2") was outside of the acceptable 0.75 to 1.33 effect size range, and thus is the only item about which we have any misfit concerns for the DaD construct. Based on the high misfit of the StateCap2 item, the item will be further scrutinized to attempt to identify some property or characteristic of the item that has caused the misfit. It was decided that the item will remain in the analysis for now. Any future administration of an instrument to assess students on the DaD construct should closely monitor the behavior of this item in future samples, and may ultimately need to eliminate the StateCap2 item if the misfit continues to occur.

The weighted mean square fit statistics for the items on the CoS construct are displayed in Table 3. Six of the eighteen items on the CoS construct had a weighted mean square fit statistic that was outside the range of the 95% confidence interval. Out of these six items that had statistically significant misfit, only one item ("FreeThrow") did not have a value between 0.75 and 1.33, and thus is the only item that requires any concern about misfit on the CoS construct. As was the case with the StateCap2 item on the DaD construct, the FreeThrow item will be further scrutinized to identify if any property or characteristic of the item has caused the misfit.

Here, the item will remain in the analysis, but in the future should be closely monitored and removed, if necessary, should the misfit continue to occur.

In addition to the item fit, we also examine the precision of the person ability estimates. For this, we look to the EAP/PV reliability coefficient. The EAP/PV reliability is the explained variance according to the estimated model divided by the total person variance (Adams, 2006), and is provided by the ConQuest software. For this data, the EAP/PV reliability is 0.64 for the CoS construct and 0.74 for the DaD constructs. These reliability estimates, however, are misleadingly low due to the design of the test forms. The different test forms given to students had items from all seven of the *Data Modeling* constructs. Due to the matrix-sampling design described above, no student responded to all the items from any construct, and few items were given to all of the students. The EAP/PV reliability estimates reported comes from an analysis of all student responses from across the forms, which includes all of the item data missing due to the test design, and thus gives an underestimate of the reliability for those constructs that one would expect in a normal administration.

Thus, in order to get meaningfully comparable reliability estimates, we estimated what the reliability would be for a five-item instrument for each construct based on simulations. The simulated data sets assumed that the entire instrument was from only one of the seven constructs. In the simulations, we used the item and item-step difficulty parameters and the distribution of

Table 3

Conceptions of Statistics Weighted Fit Statistics

Item	Weighted MNSQ	Confidence Interval Around 1.0
GottaGo1	0.92	(0.89, 1.11)
Max4MC	0.98	(0.90, 1.10)
TallestTree1	0.86	(0.75, 1.25)
TallestTree2	0.80	(0.79, 1.21)
BallMedian	0.82	(0.82, 1.18)
BallMode	0.82*	(0.83, 1.17)
BallMean	0.90	(0.87, 1.13)
Ball2	0.85*	(0.88, 1.12)
Ball3	0.81	(0.88, 1.12)
Caffeine2	1.16	(0.76, 1.24)
Corn2	1.32*	(0.86, 1.14)
Kayla1	0.90	(0.80, 1.20)
Range2	1.07	(0.76, 1.24)
Swimming1	1.30*	(0.81, 1.19)
Swimming2	0.83	(0.80, 1.20)
Swimming3	0.86	(0.81, 1.19)
Battery1	1.11	(0.80, 1.20)
Battery2	1.18*	(0.83, 1.17)
FreeThrow	1.58*	(0.83, 1.17)

Note: * indicates that the value is outside the 95% confidence interval.

person abilities from the analysis with the real data. The simulations ran with n =1000 students and assumed that there was no missing data, i.e., all students answered all 5 items for the given construct. By eliminating all of the data that was missing by the design of the test forms and by limiting the size of the test to 5 items (a realistic length for a classroom assessment), the analysis of the simulated data gives a more realistic estimate of the reliability for a more realistic context. The EAP/PV reliabilities for the DaD and CoS constructs from the sample data and the simulated data are displayed in Table 4.Note that the reliability increase for both of the constructs even with the reduction of items, due to the elimination of the missing data.

Table 4

EAP/PV Reliabilities by Construct

Construct	Original	Adjusted
Data Display (DaD)	0.74	0.80
Conceptions of Statistics (CoS)	0.64	0.83

Using a Wright Map to check the consistency of the results with the theoretical expectations—Data Display

The person ability estimates and the item difficulty estimates from the PCM analysis can be summarized graphically using a Wright Map (Wilson, 2005). By representing both the person abilities and item difficulties (and the construct map levels that they relate to) on the same scale, the results of the partial credit analysis can be related to the proposed theory of development presented by the construct maps. In this section, we first describe how to interpret a Wright Map, and the describe how we used the Wright Map results to check whether (a) the items mapped consistently to the levels they were intended to map to, and (b) whether the empirical results were consistent with the intended order of the construct levels. We use two dimensions, DaD and CoS, to illustrate some variations in our approach.

In a Wright Map, item difficulties and person proficiencies are graphed on the same scale. Lower difficulty items and lower proficiency students appear at the bottom of the scale, while higher difficulty items and higher proficiency persons are at the top. Here, in an effort to improve interpretation, the items side of the Wright Map will display the Thurstonian thresholds instead of only looking at the item difficulties and the corresponding steps. At any transition from one level of response to another on a given item, a Thurstonian threshold is the location in logits at which a person has a 50% probability of achieving a score in that category or higher (Wilson, 2005). These locations can be identified on cumulative probability plots as the points where the curves intersect with the probability equal to 0.5 line. These values tend to be more interpretable because they identify levels where students are most likely to achieve specific scores (Kennedy, 2005).

The initial Wright Map for the Data Display construct with all of the items is displayed in Figure 7. This version of a Wright Map has been ordered so that each column represents a different level of the Data Display construct. This representation helps to see which items are behaving unexpectedly compared to other items at the same hypothesized level of difficulty.

The left side of the Wright Map on Figure 7 shows the values on the logit scale, then the distribution of student proficiencies. These student proficiencies/abilities appear to be roughly normally distributed, which is one of the assumptions we made in estimating the model. Moving to the right side of Figure 7, the Wright Map displays the thresholds for each item step that have been separated into columns, which correspond with the levels of the construct map for DaD. The first of these columns is labelled "No Link" and is reserved for scores that show some sort of relevant response to the item, but are not up to the lowest level of the DaD construct (level 1). The values in the "No Link" column represent the thresholds for the boundary between the "No Link(i)" and "Missing" responses and the "No Link(ii)." Since each column represents thresholds that correspond to a given level of the construct map, if the construct map levels have a reasonably constant meaning across items, the thresholds in each column should be in a similar level of difficulty, and the difficulties should tend to increase as the levels from the construct map

increase. The Wright Map in Figure 7 suggests that most items are generally consistent in this regard, except for some noticeable overlap in the levels and some surprisingly low threshold values in level 5. Specifically, the three "Gotta Go" (GG1-3) items have threshold estimates in level 5 that are easier than expected.

Upon a closer inspection of these items, we can easily distinguish these three items from the rest of the DaD items. The "Gotta Go" items were all multiple-choice items, while the other DaD items that could be scored at the two highest levels (levels 4 and 5) were constructed response items. Students could be credited with level 5 responses to these items without having to *explain* the reasoning behind their answers. Based on this observation, we conclude that the low threshold estimates for these items are based on more than simply construct level, and instead believe that the inconsistency in ordering is primarily due to the relative easiness of choosing a response rather than explaining it. While these items might have some usefulness in assessing students' ability on the Data Display construct, being multiple-choice items they do not measure it consistently within an instrument primarily made of constructed response items. In light of these differences, these items will be removed from future assessments.

The Wright Map can also be used to classify students into the qualitatively distinct levels of understanding that were hypothesized in the construct map in Figure 6. Graphical representations of student proficiencies of this type can provide useful formative feedback to teachers for classroom planning and for diagnosing individual student needs. As an alternative to the more conventional method of convening a standard setting panel that subjectively sets cut-points

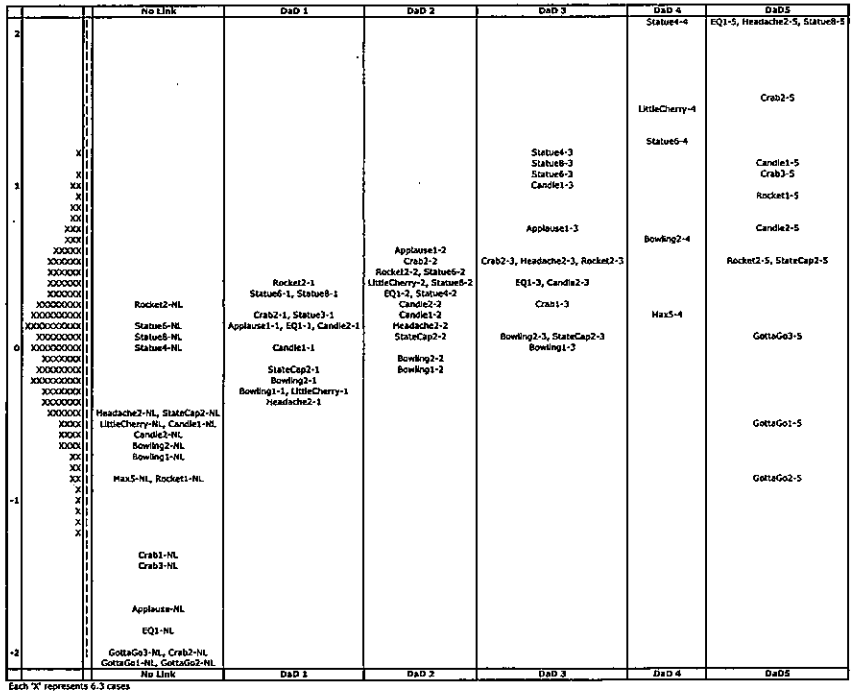

Figure 7. Data Display Initial Wright Map

along the Wright Map based on the judgments of expects (including teachers, curriculum developers, educational researchers, etc.), here we apply a quantitatively based method to set cut-points. This process of setting cut-points along the logit scale of the Wright map based on the Thurstonian Thresholds is set forth as follows by Kennedy and Wilson (2007):

1. For each level described in the construct map, compute the average Thurstonian Threshold value across items at that level.
2. Take those Thurstonian Threshold averages and find the midpoint between all of the adjacent categories.
3. Use the midpoints as the quantitative cut-point between the levels of qualitatively distinct understandings described by the construct map.

After removing the inconsistent items mentioned above, as well as any items that had a very low number of responses at a given level (which resulted in a very large standard error for that threshold estimate), cut-points were set for the Data Display construct and a new Wright Map with these levels is displayed in Figure 8.

The Wright Map in Figure 8 includes the cut-points between the construct levels, shown by the horizontal lines in the graph. The intent of these cut-points tell us what ability level a student must reach before moving to the next highest level of the construct. The items in the shaded boxes indicate items that behave within the cut-points for a given level.

As noted by Briggs and Alonzo (2009), this process of setting cut-points has some inherent potential problems, some of which are present

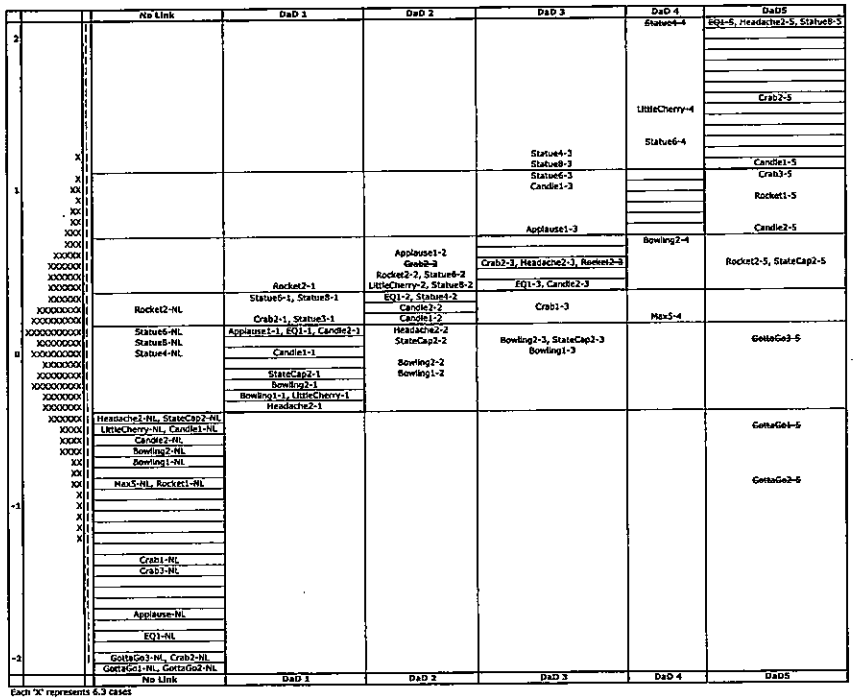

Figure 8. Data Display Wright Map with Cut-points

here. Although the mean thresholds for the items at different levels are increasing, the location of category thresholds across items is inconsistent. For example, on the Candle2 item the amount of ability necessary for a student to have a 50% probability of responding at DaD level 3 would only be enough to give that same student less than a 50% chance of scoring at the lower hypothesized DaD level 2 or higher on items Rocket2, Statue6, or Applause1. At levels 4 and 5 on the Wright Map, it appears that there is too much variability in the locations of the category thresholds. This is possibly due to the absence of students who performed near the top levels of the construct (see Figure 8). While this is of some concern, it would be premature to make conclusions on the performance of items in these top levels without more data of high performing students. In addition, more items should be included that elicit student responses on the top levels of the DaD construct, especially for level 4.

For the lower levels of the DaD construct, however, the concern is not so much that there is too much variability in the locations of category thresholds across items, but that there exist a number of overlapping items between the different construct map levels. The overlapping of items seems to occur most often between levels 1 and 2 of the construct. This suggests that these adjacent overlapping levels are not behaving distinctively from each other. In light of this, the construct map should be reconsidered to see if a level 2 response, when a student focuses on individual values in the display, is really a task that requires a higher level of ability to perform than providing a level 1 response. When taking a closer look at the lowest level 2 items, however, Bowling1 and Bowling2 are the only two DaD items where students could have scored responses at level 2(b) ("Construct / interpret data by considering ordinal properties") and not level 2(a) ("Concentrate on specific data points without relating these to any structure in the data"). This suggests that within the level 2 responses, there might exist inconsistencies. Thus, before any decision is made as to combining the entirety of level 2 with level 1, there should be an examination as to whether the lowest level 2 scores, the level 2(b) scores, require similar ability as to what is being scored in level 1.

Another issue that arises when setting cut-points is whether they are precise in classifying individual students into specific categories. It is important to keep in mind that the student ability estimates and the item category thresholds are both estimated with error. In a high-stakes testing environment, this could create great concern for misclassification of students. In a classroom environment, however, where the goal of the assessment is to provide formative feedback to the teacher, the concern is mitigated. Borderline students between 2 levels of a construct map would receive the same instruction whether or not they were classified in the higher part of the lower level or the lower part of the higher level. If the specific classification of students was a concern here, then confidence intervals could be incorporated into the Wright Map in Figure 8 using the standard errors of the threshold estimates to identify which borderline students could not be classified within a certain level of confidence.

The results presented here for the DaD construct suggest the cognitive framework theorized by the construct map is developmentally ordered as theorized, but that not all of levels may be truly distinct from each other, and that the differences in threshold estimates across items are not uniform enough to formalize cut-points.

Using a Wright Map to check the consistency of the results with the theoretical expectations—Conceptions of Statistics

As we did with the DaD construct, we now examine the Wright Map for the Conceptions of Statistics construct to examine the person ability estimates and the item threshold estimates from the PCM analysis. The Wright Map for the Conceptions of Statistics construct with all of the items is displayed in Figure 9.

When examining the CoS Wright Map, the student abilities again look to be roughly normally distributed, and for the most part, the columns of item threshold estimates for levels 1 through 3 appear to be grouped near together and to be

generally increasing in difficulty as the levels increase. Once again, however, there appears to be considerable overlap between the item threshold values across the levels. Additionally, it is difficult to make any conclusions about level 4 on this assessment, because only two items had any level 4 responses. In order to make any conclusions about how well the instrument assesses the top level of the construct map, more items must be developed that tap into level 4 of the CoS construct, and students who are at higher levels of the construct need to be included in the sample.

When taking a closer look at the level 3 items, one item appears to be easier than the other items. That item is the "Battery1" item (particularly the "Battery1-3" threshold in Figure 9), which was being piloted for the first time in this sample. Due to the low value for the level 3 threshold, we examined the item to see if we can discern why a level 3 response is easier than expected. For this item, the scoring exemplar gave students credit for a level 3(c) answer (Generalize the use of a statistic beyond its original context of application or invention) to students who answered by using a statistic to indicate typical life span. These level 3(c) responses were either point estimates such as median and mean or by estimates of an interval with a reference to the median or mean. There was no option to score a student at a level 2(a) response (Calculate statistics indicating central

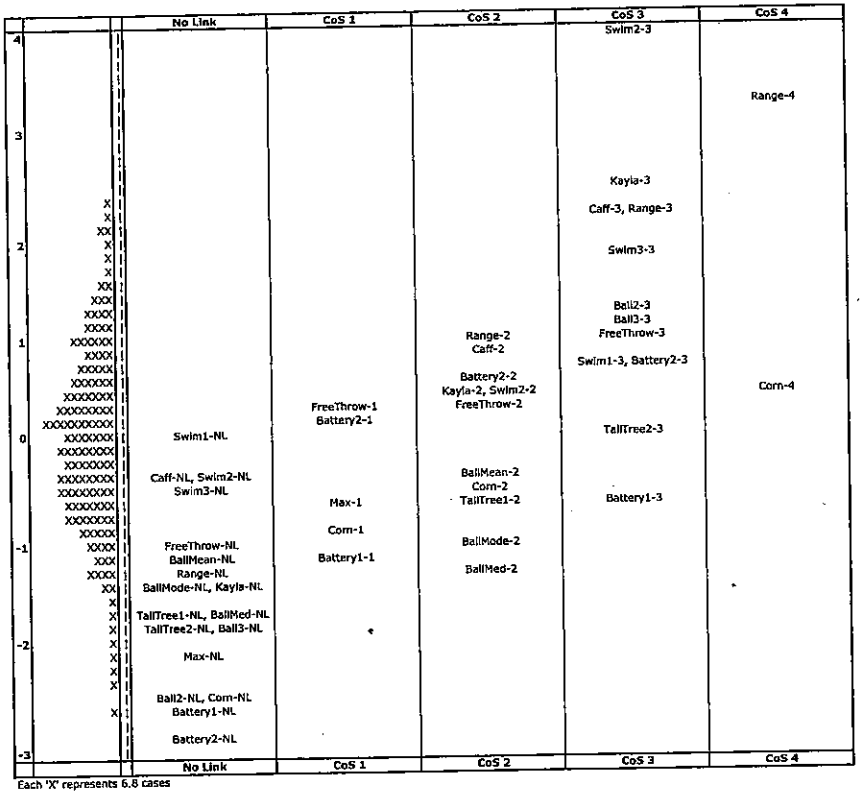

Figure 9. Conceptions of Statistics Wright Map

tendency). Thus, the item and exemplar in its current state is likely scoring some responses that should be at level 2 as level 3 responses. This would account for the low level 3 threshold estimate. Before inclusion in future assessments of the CoS construct, either this item, or the scoring exemplar, must be modified to better differentiate between a level 2(a) and level 3(c) responses.

After removing the inconsistent CoS items, as well as the items that had a very low numbers of responses at a given level, cut-points were set for the Conceptions of Statistics construct and a new Wright Map with these levels is displayed in Figure 10.

The Wright Map in Figure 10 is a modified version of Figure 9, and includes horizontal lines that represent the cut-points between the construct levels and shaded boxes that indicate items that behave within the cut-points for a given level.

Here, the mean thresholds for the items at different levels are once again increasing, and the location of category thresholds across items appears to be more consistent than the DaD construct. The one area where overlap appears to be an issue is at the lowest threshold estimates for CoS level 2. Once again, this suggests that either theses adjacent overlapping levels are not behaving distinctively from each other, or that a

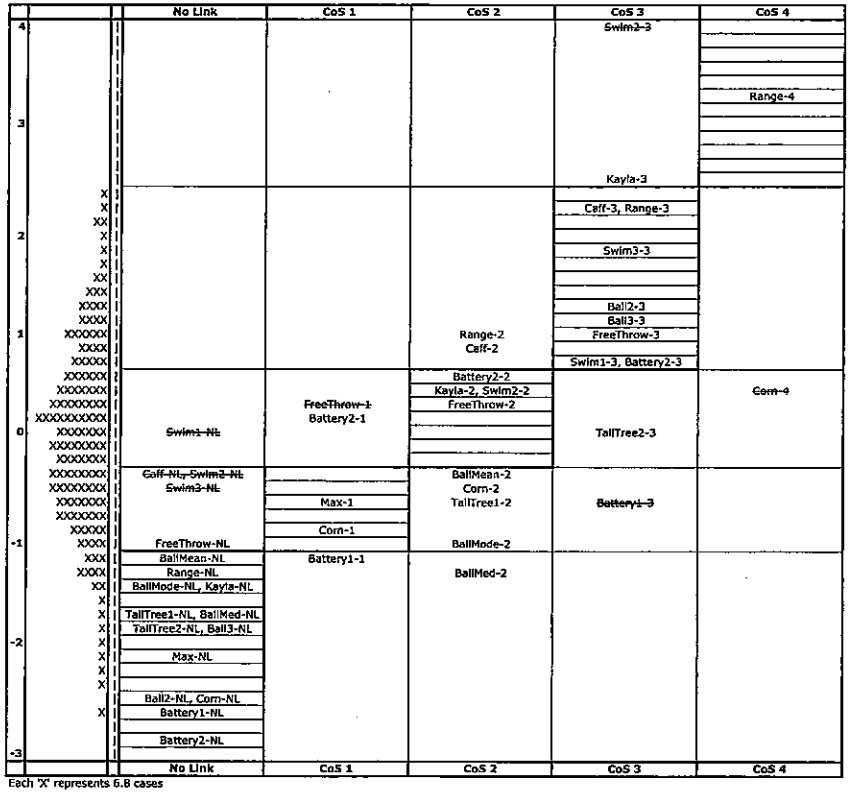

Figure 10. Conceptions of Statistics Wright Map with Cut-points

student's understanding of Conceptions of Statistics is interacting with the specific content of a given item. Since the overlap here is limited to only a few items, Battery2 on level 1 and BallMed and BallMode on level 2, we would examine these items in greater detail before making any conclusions about reconsidering the levels of the construct map.

The results presented here for the CoS construct once again support the existence of the developmentally ordered levels set forth by the construct map, although further examination must be undertaken for some overlapping items. As compared to the DaD construct, the uniformity of the threshold estimates across items provides us with more confidence in the setting of cut-points and classification of students.

The Seven ADMSR Constructs as a Learning Progression

In order to model the seven constructs together, we fit a multidimensional IRT model that models each construct as a separate but associated dimension. Multidimensional item response models describe the relationship between multiple person abilities and the probability of a certain response to an item. By modeling the seven constructs together using the Multidimensional Random Coefficients Multinomial Logit (MRCML; Adams, Wilson, and Wang, 1997; Briggs and Wilson, 2003) model, we can try to determine from a measurement perspective whether the collected data support the existence of associations between these constructs: We can test to see if the ADMSR data contained seven distinct dimensions. We test for this by seeing if the fit of the seven-dimensional model is statistically significantly better than the fit of unidimensional model of the data from all seven constructs. We also examine the correlations between the dimensions that are obtained from running the seven-dimensional MRCML model estimation. Fitting the MRCML model to the student responses should also provide the additional benefit of increased reliabilities of the seven constructs.

Furthermore, if the seven constructs can be aligned on a common scale, the relativities between the levels of the constructs, as outlined by the ADMSR theoretical framework in Figure 5, can be examined. Here, we also apply an alignment technique, Delta Dimensional Alignment, to the results from a multidimensional measurement model to examine the relationships between the seven constructs and the theorized learning progression *across* constructs. Note that we will not investigate the evidence for the specific links shown in Figure 5. The investigations of those are discussed in a separate series of papers (See Wilson, 2009b; Diakow, Irribarra, and Wilson, 2012).

The Multidimensional Random Coefficients Multinomial Logit Model

Unidimensional IRT models are based on the basic assumption that the items in the instrument measure one latent ability (Lord, 1980). Multidimensional item response models, however, are based on the assumption that more than one ability is required to respond correctly to items on a test. Generally, multidimensional IRT models have been classified as either compensatory models (Reckase, 1985, 2007) or non-compensatory models (Sympson, 1978). Compensatory models have an additive nature of the probabilities, which makes it possible for a test-taker with low ability on one dimension to compensate by having higher levels of ability on other dimensions (Ackerman, 2003). Non-compensatory multidimensional IRT models have a multiplicative nature to the probabilities, which does not allow for compensation by the other dimensions since the probability of a correct response is limited to the smallest component probability.

Compensatory multidimensional IRT models have been proposed that use the cumulative logistic function as the basis of the model (Reckase, 1985, 2007) and also the cumulative normal distribution function (McDonald, 1967). Due to the relative simplicity of mathematical calculations that can be performed using a logistic formulation, that is the method followed here.

The multidimensional IRT model we use in our analysis is a compensatory logistic model and is the multidimensional formulation of the random coefficients multinomial logit model

(RCML) (Adams, Wilson, and Wang, 1997). The RCML was designed to allow for flexibility in designing customized models and has been used for parameter estimation in the Conquest software (Wu, Adams, Wilson, and Haldane, 2007). The multidimensional formulation of the RCML is the multidimensional random coefficients multinomial logit model (MRCML; Adams, Wilson, and Wang, 1997; Briggs and Wilson, 2003).

When we fit the MRCML to the AMDSR data, we are modeling what Wang (1995) referred to as *between-item multidimensionality* (see also Adams, Wilson and Wang (1997)). Wang classified multidimensional models and tests as either having *within-item* and *between-item multidimensionality*. A between-item multidimensional test consists of items that each relate to just one dimension, while a within-item multidimensional test also includes items that relate to more than one of the I dimensions. The items in the ADMSR item bank mostly include items that relate to single dimensions, but also have some items that are designed to relate to multiple dimensions. In the latter case, however, the item responses are scored independently for each dimension and separate parameters are estimated for each of those dimensions. This independent scoring and parameter estimation for these items in effect considers them as independent items, which makes a between-item analysis reasonable.[2] Thus, even though the test is not exactly multidimensional between items, the independent scoring of an item on different dimensions allows us to treat the items as such.

Taking into account the ADMSR items (polytomous items with ordered categories), the MRCML generalization can be constrained to be a multidimensional partial credit model (see Masters, 1982). The between-items form of the multidimensional partial credit model (i.e., the between-item version) assumes that for each item i, with ordered categories of response indexed by j ($j = 0,\ldots, J$), there corresponds a unique dimension among a larger set of possible dimensions denoted by d ($d = 1,\ldots, D$). The persons responding to a given item are indexed by p ($p = 1,\ldots, P$). The log odds of the probability of a person's response in category j of item i compared to category j–1 is modeled as a linear function of a person's latent ability on that dimension (θ_{pd}), and the relative difficulty of category j (δ_{ij}, or the step difficulty):

For $j = 1,\ldots, J$,

$$\ln\frac{\Pr\left(x_{ip} = j \mid \theta_{pd}\right)}{\Pr\left(x_{ip} = j-1 \mid \theta_{pd}\right)} = \theta_{pd} - \delta_{ij}, \qquad (3)$$

When using this model, each person has a separate (though possibly correlated) latent ability estimate for each dimension d, and a vector of all of these estimates is represented by . The mean of the step difficulties (δ_{ij}) for an item i is an item's overall item difficulty δ_i. Thus, each δ_{ij} can also be formulated as $\delta_i + \tau_{ij}$, where τ_{ij} is the deviation from the mean item difficulty δ_i for item i at step j. Formulated this way, the last t parameter for each item must equal to the negative sum of the others so that the sum of all the t parameters equals zero. In the more complicated form of the multidimensional partial credit model (i.e., the within-item version) the items may each relate to more than one dimension. In the analysis here we use only the between-item version of the model.

Results—Fitting the MRCML to the ADMSR Data

After running the seven-dimensional analysis, the first test is to see if the ADMSR data contained seven distinct dimensions. We tested this for statistical significance by comparing the fit of a unidimensional model of the data from all seven constructs and the seven-dimensional model. The likelihood ratio test compares the difference in the deviances of the models with a chi-squared distribution with the degrees of freedom equal to the difference in the number of estimated parameters of the two models. The

2 A within-item multidimensional analysis was conducted on the data, and the results were compared to the results of the between-item multidimensional analysis. The two analyses yielded statistically significantly different results as a whole when examining total model fit, with the advantage to the between item model, and the parameter estimates for the affected items (those scored on more than one construct) were statistically significantly different as well. Thus, a between-item analysis is preferred to allow for the inclusion of independent scoring of an item on more than one construct.

deviances, and number of estimated parameters from the two models are given in Table 5.

As can be seen is Table 5, the difference between the deviances of these two models is 1,199.31, and the difference in degrees of freedom of 27. Applying the likelihood ratio test here, however, would not be appropriate because the null hypothesis that is being tested is that the correlation of the 7 dimensions is 1.0, which is at the boundary of the parameter space for the correlation coefficient (correlations can not be greater than 1.0). To test for significance here, we follow the suggestion by Snijders and Bosker (1999) and divide the p-value of the likelihood ratio test by two, and still get a statistically significant result. Thus, we conclude that the seven-dimensional model fits better than the unidimensional model, in a statistically significance sense.

We also need to decide whether this statistically significance difference corresponds to an important effect difference. To explore this, we looked at the estimated correlations between the constructs/dimensions that were obtained from ConQuest when running the seven-dimensional MRCML model estimation. A matrix of the correlation of the 7 dimensions is presented in Table 6.

Table 5

Relative Model Fit between Unidimensional and Seven-dimensional

	Deviance	Estimated Parameters
Unidimensional	89435.54	363
Seven-dimensional	88236.23	390
Difference	1199.31	27

Table 6

Correlations of the Seven Constructs/Dimensions

	Construct / Dimension					
	DaD	MRC	CoS	Cha	MoV	Inl
MRC	0.832					
CoS	0.785	0.872				
Cha	0.783	0.843	0.851			
MoV	0.807	0.917	0.893	0.902		
Inl	0.902	0.886	0.876	0.873	0.935	
ToM	0.778	0.811	0.818	0.806	0.847	0.811

The correlations between the constructs range from 0.778 to 0.935. Since all seven of the constructs are part of the same ADMSR curriculum and were given on a test of related material, we expected to see relatively high correlations between the dimensions but not so high as to suggest that the dimensions are the same. When the correlations are too high, however, it suggests that the constructs might not be separate dimensions after all: somewhat arbitrarily, we take 0.95 as a cut-off for dimensions being meaningfully indistinguishable. As Table 3 shows, none of the correlations between the constructs are greater than 0.95. Based on these results, which support that all seven of the constructs measure distinct dimensions, the ADMSR project continues forward with seven distinct but related constructs.

Fitting the MRCML model to the student responses also provides the benefit of increased reliabilities. The correlation structure of the model improves the reliability of the person ability estimates, as the information used for the estimates for any single dimension is augmented by the information from the estimates for the other domensions. After fitting the seven-dimensional MRCML to the data, the reliabilities for the constructs all increased, with a mean increase of almost 0.1. This increase in reliabilities leads to a practical advantage to employing multidimensional models as well, allowing the construction of shorter tests without the need to sacrifice reliability.

The person ability estimates and the item difficulty estimates from the multidimensional analysis can be summarized graphically using a multidimensional version of a Wright Map (Wilson, 2005). A seven-dimensional Wright Map for the ADMSR data is presented in Figure 11.

Figure 11 consists of seven student ability distributions in the left columns and the seven dimensions of item threshold estimates are on the right side. The thresholds in each dimension in Figure 11 are ordered by level so that each column shows all of the item thresholds for a given level on the construct map. Figure 11 shows how each dimension in the Wright Map appears to have "steps" across the dimension because the thresholds estimates are rising with each new column that represents a new level on the construct map for that dimension. While it may be convenient to view all seven ADMSR dimensions together on a single Wright Map, Figure 11 is limited because it does not allow for comparisons of the ability distributions or the item threshold estimates *across* the dimensions. This is because, in common with other multidimensional modeling approaches, the MRCML model makes the assumption that the person ability estimates are centered on zero for every dimension.[3] This is necessary for identification purposes.

Aligning the Dimensions—Delta Dimensional Alignment

If we wish to make comparisons across dimensions, then the next step in our analysis of the seven-dimensional MRCML requires aligning the dimensions. The problem with the model assumption that all the person distributions have a zero mean is that there is no *a priori* reason to assume that the students will have the same mean ability on each dimension – in fact, we hypothesize that they likely will not because we believe that some of the constructs represent more sophisticated forms of understanding. In order to

3 Equivalently, the mean of item difficulties can be set to zero for every dimension.

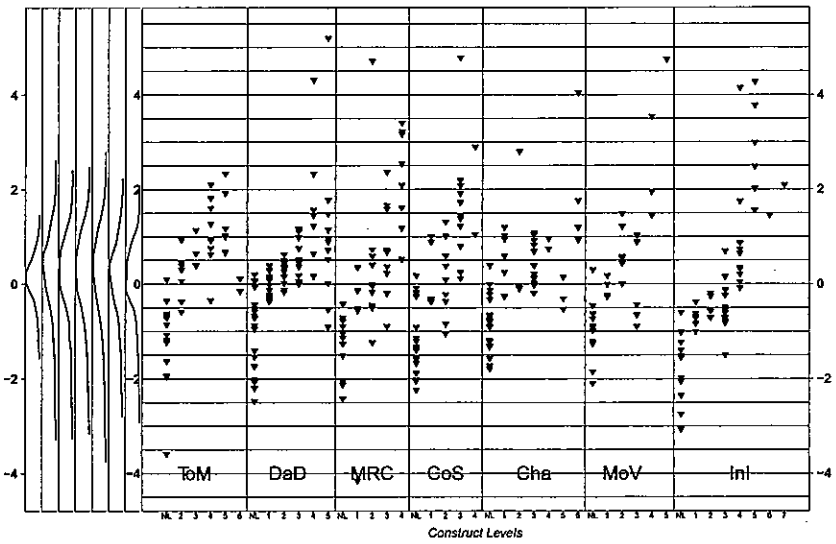

Figure 11. Seven Dimensional Wright Map of the ADMSR Learning Progression

examine whether certain levels of the construct maps are horizontally associated across the learning progression, as hypothesized in Figure 5, we need to use a technique to align the constructs/dimensions. One such alignment technique is Delta Dimensional Alignment (DDA; Schwartz and Ayers, 2011).

The Delta Dimensional Alignment method aligns multiple dimensions by transforming the item locations and step parameters obtained after running an initial multidimensional analysis. The item and step parameters of the different dimensions are transformed by using the means (m) and standard deviations (s) of the subsets of the items for each dimension (*d*), which are calculated from a unidimensional analysis (i.e., when all of the items are assumed to be from a single dimension). Thus, this technique is based on an assumption that the dimensions are somewhat (positively) correlated, and that the metric of one, composite, dimension is a reasonable one to use to align the metrics across all the dimensions. The step-by-step details of the DDA technique are described below.

The first step in DDA is to run a unidimensional analysis assuming that all items come from a single dimension to obtain item location estimates. Although we do not believe that this is exactly true, we nevertheless see it as being approximately true, as we expect all of the dimensions to be moderately to strongly correlated. Using these item location estimates, compute the mean (μ_{uni}) and standard deviation (σ_{uni}) for each subset of items by dimension. The next step is to run a multidimensional dimensional analysis to obtain another set of item location estimates. Using the item location estimates, compute the standard deviation (σ_{multi}) for each subset of items by dimension. Recall that the mean of each dimension in this second (multidimensional) analysis will be zero, due to the identification constraint. Using the estimates obtained from both analyses, transform the multidimensional item estimates using the following formulas for item location and step parameters:

Item location:

$$\delta_{id(transformed)} = \delta_{id(multi)}\left(\frac{\sigma_{d(uni)}}{\sigma_{d(multi)}}\right) + \mu_{d(uni)}, \quad (4)$$

and

Step parameters:

$$\tau_{ikd(transformed)} = \tau_{ikd(multi)}\left(\frac{\sigma_{d(uni)}}{\sigma_{d(multi)}}\right). \quad (5)$$

Note that these transformations are for the δ_i and τ_{ik} parameters values, and are not performed directly on the threshold values.

The final step of the DDA method is to run another multidimensional analysis using these transformed item estimates and step parameters as anchored values, and hence calculating new values for the Thurstonian thresholds. Since the item parameters are anchored in this final analysis, the model can estimate the student abilities without requiring the previous constraint that the person ability estimates be centered on zero for every dimension.

Examining the Links—Results from Aligning the ADMSR Data

Following the steps of the Delta Dimensional Alignment method, we ran a new MRCML analysis with the transformed and anchored item difficulty parameters. The results of this analysis can be represented with another Wright Map. Figure 12 shows the results of the aligned multidimensional Wright Map with the thresholds of all 110 items, representing all seven dimensions.

Similar to Figure 11, Figure 12 is ordered by level for each dimension and shows the rising threshold estimates across the dimension. Using the Wright Map in Figure 12, and the corresponding threshold values, we can now compare the results of the multidimensional analysis to the 13 hypothesized theoretical connections across the constructs of the learning progression that are represented in Figure 5. The evidence supporting each of the theoretical connections is indirect in that when a *source* level is below a *target* level, this is consistent with the link, but does not prove the link. On the other hand, if the opposite is true, i.e. a source level is above a target level, then this would be evidence against the existence of the link.

In the next few paragraphs we examine a particular subset of the links, as an example of the process we use, and then summarize the entire set of results across all of the dimensions.

Starting from the left hand side of Figure 5, the first two theoretical connections come out of the *Theory of Measurement* (ToM) construct at level 4. The ToM construct maps the degree to which students understand the mathematics of measurement and develop skills in measuring. At level 4 of ToM, a student is beginning to consider properties of a unit in relation to the goals of measurement. Within this level, a student starts to use standard units, consider the suitability of a certain unit, qualitatively predict inverse relation between size of unit and measure, and partition units by factors of 2 when an object cannot be measured in whole units. As Figure 5 displays, the connections from ToM level 4 go to the first levels of both the *Data Display* (DaD) and the MoV constructs. At level 1 of DaD, a student is beginning to create or interpret displays without reference to the goals of the inquiry. At level 1 of MoV, a student will be able to identify sources of variability.

To analyze the connection between level 4 of ToM and level 1 of DaD and MoV, we look at the mean and range of the threshold estimates for those levels. The ToM4 threshold estimates have a mean of 0.49 and range from –0.90 to 1.49 logits. The DaD1 threshold estimates have a mean of 0.11 and range from –0.29 to 0.49 logits, and the MoV1 estimates have a mean of 0.27 and a range from 0.12 to 0.48 logits. Comparing these estimates, it appears that the DaD1 and MoV1 items are not as difficult as many of the ToM4 items, and the data might not provide evidence to support the connection hypothesized in the learning progression that students proceed from ToM4 to DaD1 and from ToM4 to MoV1. However, looking more closely at the ToM construct, however, we see that ToM4 has three threshold estimates that are inconsistently higher than the other estimates for that level, and that the threshold estimates all are

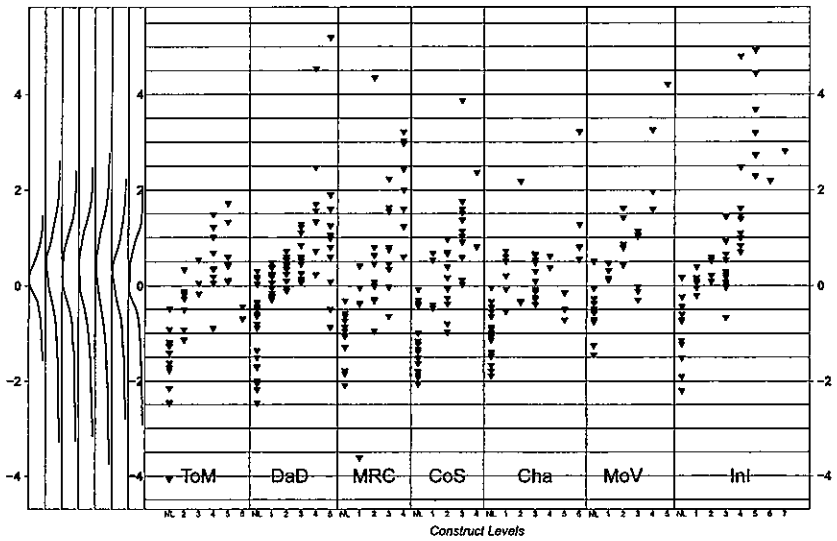

Figure 12. ADMSR Learning Progression Using DDA

for items that come from the same multi-part task called "Ruler." After flagging the Ruler task as requiring further review and removing the Ruler threshold estimates, ToM4 would only have a mean value of 0.12 logits. If we compare DaD1 and MoV1 to ToM4 after we have removed the Ruler items, then it would appear that the item thresholds of DaD1 and ToM4 are fairly comparable, while the threshold estimates for MoV1 are slightly higher. A close-up of the section of the Wright map containing the ToM, DaD and MoV constructs is presented in Figure 13.

Figure 13 shows the item threshold estimates for the three constructs being examined, and it includes red boxes that indicate the estimates that are part of the theorized link. With the removal of the Ruler items, these results provide evidence to support a progression of students from ToM4 to MoV1, and that students are performing at level 1 of DaD at about the same time as they are performing at level 4 of ToM.

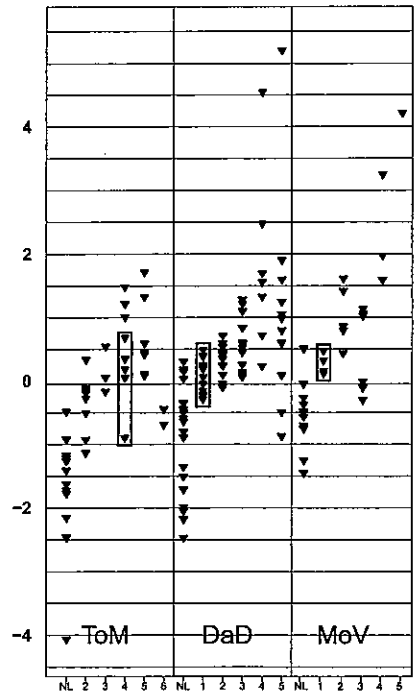

Figure 13. Link Between ToM4 with DaD1 and MoV1

Moving up the ToM construct to level 6, there are two theoretical connections that connect ToM6 to level 3 of the *Conceptions of Statistics* (CoS) construct and to level 2 of MoV. The arrows from ToM6 to CoS3 and MoV2 theorize that a student would need to progress to level 6 of ToM before he could attain the respective levels of the other constructs. At level 6 of ToM, students predict the effects of changes in unit on measure or scale. Students use relations among units to quantify the effects of a change in unit on the resulting measure and evaluate tradeoffs when selecting measurement tools. This is the top level of the ToM construct. Students at CoS3 conceive of statistics as measures of qualities of a distribution, such as its center and spread. Hence, they can reason about the effects of changes in distribution, such as the presence or absence of extreme values, on the resulting value of a statistic. At MoV2, students informally order the contributions of different sources to variability, using language such as "a lot" or "a little." Students also describe mechanisms and/or processes that account for these distinctions, and they predict or account for the effects on variability of changes in these mechanisms or processes.

To analyze the connection between ToM6 and CoS3, we look to the mean and range of the threshold estimates for those levels. The ToM6 threshold estimates have a mean of –0.57 and range from –0.69 to –0.44 logits. The CoS3 threshold estimates have a mean of 1.31 and range from 0.04 to 3.87 logits. Comparing these estimates, it is clear that the ToM6 estimates are lower than the CoS3 estimates. Based on this comparison alone, this at least does not contradict the hypothesized connection in the learning progression that students reach level 6 of the ToM construct before reaching level 3 of the CoS construct. The ToM construct, however, only has two threshold estimates at level 6. Therefore, we also want to compare ToM4 and ToM5 to CoS3 to have more confidence in our comparison. Figure 14 contains a close-up of these constructs on the

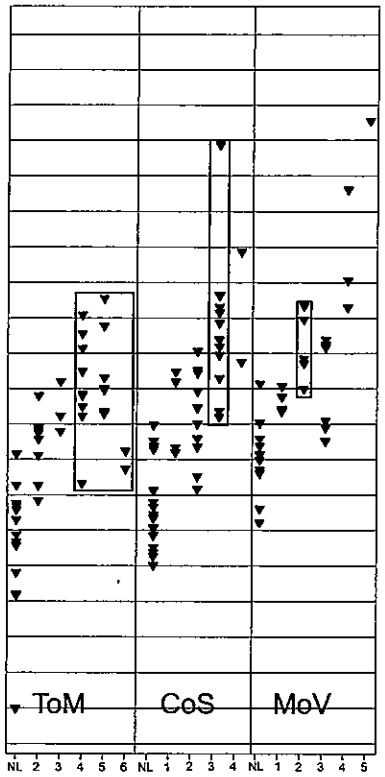

Figure 14. Link Between ToM6 with CoS3 and MoV2

Wright map, with the levels of the theoretical link highlighted.

By examining Figure 14 and the means of the threshold estimates for ToM4 (0.15) and ToM5 (0.49), we find that ToM levels 4 and 5 are both more difficult than ToM6, but are still clearly lower than the CoS3 estimates. This supports the hypothesized connection in the learning progression that students reach the top levels of the ToM construct before reaching level 3 of the CoS construct. Thus, the ability to predict the effects of changes in unit on measure or scale (ToM6) precedes the ability to conceive statistics as measures of qualities of a distribution (CoS3).

To analyze the connection between ToM6 with MoV2, we again compare the means and ranges of the threshold estimates for those levels. The MoV2 threshold estimates have a mean of 1.0 and range from 0.44 to 1.62 logits, and are clearly higher than the ToM4–6 thresholds. This supports the hypothesized connection in the learning progression that students reach the top levels of the ToM construct before reaching level 2 of the MoV construct. Thus, as theorized, it was easier for the students in the sample to predict the effects of changes in unit on measure or scale (ToM6) than to informally order the contributions of different sources to variability (MoV2).

As mentioned above, the shaded area labeled "Bootstrapping" in Figure 5 between MRC3–5 and DaD4–5 represents the aspect of ADMSR theory of learning that these levels of the two constructs are where a student's ability on one construct increases with a coordinated ability increase on the opposing construct. This theory behind this connection is not yet as clear as those designated by the arrows, and this interaction is still being explored to determine how to further model this relationship. The estimates for the bootstrapping levels are fairly similar, with the DaD4–5 estimates have a mean of 1.47 logits and the MRC3–5 threshold estimates have a mean of 1.43 logits. A close-up of the section of the Wright Map featuring the item threshold estimates for DaD and MRC is displayed in Figure 15.

Figure 15 shows the item threshold estimates for DaD and MRC, and it includes two red boxes that indicate the estimates that are a part of the bootstrapping section of Figure 5. Note that these boxes do not include all of the thresholds—two outlier threshold estimates for the DaD construct have been excluded. The distributions of the different levels of the two constructs appear to be at similar values with a fair amount of overlap, with MRC4 appearing to be slightly higher than the other levels. Unfortunately, this sample contained no MRC5 scores. Only two items had responses that could have been scored at this level, but none of the students produced responses that warranted a MRC5 score. Even though the connection betweens the DaD4–5 and MRC3–4 levels

are supported by the data here, future samples of higher scoring students are needed in order to make any conclusions regarding the difficulty of MRC5 items and how they are related to the DaD construct and the other levels of MRC.

Another of the connections represented in Figure 5 starts at level 6 of Cha and goes to level 4 of CoS. At Cha6, students develop probabilities for compound (aggregate) events as a ratio of target outcome(s) and the total number of outcomes. The level culminates in coordinating relative frequencies of observed outcomes for aggregate events with the probabilities of these outcomes. Students at CoS4 start expecting sample-to-sample variability in a statistic, and attribute this variability to chance. The arrow from Cha6 and CoS4 theorizes that a student would need to progress to level 6 of Cha before he could attain level 4 of CoS.

To analyze the connection between Cha6 and CoS4, we once again look at the mean and range of the threshold estimates for those levels. After eliminating one outlier, the Cha6 threshold estimates have a mean of 0.86 and range from

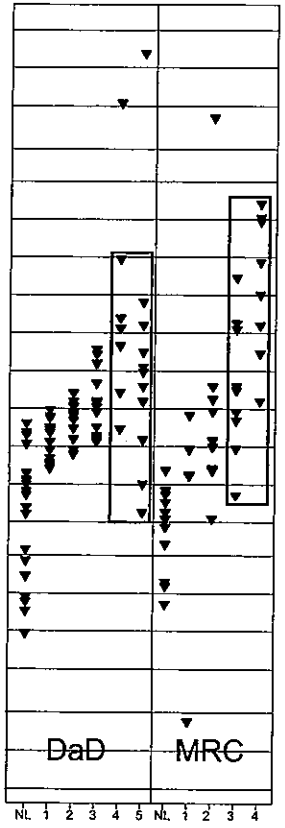

Figure 15. Bootstrapping Between DaD and MRC

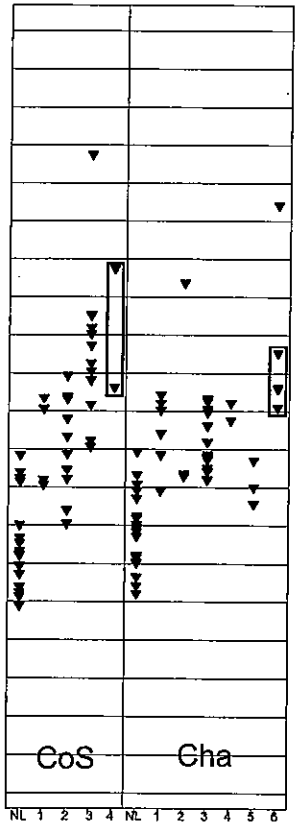

Figure 16. Link Between Cha6 with CoS4

0.56 to 1.28 logits. The CoS4 threshold estimates have a mean of 1.59 and range from 0.82 to 2.38 logits. A close-up of the section of the Wright map containing the Cha and CoS constructs is presented in Figure 16.

While there exists a small amount of overlap between the threshold estimates, looking at Figure 16 and comparing the means and ranges of the thresholds show that the Cha6 threshold estimates are lower as a whole than the CoS4 estimates. This supports the hypothesized connection in the learning progression that students reach the top level of the Cha construct before reaching level 4 of the CoS construct. As mentioned previously, however, there are only two CoS4 threshold estimates being compared in this link. So, even though the data supports the connection, there is not enough data to make a conclusion until more CoS4 items are examined.

Through similar analyses, we can examine the remaining theorized connections of the ADMSR learning progression represented in Figure 5. The results discussed above and these remaining results are summarized in Table 7. These results provide evidence either supporting or rejecting the theorized relationships between the constructs. At this time, the ADMSR project is not removing any of the connections that are not supported by the data until they are examined with another sample. Regardless of whether the

Table 7

Summary of Results for Theorized Connections in ADMSR Learning Progression

Hypothesized Link	Observation	Conclusion
ToM4 to DaD1	Partially supported by data. When Ruler items are removed, data supports that ToM4 has similar difficulty as DaD1.	Examine Ruler items. Retain theorized connection
ToM4 to MoV1	Supported by data. When Ruler items are removed, data supports that ToM4 precedes MoV1	Examine Ruler items. Retain theorized connection
ToM6 to CoS3	ToM6 thresholds are less than ToM4 and ToM5. Data supports that ToM4-6 all precede CoS3.	Retain theorized connection
ToM6 to MoV2	ToM6 thresholds are less than ToM4 and ToM5. Data supports that ToM4-6 all precede MoV2	Retain theorized connection
DaD4 to CoS3	Partially supported by data. Similar threshold estimates for DaD4 and CoS3.	Retain theorized connection
DaD5 to Cha4	Data does not support connection. DaD5 appears more difficult than Cha4, but only two Cha4 thresholds.	Test on another sample to gather more Cha4 data.
CoS3 to Cha3	Data does not support connection. CoS3 is more difficult than Cha3.	Test on another sample before removing connection
CoS3 to InI3	Data does not support connection. CoS3 is more difficult than InI3.	Test on another sample before removing connection
CoS4 to InI7	Only one threshold estimate for InI7 and two estimates for CoS4. Not enough data for InI7 in sample.	Test higher performing sample to gather CoS4 and InI7 data.
Cha6 to CoS4	Only two CoS4 threshold estimates. Data supports connection, but not enough data to make conclusion.	Retain theorized connection. Examine more CoS4 items.
Cha6 to InI5	Data supports that Cha6 precedes InI5.	Retain theorized connection.
MoV5 to InI5	Only one threshold estimates for MoV5 in sample.	Test higher performing sample to gather MoV5 data.

connections were supported by the data in this sample, examining the behavior of the students and the relationships of the items across the constructs will aid in the future development of the ADMSR learning progression.

Conclusions and Future Work

The development of the ADMSR learning progression relies on both unidimensional and multidimensional analysis of its seven constructs. By looking at the constructs of the ADMSR learning progression individually, results from the partial credit analysis lead to refinements of the items, the scoring exemplars, and to the construct maps. For this sample of students, the unidimensional partial credit analysis has led to the removal and modification of some items and scoring exemplars, as well as providing validity evidence relating to the hypothesized theory of the construct maps.

With varying degrees of success, we applied a quantitative process for setting cut-points for student ability levels based on the thresholds of the item levels. While some concerns may arise concerning the classification of specific students of abilities near the cut-points, the Wright Maps with cut-point scores can provide meaningful feedback that can assist teachers in determining the ability levels of their students and subsequently inform instruction and curriculum decisions.

By examining all seven of the ADMSR constructs together in a multidimensional analyses, and aligning the dimensions, we examined the estimates of the different levels of the constructs as hypothesized in the learning progression shown in Figure 5. For this sample of students, some of the hypothesized connections in the learning progression were supported by the analysis, while others were not. For the unsupported connections, we hesitate to dismiss them based on the results of the two samples examined here. It might be the case that while no support for the connections between construct levels was present here, there could be evidence of the existence of the connection when we analyze the responses for other students. Examining these connections between constructs helps validate the theory of the AMDSR learning progression, and it also influences the curriculum and instruction.

The development of constructs, and a learning progression, in the BEAR Assessment System is an iterative process. Refinements are made after each sample in analyzed, and then tested on a new sample. The ADMSR project is ongoing in its development of a curriculum and assessment for the data modeling learning progression, and will continue this process with each successive sample.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U. S. Department of Education, through grant R305B110017 to University of California, Berkeley. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Ackerman, T., Gierl, M. J., and Walker, C. M. (2003). Using multidimensional IRT to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, Fall 2003, 37-53.

Adams, R. J. (2006). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*,162–172.

Adams, R. J., and S.-T. Khoo (1993). *Quest: The Interactive Test Analysis System*. Australian Council for Educational Research: Hawthorn, VIC., Australia.

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

Briggs, D. C., and Alonzo, A. C. (2009, June). *The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Briggs, D. C., and Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87-100.

Burmester, K., Zheng, X., Karelitz, T. M., and Wilson, M. R. (2006, March). *Measuring statistical reasoning: Development of an assessment system for data modeling*. Paper presented at the American Education Research Association Annual Meeting, San Francisco, CA.

Diakow, R., Irribarra, D. T., and Wilson, M. (2012, April). *Analyzing the complex structure of a learning progression: Structured construct models*. Paper presented at the National Council on Measurement in Education Annual Meeting, Vancouver, Canada.

Draney, K. (2009, June). *Designing learning progressions with the BEAR assessment system*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Duncan, R. G., and Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching, 46*, 606-609.

Embretson, S.E., and Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Horvath, J., and Lehrer, R. (1998). A model-based perspective on the development of children's understandings of chance and uncertainty. In S. P. Lajoie (Ed.), *Reflections on statistics* (pp. 121-148). Mahwah, NJ: Lawrence Erlbaum.

Kennedy, C. A. (2005). Constructing Measurement Models for MRCML Estimation: A Primer for Using the BEAR Scoring Engine. *BEAR Technical Report Series 2005-04-02*. Berkeley, CA: University of California, BEAR Center.

Kennedy, C. A., and Wilson, M. (2007). Using progress variables to map intellectual development. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in schools: Intellectual growth and standard setting* (pp. 271-298). Maple Grove, MN: JAM Press.

Lehrer, R., and Kim, M. J. (2009). Structuring variability by negotiating its measure. *Mathematics Education Research Journal, 21*, 116-133.

Lehrer, R., Kim, M-J., Ayers, E., and Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey, and K. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31-60). Charlotte, NC: Information Age Publishers..

Lehrer, R., and Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14*, 69-108.

Lehrer, R., Schauble, L., Wilson, M. R., Lucas, D. D., Karelitz, T. M., Kim, M., et al. (2007, March). *Collaboration at the boundaries: Brokering learning and assessment. improves the quality of education*. Paper presented at the American Education Research Association Annual Meeting in Chicago, IL.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: MESA press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, No. 15.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction, 16*, 285.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics 26 - Psychometrics* (pp. 607-642). Amsterdam, the Netherlands: Elsevier.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike information criterion statistics*. Tokyo: KTK Scientific Publishers.

Schwartz, R., and Ayers, E. (2011). *Delta dimensional alignment: Comparing performances across dimensions of the learning progression for assessing data modeling and statistical reasoning*. Unpublished manuscript, University of California, Berkeley.

Smith, C., Wiser, M., Anderson, C. W., and Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 14*(1 and 2), 1-98.

Snidjers, T., and Bosker, R. (1999). *Mulitlevel analysis*. London, UK: Sage.

Sympson, J. B. (1978). *A model for testing with multidimensional items*. Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota.

van der Linden, W. J., and Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

Wang, W. C. (1995). *Implementation and application of the multidimensional random coefficients multinomial logit*. Unpublished doctoral dissertation. University of California, Berkeley, Berkeley, CA.

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research, 8*, 1-22.

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

Wilson, M. (2009a). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching, 46*, 716-730.

Wilson, M. (2009b, June). *Structured Constructs Models (SCM): A family of statistical models related to learning progressions*. Invited plenary address at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Wilson, M., and Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.

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

Wu, M., Adams, R., Wilson, M., and Haldane, S. (2007). ACER Conquest 2.0 [Computer software and manual]. Hawthorn, VIC, Australia: ACER.