Title
The Development and Psychometric Modeling of an Embedded Assessment for a Data Modeling and Statistical Reasoning Learning Progression

Permalink
https://escholarship.org/uc/item/96j6w7xk

Author
Schwartz, Robert Andrew

Publication Date
2012-01-01

Peer reviewed|Thesis/dissertation
The Development and Psychometric Modeling of an Embedded Assessment for a Data Modeling and Statistical Reasoning Learning Progression

By

Robert Andrew Schwartz

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Sophia Rabe-Hesketh
Professor Steve Selvin

Fall 2012
Abstract

The Development and Psychometric Modeling of an Embedded Assessment for a Data Modeling and Statistical Reasoning Learning Progression

by

Robert Andrew Schwartz

Doctor of Philosophy in Education
University of California, Berkeley

Professor Mark Wilson, Chair

Data modeling is an approach that helps students to transform 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 single 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 research applies different measurement models to the data from the ADMSR project. Unidimensional item response (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. Longitudinal and multilevel IRT models are applied to model the change in students’ abilities to see if students improved on the constructs after instruction. Finally, 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.
# Table of Contents

List of Tables.................................................................................................................... iii

List of Figures................................................................................................................... iv

## Chapter

1. **Data Modeling and the BEAR Assessment System**........................................... 1
   - Assessing Data Modeling and the BEAR Assessment System.............................. 1

2. **Construct Development and Validation**......................................................... 11
   - The Constructs........................................................................................................... 11
   - The Data Display (DaD) Construct........................................................................ 11
   - The Conceptions of Statistics (CoS) Construct..................................................... 12
   - Sample and Assessment Data............................................................................... 13
   - Modeling the Data using Item Response Models............................................... 14
   - Instrument Properties: Item Fit and Reliability.................................................... 15
   - Using Wright Maps to Validate the Proposed Theory of Development of the Construct Maps – Data Display................................................................. 18
   - Using Wright Maps to Validate the Proposed Theory of Development of the Construct Maps – Conceptions of Statistics....................................................... 21
   - Conclusions and Implications................................................................................ 23

3. **Testing for Gains over Time Using Multilevel Item Response Models**............. 30
   - Item Response Models to Test for Gains Over Time.......................................... 30
   - Pre-Post Design and Method.................................................................................. 33
   - Results – Model Fit Comparisons........................................................................ 34
   - Results – Conceptions of Statistics........................................................................ 36
Results – Data Display...........................................................................................................39

Discussion and Conclusions.................................................................................................41

4 Analyzing the Learning Progression Using a Multidimensional Item Response Model..........................................................................................................................47

Introduction...............................................................................................................................47

The Multidimensional Random Coefficients Multinomial Logit Model...............................48

Sample and Assessment Data.................................................................................................50

Results – Fitting the MRCML to the ADMSR Data.................................................................51

Aligning the Dimensions – Delta Dimensional Alignment.......................................................53

Simulation of Delta Dimensional Alignment Technique.......................................................55

Examining the Links – Results from Aligning the ADMSR Data..........................................57

Conclusions and Future Work.................................................................................................63

References...............................................................................................................................75

Appendix Chapter 3 Stata/GLLAMM Commands.................................................................80
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Data Display Weighted Fit Statistics</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>Conceptions of Statistics Weighted Fit Statistics</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>EAP/PV Reliabilities by Construct</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Conceptions of Statistics Model Fit</td>
<td>35</td>
</tr>
<tr>
<td>3.2</td>
<td>Data Display Model Fit</td>
<td>36</td>
</tr>
<tr>
<td>3.3</td>
<td>Conceptions of Statistics Estimated Parameter Values</td>
<td>37</td>
</tr>
<tr>
<td>3.4</td>
<td>Data Display Estimated Parameter Values</td>
<td>39</td>
</tr>
<tr>
<td>4.1</td>
<td>ADMSR Post-test Items</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Relative Model Fit between Unidimensional and Seven-dimensional</td>
<td>51</td>
</tr>
<tr>
<td>4.3</td>
<td>Correlations of the Seven Constructs/Dimensions</td>
<td>52</td>
</tr>
<tr>
<td>4.4</td>
<td>Threshold Estimates Before Alignment for 30 Replications</td>
<td>55</td>
</tr>
<tr>
<td>4.5</td>
<td>Threshold Estimates After DDA Technique for 30 Replications</td>
<td>56</td>
</tr>
<tr>
<td>4.6</td>
<td>Summary of Results for Theorized Connections in ADMSR Learning Progression</td>
<td>62</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Data Modeling Integrates Inquiry, Data, Chance and Inference</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>Conceptions of Statistics (CoS) Construct Map from the ADMSR Learning Progression</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>The &quot;Kayla's Project&quot; Item from ADMSR</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>The Scoring Exemplar for &quot;Kayla's project&quot;</td>
<td>9</td>
</tr>
<tr>
<td>1.5</td>
<td>A Map of the Theoretical ADMSR Learning Progression</td>
<td>10</td>
</tr>
<tr>
<td>2.1</td>
<td>Data Display (DaD) Construct Map from the ADMSR Learning Progression</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>Data Display Initial Wright Map</td>
<td>26</td>
</tr>
<tr>
<td>2.3</td>
<td>Data Display Wright Map with Cut-points</td>
<td>27</td>
</tr>
<tr>
<td>2.4</td>
<td>Conceptions of Statistics Wright Map</td>
<td>28</td>
</tr>
<tr>
<td>2.5</td>
<td>Conceptions of Statistics Wright Map with Cut-points</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>Scatterplot of Conceptions of Statistics Classroom-Level Effects</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>Scatterplot of Conceptions of Statistics Step Difficulty Estimates</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>Scatterplot of Data Display Classroom-Level Effects</td>
<td>45</td>
</tr>
<tr>
<td>3.4</td>
<td>Scatterplot of Data Display Step Difficulty Estimate</td>
<td>46</td>
</tr>
<tr>
<td>4.1</td>
<td>A Map of the Theoretical ADMSR Learning Progression</td>
<td>64</td>
</tr>
<tr>
<td>4.2</td>
<td>Seven Dimensional Wright Map of the ADMSR Learning Progression</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>Threshold Estimates for Items 5 and 6 with No Alignment</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>Threshold Estimates for Items 5 and 6 with Delta Dimensional Alignment</td>
<td>67</td>
</tr>
<tr>
<td>4.5</td>
<td>Threshold Estimates for Items 10 and 11 with No Alignment</td>
<td>68</td>
</tr>
</tbody>
</table>
4.6 Threshold Estimates for Items 10 and 11 with Delta Dimensional Alignment.....69
4.7 Aligned Wright Map of the ADMSR Learning Progression Using DDA.........70
4.8 Link Between ToM4 with DaD1 and MoV1.......................................................71
4.9 Link Between ToM6 with CoS3 and MoV2.......................................................72
4.10 Bootstrapping Between DaD and MRC.........................................................73
4.11 Link Between Cha6 with CoS4.................................................................74
Chapter 1

Data Modeling and the BEAR Assessment System

In order to master both formal statistics and informal inference, one must first fully understand the different concepts underlying data analysis or 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, that can serve as a basis for more advanced statistical reasoning. Data Modeling (Lehrer & Romberg, 1996; Horvath & 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 coordinates chance together with variability. It accomplishes this by designing learning environments that support this kind of reasoning by allowing students to invent and revise models (Lehrer & Kim, 2009).

The different components of statistical reasoning are integrated to form the Data Modeling approach to learning, and are represented by Figure 1.1.

As Figure 1.1 illustrates, Data Modeling arises out of an inquiry about some 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 & Wilson, in press).

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 and Data Modeling (Burmester, Zheng, Karelitz, & Wilson, 2006; Lehrer 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 & 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.

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 represent a cognitive theory of learning consistent with a developmental perspective. Figure 1.2 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, represents a continuum of ability where lower levels of the latent trait at the bottom of the construct map going up toward more expert levels at the top end of the map. This continuum of ability is broken up into representing “reference points” along the continuum (i.e., the “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 be described in greater detail in the following chapters.

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 Model Measure 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 construct represents the basic area of knowledge in which the rest of the constructs are played out. 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 performances 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 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 by type. 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 1.3.

![Insert Figure 1.3 here](image)

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 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 1.4.
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 relevant responses, but 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 as seven individual constructs and also all together as a single learning progression.

A collection of construct maps taken together can comprise a learning progression (Draney, 2009; Wilson, 2009). 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, & 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 & Hmelo-Silver, 2009).

There are different ways to conceive and measure learning progressions. The Berkeley Evaluation and Assessment Research (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 1.5. An arrow represents 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.

-----------------------------------

Insert Figure 1.5 here

-----------------------------------

Conclusion

The following chapters apply different measurement models to the data from the ADMSR project. In the next chapter, Construct Development and Validation, two of the constructs will be examined to see if the proposed theory of development presented by the construct map is supported by the results from the administration of the instrument to the sample of students. These constructs are analyzed individually by applying a unidimensional measurement model to a given construct. In the chapter Pre-Post Analysis Using Multilevel IRT Models, the change in students’ abilities from before until after the treatment will be modeled to see if students improved on the constructs after instruction. These changes in abilities will be examined using measurement models that take into account the multilevel structure of the data. In the final chapter, Analyzing the Learning Progression Using Multidimensional IRT, 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.
Figure 1.1 – *Data Modeling Integrates Inquiry, Data, Chance and Inference* (Lehrer, Kim, Ayers & Wilson, *in press*).
<table>
<thead>
<tr>
<th>Conceptions of Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CoS4 - Investigate and anticipate qualities of a sampling distribution.</strong></td>
</tr>
<tr>
<td><strong>CoS4D</strong></td>
</tr>
<tr>
<td><strong>CoS4C</strong></td>
</tr>
<tr>
<td><strong>CoS4B</strong></td>
</tr>
<tr>
<td><strong>CoS4A</strong></td>
</tr>
<tr>
<td><strong>CoS3 - Consider statistics as measures of qualities of a sample distribution.</strong></td>
</tr>
<tr>
<td><strong>CoS3F</strong></td>
</tr>
<tr>
<td><strong>CoS3E</strong></td>
</tr>
<tr>
<td><strong>CoS3D</strong></td>
</tr>
<tr>
<td><strong>CoS3C</strong></td>
</tr>
<tr>
<td><strong>CoS3B</strong></td>
</tr>
<tr>
<td><strong>CoS3A</strong></td>
</tr>
<tr>
<td><strong>CoS2 - Calculate statistics.</strong></td>
</tr>
<tr>
<td><strong>CoS2B</strong></td>
</tr>
<tr>
<td><strong>CoS2A</strong></td>
</tr>
<tr>
<td><strong>CoS1 - Describe qualities of distribution informally.</strong></td>
</tr>
<tr>
<td><strong>CoS1A</strong></td>
</tr>
</tbody>
</table>

Figure 1.2 – Conceptions of Statistics (CoS) Construct Map from the ADMSR Learning Progression
Kayla’s project

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

<table>
<thead>
<tr>
<th>Kayla’s Projects – Points Earned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project 1</td>
</tr>
<tr>
<td>Project 2</td>
</tr>
<tr>
<td>Project 3</td>
</tr>
<tr>
<td>Project 4</td>
</tr>
</tbody>
</table>

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

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

Figure 1.3 – The "Kayla's Project" Item from ADMSR
<table>
<thead>
<tr>
<th>Levels</th>
<th>Response Examples</th>
<th>Example of Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoS3D</td>
<td>Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among its components. Students use strategies that indicate they understand the relations between the individual scores and the mean. Students might • compare the distance of each value to the mean of 17 • multiply 17 x 4 to find the total score that the individual values need to sum to</td>
<td>• “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.”</td>
</tr>
<tr>
<td>CoS3D-</td>
<td>Predict how a statistic is affected by changes in its components or otherwise demonstrate knowledge of relations among its components. Students use a strategy described in CoS3D but make mistakes in calculations.</td>
<td>Use this information to find the mean of points Kayla received on Project 4. Show your work.</td>
</tr>
<tr>
<td>CoS2A</td>
<td>Calculate statistics indicating central tendency. Students show that they are able to calculate the mean. • Students who only show dividing 56 by 4 and addition of all the values belong in this level. • Students might write the correct answer but not show work</td>
<td>She received 19 points.</td>
</tr>
</tbody>
</table>

**Exemplar: Kayla's Project**

**Question 1:** The number of points Kayla received on project 4 (mean of 17)

<table>
<thead>
<tr>
<th>Levels</th>
<th>Response Examples</th>
<th>Example of Student Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoS2A-</td>
<td>Calculate statistics indicating central tendency. Students write correct equation to calculate the missing measure. Students provide incorrect answer, but their work shows that they know the procedure.</td>
<td></td>
</tr>
<tr>
<td>NLI(ii)</td>
<td>Answer is incorrect with no work shown but within a relevant range of answers. Students do some idiosyncratic calculation with the measures. Students begin to carry out a strategy correctly but not to completion.</td>
<td></td>
</tr>
</tbody>
</table>

**Example 1:**

**Example 2:**

**Example 3:**

**Example 4:**

**Example 5:**

**Example 6:**

**NLI(i)** Attempt item but answers are irrelevant, unclear, implausible, unreasonable or demonstrate that student did not understand the item

| M | Missing response |

*Mock student responses*
Figure 1.5 – A Map of the Theoretical ADMSR Learning Progression
Chapter 2

Construct Development and Validation

The 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 facilitated the use of the student level results to evaluate progress on the latent variable. This chapter will examine two of the seven constructs, Data Display (DaD) and Conceptions of Statistics (CoS), 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., 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 2.1.

-------------------------------------------

Insert Figure 2.1 here

-------------------------------------------

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 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 1.2, 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. Of these parts, 18 were CoS items and 21 were DaD items.

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 scores in common. Rater analyses were performed on the data for each of the seven constructs. Differences in rating patterns were identified using a multi-faceted Item Response 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 conflicting scores was from a rater who was part of the creation of that item’s scoring guide, then that score was expected to be more reliable and chosen. If neither of the scorers was involved in the creation of the scoring guide, then neither score was deemed more reliable, and one of the conflicting 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.

**Modeling the Data using Item Response Models**

To determine if each construct is being properly measured, the data is analyzed within an Item Response Theory (IRT) framework. An item response model (IRT; Rasch, 1960/1980; Andrich, 1978; Masters, 1982; Wright & Masters, 1982; van der Linden & Hambleton, 1997; Embretson & Reise, 2000) 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 (which would be required by the stricter Rating Scale Model). The PCM, specified in Equation 2.1, models the probability of person p responding in category j of item i as a function of the person ability \( \theta_p \) and step parameters \( \delta_{ij} \),

\[
Pr(x_{ip} = j | \theta_p) = \frac{\exp \sum_{l=0}^{j} (\theta_p - \delta_{il})}{\sum_{l=0}^{m_i} \exp \sum_{l=0}^{j} (\theta_p - \delta_{il})}, \quad j = 0,1,\ldots,m_i \tag{2.1}
\]

where \( \sum_{l=0}^{0} (\theta_p - \delta_{il}) = 0 \), \( \theta_p \sim N(0, \psi) \), 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 2.2) 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 \( \theta_p \) and step difficulty \( \delta_{ij} \),

\[
\ln \frac{Pr(x_{ip} = j | \theta_p)}{Pr(x_{ip} = j-1 | \theta_p)} = \theta_p - \delta_{ij}, \quad j = 0,1,\ldots,m_i \tag{2.2}
\]
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, & 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 statistics (Wright and Masters, 1982) for the item parameter estimates. This statistic is also sometimes known as the “infit” statistic. An infit statistic at a value of 1.0 represents “perfect infit”. 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 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 is behaving more consistently (i.e. less randomness) with the rest of the items than was expected—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 & 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.

The weighted mean square fit statistics for the items on the DaD construct are displayed in Table 2.1. 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.

Table 2.1. *Data Display Weighted Fit Statistics*

<table>
<thead>
<tr>
<th>Item</th>
<th>Weighted Fit Statistic</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>GottaGo1</td>
<td>0.92</td>
<td>(0.89, 1.11)</td>
</tr>
<tr>
<td>GottaGo2</td>
<td>0.93</td>
<td>(0.91, 1.09)</td>
</tr>
<tr>
<td>GottaGo3</td>
<td>0.88</td>
<td>(0.86, 1.14)</td>
</tr>
<tr>
<td>App1</td>
<td>1.14*</td>
<td>(0.90, 1.10)</td>
</tr>
<tr>
<td>Bowl1</td>
<td>0.98</td>
<td>(0.90, 1.10)</td>
</tr>
<tr>
<td>Bowl2</td>
<td>1.02</td>
<td>(0.91, 1.09)</td>
</tr>
<tr>
<td>Crab1MC</td>
<td>1.02</td>
<td>(0.88, 1.12)</td>
</tr>
<tr>
<td>Crab2</td>
<td>1.17*</td>
<td>(0.84, 1.16)</td>
</tr>
<tr>
<td>Crab3</td>
<td>0.99</td>
<td>(0.86, 1.14)</td>
</tr>
<tr>
<td>EQ1</td>
<td>0.91</td>
<td>(0.89, 1.11)</td>
</tr>
<tr>
<td>Head2</td>
<td>1.03</td>
<td>(0.91, 1.09)</td>
</tr>
<tr>
<td>LtCherry</td>
<td>1.21*</td>
<td>(0.87, 1.13)</td>
</tr>
<tr>
<td>Max5</td>
<td>0.99</td>
<td>(0.93, 1.07)</td>
</tr>
<tr>
<td>Rocket1</td>
<td>0.76*</td>
<td>(0.83, 1.17)</td>
</tr>
<tr>
<td>Rocket2</td>
<td>1.14</td>
<td>(0.75, 1.25)</td>
</tr>
<tr>
<td>Candle1</td>
<td>0.93</td>
<td>(0.89, 1.11)</td>
</tr>
<tr>
<td>Candle2</td>
<td>0.98</td>
<td>(0.88, 1.12)</td>
</tr>
<tr>
<td>Statue4</td>
<td>0.98</td>
<td>(0.88, 1.12)</td>
</tr>
<tr>
<td>Statue6</td>
<td>1.01</td>
<td>(0.86, 1.14)</td>
</tr>
<tr>
<td>Statue8</td>
<td>0.93</td>
<td>(0.88, 1.12)</td>
</tr>
<tr>
<td>StateCap2</td>
<td>1.39**</td>
<td>(0.87, 1.13)</td>
</tr>
</tbody>
</table>

The weighted mean square fit statistics for the items on the CoS construct are displayed in Table 2.2. 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.
Table 2.2. Conceptions of Statistics Weighted Fit Statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Weighted Fit Statistic</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max4MC</td>
<td>0.98</td>
<td>(0.90, 1.10)</td>
</tr>
<tr>
<td>TallestTree1</td>
<td>0.86</td>
<td>(0.75, 1.25)</td>
</tr>
<tr>
<td>TallestTree2</td>
<td>0.80</td>
<td>(0.79, 1.21)</td>
</tr>
<tr>
<td>BallMedian</td>
<td>0.82</td>
<td>(0.82, 1.18)</td>
</tr>
<tr>
<td>BallMode</td>
<td>0.82*</td>
<td>(0.83, 1.17)</td>
</tr>
<tr>
<td>BallMean</td>
<td>0.90</td>
<td>(0.87, 1.13)</td>
</tr>
<tr>
<td>Ball2</td>
<td>0.85*</td>
<td>(0.88, 1.12)</td>
</tr>
<tr>
<td>Ball3</td>
<td>0.81</td>
<td>(0.88, 1.12)</td>
</tr>
<tr>
<td>Caffeine2</td>
<td>1.16</td>
<td>(0.76, 1.24)</td>
</tr>
<tr>
<td>Corn2</td>
<td>1.32*</td>
<td>(0.86, 1.14)</td>
</tr>
<tr>
<td>Kayla1</td>
<td>0.90</td>
<td>(0.80, 1.20)</td>
</tr>
<tr>
<td>Range2</td>
<td>1.07</td>
<td>(0.76, 1.24)</td>
</tr>
<tr>
<td>Swimming1</td>
<td>1.30*</td>
<td>(0.81, 1.19)</td>
</tr>
<tr>
<td>Swimming2</td>
<td>0.83</td>
<td>(0.80, 1.20)</td>
</tr>
<tr>
<td>Swimming3</td>
<td>0.86</td>
<td>(0.81, 1.19)</td>
</tr>
<tr>
<td>Battery1</td>
<td>1.11</td>
<td>(0.80, 1.20)</td>
</tr>
<tr>
<td>Battery2</td>
<td>1.18*</td>
<td>(0.83, 1.17)</td>
</tr>
<tr>
<td>FreeThrow</td>
<td>1.58**</td>
<td>(0.83, 1.17)</td>
</tr>
</tbody>
</table>

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 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 2.3.

Table 2.3. EAP/PV Reliabilities by Construct

<table>
<thead>
<tr>
<th>Construct</th>
<th>Real Data</th>
<th>Simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Display (DaD)</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>Conceptions of Statistics (CoS)</td>
<td>0.64</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Note the reliability increase for both of the constructs even with the reduction of items, due to the elimination of the missing data.

Using Wright Maps to Validate the Proposed Theory of Development of the Construct Maps – 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 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 = 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 2.2. 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 2.2 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 2.2, 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 called “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 level “No Link(i)” and “Missing” responses to “No Link(ii)”. The lower level “No Link(i)” represents responses that do not even have a relevance to the lowest level of the construct. “Missing” responses represent where a student saw the item, but did not provide an answer. These responses do not include missing by design, where students did not have an opportunity to answer the item. The level of “No Link(ii)” represents those responses that are at least somewhat relevant to the item, but are not high enough to be considered level 1 responses. Since each column represents thresholds that correspond to a given level of the construct map, 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 2.2 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 2.1.
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 along the Wright Map based on the judgments of experts (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 & Wilson (2007b):

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 numbers of responses at a given level that rendered the threshold estimate unreliable, cut-points were set for the Data Display construct and a new Wright Map with these levels is displayed in Figure 2.3.

---

Insert Figure 2.3 here

---

The Wright Map in Figure 2.3 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 & Alonzo (2009), this process of setting cut-points has some inherent potential problems, some of which are present 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 likely due to the absence of students who
performed near the top levels of the construct. 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 either theses 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 highest part of the lower level or the lowest 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 2.3 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 Wright Maps to Validate the Proposed Theory of Development of the Construct Maps – 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 2.4.

---------------------------
Insert Figure 2.4 here
---------------------------

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.

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 2.4), 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 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, rendering the threshold estimate unreliable, cut-points were set for the Conceptions of Statistics construct and a new Wright Map with these levels is displayed in Figure 2.5.
The Wright Map in Figure 2.5 is a modified version of Figure 2.4, 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 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.

**Conclusions and Implications**

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 provided 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.

The development of constructs, and a learning progression, in the Bear Assessment System is an iterative process. The results presented here represent the
current state of the project and what can be learned after an iteration in the overall process, and not the final result. 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.
<table>
<thead>
<tr>
<th>Data Display</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DaD6</strong> – Integrate case with aggregate perspectives.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD6A</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD5</strong> – Consider the data in aggregate when interpreting or creating displays.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD5B</strong></td>
</tr>
<tr>
<td><strong>DaD5A</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD4</strong> – Recognize or apply scale properties to the data.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD4B</strong></td>
</tr>
<tr>
<td><strong>DaD4A</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD3</strong> – Notice or construct groups of similar values.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD3A</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD2</strong> – Focus on individual values when constructing or interpreting displays of data.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD2B</strong></td>
</tr>
<tr>
<td><strong>DaD2A</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD1</strong> – Create displays or interpret displays without reference to goals of data creation.</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>DaD1A</strong></td>
</tr>
</tbody>
</table>

Figure 2.1 – Data Display (DaD) Construct Map from the ADMSR Learning Progression
Figure 2.2 – Data Display Initial Wright Map
Figure 2.3 – Data Display Wright Map with Cut-points
Figure 2.4 – Conceptions of Statistics Wright Map
Figure 2.5 – Conceptions of Statistics Wright Map with Cut-points
Chapter 3

Testing for Gains over Time Using Multilevel Item Response Models

In order to make any conclusions regarding the effectiveness of the Data Modeling instruction and curriculum, we must explore whether the intervention benefitted the students. We do this by examining whether or not students’ had significant gains in ability during their exposure to the curriculum. To test whether there were gains, the ADMSR project administered pre-tests to the students before instruction, and post-tests afterward.

In this chapter, we introduce a number of different item response models that can be fit to the pre-test and post-test data to determine if there are gains in student abilities. We then fit these models to the previously introduced Conceptions of Statistics (CoS) and Data Display (DaD) constructs. After model comparison and the selection of the best fitting model, we further examine the results of the analysis to determine the amount of gain in student ability between the pre-test and post-test and if the gain varies for different classrooms. In addition to examining the effects of time and class membership on student abilities, we will also examine how the item difficulty estimates change as we apply different models to the data, since one of the goals in developing the curriculum is to have a “bank” of items with anchored item difficulties that can be used to test other populations.

Item Response Models to Test for Gains Over Time

The item response model applied in the previous chapter (the Partial Credit Model; “PCM”) and the others that will be presented here all attempt to measure a certain latent ability – i.e. an ability that is not directly observable. To measure latent ability, item response models do so indirectly, by examining a student’s responses to a set of items. The PCM in the prior chapter measured a student’s ability based solely on the post-test. When the data is more complex and has repeated measures of the students, such as a pre-test and post-test, the method of measuring the latent ability is more complex. The student’s ability could be measured separately at the different time points (pre-test or post-test), or as a combination of the student’s responses at different time points (both pre-test and post-test).

The first model that we will fit to the data here is once again the PCM. The specification of the PCM is once again given in Equation 3.1, this time as the log ratio of the probability of the test taker (or person) $p$ at occasion $o$ responding in category $j$ of item $i$ to the probability of responding in category $j$ - 1 as a function of the ability of the test taker at that occasion $\theta_{op}$ and step difficulty $\delta_{ij}$,
where $\theta_{op} \sim N(0, \psi_1)$, and $m_i$ is the total number of steps in item $i$. This is slightly different than the formulation given in Equation 2.2 in that it models the ability for a person at an occasion rather than for a given person. Thus, a person who has taken the test at both occasions will have two ability values for this model, one for the pre-test and one for the post-test. Take note, however, that modeled this way, the PCM does not take into account any relationship or association between a student’s two abilities at the different occasions. The results of the PCM analysis could be examined to determine student gain by running a analysis (such as a regression analysis) on the student abilities using the time the test was taken as a predictor. Recent research, however, has found bias in the standard errors and the parameter estimates when conducting such a two-step analysis (Koren & Wang, 2012).

Thus, in this chapter, we apply a series of IRT models that examine the growth of student abilities directly from their scored item responses.

The next model to be fit to the data is an explanatory extension of the PCM, a latent regression PCM. An explanatory model uses either person or item properties to explain the effects of either persons or items (DeBoeck & Wilson, 2004). A latent regression PCM is an example of a person explanatory model. The latent regression PCM model applied here acknowledges a relationship between the pre-test and post-test of a given student and models the longitudinal change of person ability between the two tests as a fixed effect. In reality, the person ability $\theta_p$ from the previous chapter’s PCM was really the person ability at a given occasion $o$, because that dataset only had persons on one occasion. In Equation 3.2, that ability is now modeled to include a fixed effect $\beta_1$ where $\text{time}_{op}$ has a value of 0 if the item was taken on the pre-test and a value of 1 if the item was taken on the post-test, and a random effect for each student $\zeta_{1p}$:

$$\theta_{op} = \beta_1 \text{time}_{op} + \zeta_{1p}$$  \hspace{1cm} (3.2)

where $\zeta_{1p} \sim N(0, \psi_1)$. The fixed effect $\beta_1$ parameter for the model represents the mean change in student ability from pre-test to post-test. We expect to see a significant positive change from pre-test to post-test to represent the learning of the students from being exposed to the Data Modeling curriculum. The ability estimate for a student when taking the pre-test (where $\text{time}_{op} = 0$) is simply the random effect $\zeta_{1p}$, while the ability estimate for a student when taking the post-test (where $\text{time}_{op} = 1$) is that pre-test estimate plus the mean change in ability $\beta_1$.

While accounting for a change in time in the ability of students taking the test, the model given in Equation 3.2 has some obvious limitations. Specifically, it does not allow for variation of the change in ability between students. While we would hope that all students will have a similar large change in ability from the pre-test to the post-test, this
would not realistically occur. Thus, a model is needed that allows for variation in individual changes in ability over time.

Multilevel modeling is a general framework that allows for the modeling of nested data structures (Goldstein, 1995; Raudenbush & Bryk, 2002). Repeated measures over time by one respondent are one type of nested data, and thus the multilevel framework can be applied to growth curve modeling of longitudinal data (Bryk & Raudenbush, 1987). Modeling change in longitudinal data for individuals has previously been explored in the context of IRT (Andersen, 1985; Embretson, 1991; Fischer, 2001). Anderson (1985) applied the multidimensional Rasch model to different occasions by modeling different correlated abilities at each occasion. Embretson (1991) also applied a multidimensional Rasch model to repeated measures, but in contrast to Andersen, provided parameters for individual differences in change by assuming that at each additional time point a new additional dimension (a change ability) is involved in the performance. Embretson’s model has since been extended for use of polytomous items (Fischer, 2001).

The next model fit to the data, specified in Equation 3.3, builds on Equation 3.2 but also models the longitudinal structure of the data by allowing for individual variation in longitudinal change by adding a random effect for time $\zeta_{2p}$ for each student. $\zeta_{2p}$ is the individual deviation from the mean of a student’s change in ability from pre-test to post-test, and $\beta_1$ is the mean change in ability. In effect, Equation 3.3 extends Equation 3.2 from a random intercept model to a random slope model.

\[
\theta_{op} = (\beta_1 + \zeta_{2p}) \text{time}_{op} + \zeta_{1p}
\]  

(3.3)

where $\zeta_{1p} \sim N(0,\psi_1)$, $\zeta_{2p} \sim N(0,\psi_2)$, and $\text{Cor}(\zeta_{1p}, \zeta_{2p}) = \rho$. Applying this model, a student’s ability on the pre-test would be estimated by the random effect $\zeta_{1p}$, and a student’s ability on the post-test would be the sum of their pre-test ability, the mean change in ability, and their deviation from that mean change in ability. Equation 3.3 is an application of the Fischer extension of the Embretson model – it models multiple abilities, an initial ability ($\zeta_{1p}$) and ability for change ($\beta_1 + \zeta_{2p}$).

Other, more traditional, types of multilevel data structures are common in educational research, and the data here are no exception - students in the ADMSR dataset are clustered in classes. When clustering of data occurs, the independence assumption that underlies the modeling of the data may be violated (Raudenbush & Byrk, 2002). In order to account for the students clustered in classes, the next model to be fitted extends Equation 3.3 by adding a multilevel structure, in addition to the previously described longitudinal data structure, that takes into consideration the dependence of units in the
same cluster, students in classrooms. The model specified in Equation 3.4 extends the model specified in Equation 3.3 by considering the effects of time that were added to the PCM and by adding two additional random effects: (1) a classroom-level random effect for students’ ability on the construct, and (2) a classroom-level random effect for the longitudinal change in ability.

\[
\theta_{opc} = \left( \beta_1 + \zeta_{2pc} + \xi_{4c}^{(2)} \right) \text{time}_{opc} + \xi_{1pc} + \xi_{3c}^{(2)} 
\]

In Equation 3.4, \( \zeta_{1p} \sim N(0, \psi_1) \), \( \zeta_{2p} \sim N(0, \psi_2) \), \( \xi_{3c}^{(2)} \sim N(0, \psi_3) \), \( \xi_{4c}^{(2)} \sim N(0, \psi_4) \), Cor(\( \zeta_{1pc} \), \( \zeta_{2pc} \)) = \rho_1 \), and Cor(\( \xi_{3c}^{(2)}, \xi_{4c}^{(2)} \)) = \rho_2 \). In this formulation, \( \xi_{3c}^{(2)} \) represents the classroom level random effect for ability and \( \xi_{4c}^{(2)} \) represents the classroom level random effect for change in ability from pre-test to post-test. This model is multidimensional not only at the individual level, but also at the classroom-level. In this model, a student’s pre-test ability is estimated by the classroom level mean ability on the pre-test and the student’s individual deviation from this mean. The student’s post-test ability will be the sum of this pre-test ability, the mean change in ability, the classroom-level deviation from the mean ability, and the individual student deviation from that classroom-level change in ability. By fitting such a model that includes classroom level effects, not only can we separate out student level versus classroom level effects, but also we can examine and compare the differences in abilities and changes in abilities between the different classrooms.

**Pre-Post Design and Method**

Assessments were administered at the beginning and ending of the student’s exposure to the data modeling curriculum. 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. Both the pre-tests and the post-tests were designed to cover all 7 of the data modeling constructs. The students ranged in grade level from grades four through seven, and were located in Arkansas and Wisconsin public schools. Once again, this chapter only fits the above-described models and discusses the results from two of the seven constructs from the ADMSR project, the Data Display (DaD) and Conceptions of Statistics (CoS) constructs.

Both the pre-test and post-test were designed to have multiple forms of the test so that all 7 of the constructs could be covered. Given that a minimum number of items needs to be administered for a given construct in order to be able to estimate individual student abilities, a single test form would have been too lengthy. Since only 50 minutes

---

1 Data in item response modeling often follow a multilevel structure, including the previous chapter’s data that was analyzed using the Partial Credit Model. The PCM can be considered as a multilevel model with items that are nested in students, where the item difficulties are fixed effects and the person abilities are random effects. In this chapter, however, when we refer to “multilevel IRT models” we are describing models with an additional level structure in addition to items nested in students.
were allotted for each assessment, a matrix sampling design was used. This design did not administer items from every construct to every student. Each test form was designed to cover different combinations of at least four of the seven constructs. For the Arkansas classrooms, three of the five pre-test forms and three of the five post-test forms covered the Conceptions of Statistics construct. Similarly, for the Data Display construct the Arkansas pre-test included DaD items on three of five forms. The Arkansas post-test, however, included enough DaD items to estimate individual student ability on all five of the forms. In Wisconsin, all of the students were given pre-tests and post-tests that included the CoS and DaD constructs.

The scoring process and rater analysis for the pre-tests was the same as the method for the post-test scoring and rater analysis described in the prior chapter. After the pre-tests and post-tests from all sites were scored and rater reliabilities were checked, student scores were combined into a single dataset containing all the students.

For the combined Arkansas and Wisconsin sites, there were a total of 687 students who were administered a form that tested CoS on either the pre-test or post-test. Of those 687 students, 293 of them were tested for CoS on both the pre-test and the post-test. There were a total of 12 multi-part CoS items, with 19 individually scored parts. For the DaD construct, 956 different students took at least one test form containing the DaD items. 497 of those students were administered DaD items on both the pre-test and post-test. There were a total of 12 multi-part DaD items, with 22 individually scored parts.

The earlier described models can all be placed within the generalized linear latent and mixed modeling (GLLAMM) framework and fit to the data by using the program gllamm (Rabe-Hesketh, Skrondal, and Pickles, 2004; Zheng & Rabe-Hesketh, 2007) in the Stata software package (StataCorp, 2011). All models were fit using adaptive quadrature as the integration method and Newton-Raphson for maximum likelihood estimation (Rabe-Hesketh et al., 2005).

Results – Model Fit Comparisons

After fitting all four of the IRT models to the Data Display and Conceptions of Statistics data, we compare how well the models fit the data. To assess and compare the model fit, we look at the log-likelihood (LL), the Akaike Information Criterion (AIC) (Sakamoto, Ishiguro & Kitigawa, 1986), and the Bayesian Information Criterion (BIC) (Schwarz, 1978). The AIC is defined as $AIC = 2k - 2LL$, where $k$ is the number of free parameters in the model and $LL$ is the log of the maximum likelihood of the parameter estimates. The BIC is defined as $BIC = k \cdot \ln(n) - 2LL$, where $n$ denotes the number of observations in the dataset. When comparing the AIC and BIC criteria across models, a lower value indicates better fit.

The log-likelihood, degrees of freedom, AIC and BIC for all four of the item response models fit to the Conceptions of Statistics data are presented in Table 3.1.
Table 3.1. Conceptions of Statistics Model Fit

<table>
<thead>
<tr>
<th>Model</th>
<th>Log-likelihood</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3.1) PCM</td>
<td>-7377.081</td>
<td>51</td>
<td>14856.16</td>
<td>15277.31</td>
</tr>
<tr>
<td>(3.2) Fixed Effect Pre-Post</td>
<td>-7443.463</td>
<td>52</td>
<td>14990.93</td>
<td>15420.33</td>
</tr>
<tr>
<td>(3.3) Random Effects Pre-Post</td>
<td>-7323.647</td>
<td>54</td>
<td>14755.29</td>
<td>15201.22</td>
</tr>
<tr>
<td>(3.4) Class-level Random Effects</td>
<td>-7101.742</td>
<td>57</td>
<td>14317.48</td>
<td>14788.18</td>
</tr>
</tbody>
</table>

The likelihood ratio test uses the log-likelihood and the degrees of freedom to compare nested models. Models are considered to be nested models when the more complex model can be transformed into the simpler model by imposing a set of constraints on the parameters. Here, Model 3.2 through Model 3.4 are nested models, but Model 3.1 is not due to the fact that it is specified to estimate the first level random effects as item responses nested in occasions while the others estimate the first level random effects as item responses nested in students (with covariates and additional levels of random effects). Thus, no comparisons can be made using the likelihood ratio test between the PCM in Model 3.1 and any of the other models. Using the likelihood ratio test to compare the differences in log-likelihood on Model 3.2 through Model 3.4, however, tells us that each time more parameters were added to the models, the difference in model fit was statistically significant ($p < 0.001$) when approximating the probability distribution using a chi-squared distribution. Thus, the results of the likelihood ratio test suggest that it is significantly better to fit a model with random effects for time than one with a fixed effect, and that it is significantly better to fit a model that includes classroom-level effects for ability and for change in ability over time than a model that does not account for classroom-level effects.

When comparing the models fit to the CoS data, we also look at the AIC and BIC criteria. The lowest values for both the AIC and the BIC criteria occurred when fitting the model that included classroom-level random effects. Thus, these additional criteria further suggest that Model 3.4 is the best fitting model. One interesting thing to note when examining the AIC and BIC criteria for the different models is that the highest values occurred for the model with the fixed effect for change in ability from pre-test to post-test. This implies that a model (Model 3.2) that artificially constrains differences over time to a fixed effect actually fits worse than a model that ignores repeated measures over time completely (Model 3.1).

The log-likelihood, degrees of freedom, AIC and BIC for all four of the item response models fit to the data from the Data Display construct are presented in Table 3.2.
Once again, when fitting the models for the DaD data, Model 3.2 through Model 3.4 are nested models, but Model 3.1 is not, and thus no comparisons can be made using the likelihood ratio test between the PCM in Model 3.1 and any of the other models. When using the likelihood ratio test to compare the differences in log-likelihood on Model 3.2 through Model 3.4, the difference in model fit was once again statistically significant ($p < 0.001$) each time more parameters were added to the models. Thus, for the DaD data, the results of the likelihood ratio test again suggests that the best fitting model of those fit here is Model 3.4, which includes classroom-level effects for ability and for change in ability over time than a model that does not account for classroom-level effects.

Similarly to the CoS data, when a more complex model (i.e. more parameters) was fit to the data the AIC value decreased. The only difference here is that the AIC value for the fixed effect model was less than the Partial Credit model AIC. This difference suggests that it is perhaps more reasonable to model a constant change in ability for the DaD construct than for the CoS construct. We will examine this further when looking at the parameter estimates from the models.

The BIC values for the DaD data also did not behave in the same fashion as the CoS data. Here, the BIC was not the same as when it was fit to the CoS data in that it slightly decreased from Model 3.1 to Model 3.2, and the lowest overall BIC value was for Model 3.3. Thus, when using the likelihood ratio test and the AIC as the methods of comparison Model 3.4 is considered the best fitting model, but when using the BIC as the method of comparison we have a contrary result – Model 3.3 is the best fitting model.

**Results – Conceptions of Statistics**

Now that we have fit all of the models to the data for the Conceptions of Statistics construct and compared the relative fit between the models, we now examine the estimated values of the parameters. By comparing the parameter values, we can see how the estimates of student ability, item difficulty, and the change in ability over time differ between the models. The estimated parameters of the models (not including the item step-difficulty estimates), and their standard errors, are presented in Table 3.3. Standard errors are not provided when reporting correlations.
Table 3.3 Conceptions of Statistics Estimated Parameter Values

<table>
<thead>
<tr>
<th>(3.1) PCM</th>
<th>(3.2) Fixed Effect Pre-Post</th>
<th>(3.3) Random Effects Pre-Post</th>
<th>(3.4) Class-level Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_1$</td>
<td>1.705 (0.119)</td>
<td>1.191 (0.092)</td>
<td>1.610 (0.145)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.652 (0.044)</td>
<td>0.742 (0.076)</td>
<td>0.556 (0.135)</td>
</tr>
<tr>
<td>$\psi_2$</td>
<td></td>
<td>1.505 (0.209)</td>
<td>0.863 (0.136)</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td>-0.506</td>
<td></td>
<td>-0.562</td>
</tr>
<tr>
<td>$\psi_3$</td>
<td></td>
<td></td>
<td>0.652 (0.186)</td>
</tr>
<tr>
<td>$\psi_4$</td>
<td></td>
<td></td>
<td>0.364 (0.123)</td>
</tr>
<tr>
<td>$\rho_2$</td>
<td></td>
<td></td>
<td>-0.385</td>
</tr>
</tbody>
</table>

As previously stated, the $\psi_1$ represents the variance of the first level random effect. We would expect to see this value decline as the model becomes more complex and attributes this variance to other effects. Here, this is basically what we have observed with the exception of the results from Model 3.2, the worst fitting model with a fixed effect for time.

All of the models that include changes over time also have estimates for the fixed effect $\beta_1$, the coefficient for time. For Models 3.2 and 3.3, these values represent the mean change in ability in logits over all the students, while for Model 3.4 it represents the mean change in ability in logits for students when controlling for the effect of change over time at the classroom-level. For all the models, the change in ability from pre-test to post-test is statistically significant at $p < 0.01$. In addition to examining whether the gains are statistically significant, we must also examine whether the effect size is large enough to claim that these are educationally significant gains. At the current stage of the ADMSR project, we have not yet concluded exactly what would entail meaningful effect sizes for the seven constructs, but we do know from experience that typical gains over a year in achievement tests are approximately half a logit. Thus, we can conclude that at this point the gains in ability on the CoS construct represent educationally important gains over the brief period of time the students were exposed to the AMDSR curriculum.

Table 3.3 also lists the correlations between the random effects at each level, with $\rho_1$ representing the correlation between the random effects at the student-level and $\rho_2$ representing the correlation between the random effects at the classroom-level. Both Model 3.3 and Model 3.4 had similar values for $\rho_1$ around -0.5, indicating a moderate to strong negative correlation between student’s abilities on the pre-test and their gains from pre-test to post-test. This negative correlation is not surprising in an educational sense - those students with lower abilities at the pre-test have a lot of room for improvement on the post-test, while students with high abilities might reach a “ceiling” for their ability which limits their gains. This “ceiling effect” can come about due to a number of factors, including possible deficiencies of the instrument. In the earlier chapter, we discussed how the measurement of the CoS construct was limited by having too few items at the top end...
of the construct (CoS level 4) in addition to the presence of few high performing students. Both of these factors likely contributed to the negative correlation between pre-test ability and change in ability over time. Similarly, the correlation at the classroom-level between pre-test ability and change in ability over time, \( \rho_2 \), also implied a moderately negative correlation.

One of the advantages of fitting a model that includes classroom-level effects is that it allows for comparisons to be made at that level. A scatter plot of the empirical Bayes predictions of classroom-level effects for pre-test ability and change in ability over time is presented in Figure 3.1. The classrooms in Figure 3.1 are ordered on the horizontal-axis by their predicted value for the random effect for pre-test ability \( \zeta^{(2)}_{3c} \), and the vertical-axis represents the predicted value for the random effect \( \zeta^{(2)}_{4c} \), the classroom-level change in ability over time.

---

Insert Figure 3.1 here

---

The negative correlation between the classroom-level random effects is apparent when looking at the predicted values of the classroom-level effects. Out of the 15 classrooms with negative effects for change over time, 12 of them have positive effects for pretest ability. For the classrooms with positive effects for changes over time, 7 of the 10 classrooms had negative effects for pre-test ability (i.e. were below the average classroom ability at the time of the pre-test).

The graph also clearly shows that classroom 0 had the greatest change in ability from pre-test to post-test. Classroom 0 is the only classroom from the Wisconsin site, while the other classrooms are from Arkansas. While this could be accounted for in part by the fact that the classroom 0 had one of the lower effects for pre-test ability, it invites further analysis as to why the Wisconsin classroom had a markedly greater improvement from pre-test to post-test than the Arkansas classrooms.

As mentioned earlier, Table 3.3 did not include all of the parameters of the models. Each model also included 50 fixed effects for the item-step difficulties. Since a goal of the ADMSR project is to create an item bank of items with invariant difficulties, we want to explore whether these estimates are significantly different if we apply a different model to the data. For the most part, the item-step difficulties did not significantly change as new models were applied. We will take a closer look at the differences between the estimates from the two best fitting models, Model 3.3 and Model 3.4. The scatterplot of the item-step difficulty estimates for Model 3.3 and Model 3.4 is shown in Figure 3.2.
In Figure 3.2, the 50 item-step difficulties are represented by the blue dots and the line represents the identity $y = x$ line. None of the item-step difficulty estimates in Figure 3.2 are significantly different. Three of the estimates, however, have been marked by orange squares as estimates with the greatest differences, ranging from 0.386 to 0.695 logits. Even though these are the largest differences in estimates, they are not significant because they are the most imprecise estimates (i.e. have the largest standard errors of measurement). These three estimates are the three step difficulties for the Battery1 item. Not surprisingly, this was one of the items identified in Chapter 2 as a problem item that should be considered for removal from the instrument. Given the variability of the item-step difficulty estimates here in comparison to the other items, this provides even more evidence to support not including this item in future administrations of the instrument.

In addition to comparing the item estimates, we also compared the person ability estimates between the two best fitting models. Although there were some differences, the person abilities produced by fitting Model 3.3 and Model 3.4 were very similar. The estimates between the two models were correlated at $\rho = 0.98$, and none of the estimates were statistically significantly different.

**Results – Data Display**

All four models were also fit to the Data Display construct. The estimated parameters of the models, not including the item step-difficulty estimates, and applicable standard errors are presented in Table 3.4.

<table>
<thead>
<tr>
<th></th>
<th>(3.1) PCM</th>
<th>(3.2) Fixed Effect Pre-Post</th>
<th>(3.3) Random Effects Pre-Post</th>
<th>(3.4) Class-level Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_1$</td>
<td>0.324 (0.021)</td>
<td>0.228 (0.016)</td>
<td>0.220 (0.024)</td>
<td>0.151 (0.018)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.336 (0.021)</td>
<td></td>
<td>0.343 (0.029)</td>
<td>0.265 (0.044)</td>
</tr>
<tr>
<td>$\psi_2$</td>
<td></td>
<td>0.241 (0.031)</td>
<td></td>
<td>0.137 (0.023)</td>
</tr>
<tr>
<td>$\rho_1$</td>
<td></td>
<td>-0.265</td>
<td></td>
<td>-0.322</td>
</tr>
<tr>
<td>$\psi_3$</td>
<td></td>
<td></td>
<td>0.233 (0.046)</td>
<td></td>
</tr>
<tr>
<td>$\psi_4$</td>
<td></td>
<td></td>
<td>0.129 (0.024)</td>
<td></td>
</tr>
<tr>
<td>$\rho_2$</td>
<td></td>
<td></td>
<td></td>
<td>-0.353</td>
</tr>
</tbody>
</table>
When examining the parameter values for the DaD construct, the first noticeable difference from the CoS construct is that the student-level variances, the $\psi_1$ values, are much smaller. The range of $\psi_1$ values for the CoS construct was from 0.726 to 1.705, while the range for the DaD construct is from 0.151 to 0.324. Thus, the student’s pre-test abilities on the DaD construct were much more consistent with each other than their abilities on the CoS construct. Another difference that we notice in the parameters from the DaD construct data is that the student-level variances monotonically decreased with the addition of parameters to the models.

The coefficient for time, $\beta_1$, which represents the change in ability from pre-test to post-test, is once again statistically significant at $p < 0.01$ for all of the models. The estimated values for these parameters, however, are less than the values for the CoS construct. Here, they are all well below the half a logit threshold mentioned earlier. Although the values are below this threshold, it is still possible that these gains are important gains. Given these low values, we must investigate further before we can conclude if the gains in ability on the DaD construct represent educationally important gains.

We now turn to the correlations listed in Table 3.4 for Model 3.3 and Model 3.4. Both models had similar values for $\rho_1$ close to -0.3, indicating a low to moderate negative correlation between student’s abilities on the pre-test and their gains from pre-test to post-test. Once again, we are not surprised at this negative correlation, but we note that this correlation is weaker than what existed for the CoS construct. This weaker correlation may be explained in part due to the fact that, unlike the CoS construct, the DaD construct had sufficient items measuring the highest levels, and thus did not suffer from as much of a ceiling effect as in the prior analysis. In addition, the correlation at the classroom-level between pre-test ability and change in ability over time, $\rho_2$, once more is similar to the student-level correlation with a low to moderate negative correlation of -0.353.

The empirical Bayes predictions for the classroom-level random effects for the DaD analysis when fitting Model 3.4 are presented in Figure 3.3. The classrooms in Figure 3.3 are ordered on the horizontal-axis by their value (in logits) for the predicted random effect for pre-test ability $\xi_{c}^{(2)}$, and the vertical-axis represents the value (in logits) for the predicted random effect $\xi_{c}^{(2)}$, the classroom-level change in ability over time.

---------------------------

Insert Figure 3.3 here

---------------------------
Figure 3.3 clearly shows that, once again, the classroom from the Wisconsin test site (classroom 0) had the greatest change in ability from pre-test to post-test. The empirical Bayes prediction for the random effect of pre-test ability for classroom 0 was just below the average effect for classrooms. With the smaller negative correlation coefficients for the DaD construct and the near average pre-test ability for classroom 0, it is evident that there is something other than pre-test ability that accounts for the Wisconsin classroom’s greater improvement from pre-test to post-test than the Arkansas classrooms.

Each model that was fit to the DaD data also included 83 fixed effects for the item-step difficulties. The scatterplot of the item-step difficulty estimates for Model 3.3 and Model 3.4 when fit to the DaD construct data is shown in Figure 3.4.

---

In Figure 3.4, the 83 item-step difficulties are represented by the blue dots and the line represents the identity y = x line. None of the item-step difficulty estimates in Figure 3.4 are significantly different. In fact, all of the estimates from Model 3.3 and Model 3.4 differ by less than 0.08 logits. This is good evidence that the item estimates are invariant, even if different models are applied to the data.

For the DaD construct, we evaluated the person ability estimates between the two best fitting models. Once again, there were some differences, but the student ability estimates generated when fitting Model 3.3 and Model 3.4 were very similar. The student ability estimates between the two models were correlated at $\rho = 0.96$, and none of the estimates were statistically significantly different.

Discussion and Conclusions

We have fit a variety of models to the Data Display and Conceptions of Statistics constructs from the ADMSR project. For the CoS construct, all criteria suggest that the best fitting model is one that takes into account classroom membership for both initial ability and change in ability over time. For the DaD construct, however, some criteria indicated that a model that includes random effects for ability and change in ability over time is sufficient without having to take into account classroom-level effects.

When choosing a model, the purpose of the modeling and practical considerations must also be taken into account. If the purpose of the modeling is to create a set of usable items or to only estimate the ability of students, then including classroom-level effects may not be needed. As we have shown above, the results of the item estimates and
student abilities were not statistically significantly different when we fit the final two models. Thus, *Model 3.3* would be sufficient for these purposes. *Model 3.3* also has practical advantages over *Model 3.4*. It is more straightforward to interpret the results from *Model 3.3* as compared to *Model 3.4*, and *Model 3.3*’s relative simplicity requires less computational power to fit the model. For the models fit here to CoS and DaD, *Model 3.3* required 1 day and 2 days to run for each construct, and *Model 3.4* required 9 days and 21 days to run for each construct, respectively.

If the purpose of the analysis is to analyze the differences between classrooms, however, then it would be advantageous to apply a multi-level IRT model such as *Model 3.4*. By fitting this model, we were able to identify how changes in ability over time were affected at the classroom-level. The ADMSR project, like most educational interventions, is implemented at the classroom level, and thus one would expect variation to naturally occur at this level based on the different instruction received by the students. Identifying this variation is only the first step and can lead to the discovery of why this variation occurs, which can ultimately improve the implementation of the ADMSR curricula. One possible hypothesis of why some classes have greater changes in ability is variation in the levels of fidelity of the implementation of the classroom instruction to the curricula. Currently, the ADMSR project is designing an instrument to measure the degree of fidelity of the classroom instruction, which when used in conjunction with the multi-level IRT model used here will hopefully explain some of this classroom-level variance.
Figure 3.1 – Scatterplot of Conceptions of Statistics Classroom-Level Effects
Figure 3.2 – Scatterplot of Conceptions of Statistics Step Difficulty Estimates
Figure 3.3 – Scatterplot of Data Display Classroom-Level Effects
Figure 3.4 – Scatterplot of Data Display Step Difficulty Estimate
Chapter 4
Analyzing the Learning Progression Using a Multidimensional Item Response Model

Introduction

The theoretical framework of the seven ADMSR constructs outlined in chapter 1, and presented again here in Figure 4.1, implies 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 4.1 illustrate, this construct relies on an orchestration of the components of the other data modeling constructs. The arrows in Figure 4.1 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 opposing construct, and vice-versa, but where there is no theorized one-way causal connection.

In order to model the seven constructs together, we must 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, & Wang, 1997; Briggs & 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 specific relationships between the individual levels of the constructs, as outlined by the ADMSR theoretical framework in Figure 4.1, can be examined. In this chapter, we 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.

*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 on the logistic model, it is preferred 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, & 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, & Haldane, 2007). The multidimensional formulation of the RCML is the Multidimensional Random Coefficients Multinomial Logit Model (MRCML; Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003).

The MRCML formulates a generalized solution for a family of multidimensional, polytomous Rasch-based models. The general MRCML formulation for the probability of
a response in category $k$ of item $i$ is:

$$
\text{Pr}(X_{ik} = 1; A, B, \xi | \theta) = \frac{\exp(b_k \theta - a_{ik} \xi)}{\sum_{k=1}^{K} \exp(b_k \theta - a_{ik} \xi)}
$$

(Adams, Wilson, & Wang, 1997), where $\xi = (\xi_1, ..., \xi_n)$ is a vector of the item and step parameters, $A$ is the design matrix, $B$ is the scoring matrix, and $\theta$ is the vector of abilities on each dimension. The design matrix, $A$, represents the relationships between item-category combinations (rows $a'_{ik}$) and the model parameters (the columns), the item difficulties and step difficulties. The scoring matrix, $B$, represents the relationships between items (the rows $b_{ik}$) and dimensions (the columns) (Kennedy, 2005).

A vector of ability parameters represents a person’s ability in the MRCML model, with a unique ability parameter for each dimension. The MRCML therefore allows for multiple abilities for each person. When using the MRCML model on the data from the ADMSR project, each student can have up to seven different predicted abilities, one for each of the seven constructs.

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 on several unidimensional subscales, while a within-item multidimensional test also includes items that relate to more than one of the latent 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 appropriate. 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. The simpler form of the Multidimensional Partial Credit Model (i.e. the between-item version) assumes that for each item $i$, with ordered categories of $^1$\

\[^1\] 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 results as a whole when examining total model fit, and the parameter estimates for the affected items (those scored on more than one construct) were statistically significantly different as well.
response indexed by \( j (j = 1, \ldots, J) \), 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 \((\theta_{pd})\), and the relative difficulty of category \( j \) \((\delta_{ij})\) or the step difficulty:

\[
\ln \frac{\Pr(x_{ip} = j | \theta_{pd})}{\Pr(x_{ip} = j-1 | \theta_{pd})} = \theta_{pd} - \delta_{ij}
\]

(4.2)

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 \( \theta = (\theta_1, \ldots, \theta_d) \). The mean of the step difficulties \((\delta_{ij})\) for an item \( i \) is an item’s overall item difficulty \( \delta_i \). Thus, each \( \delta_{ij} \) can also be formulated as \( \delta_i + \tau_{ij} \), where \( \tau_{ij} \) is the deviation from the mean item difficulty \( \delta_i \) for item \( i \) at step \( j \). Formulated this way, the last tau parameter for each item must equal to the negative sum of the others so that the sum of all the tau 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 that case, the equivalent of Equation 4.2 would indicate a compensatory relationship of the ability on the respective dimension (See Adams, Wilson, & Wang, 1997). In the analysis here we use only the between-item version of the model.

**Sample and Assessment Data**

For the analysis presented here, the sample of students was the same as those described in Chapter 2. Once again, the data presented here is exclusively from the post-tests. The post-tests were administered to 1002 students, who were all exposed to the Data Modeling curriculum. A total of seven different test forms were used. Each test form contained 20 multi-part items that covered at least four of the seven constructs. A total of 53 multi-part items, with 110 individually scored parts, were administered. The distribution of these items across the seven ADMSR constructs is presented in Table 4.1.
Table 4.1. *ADMSR Post-test Items*

<table>
<thead>
<tr>
<th>Construct</th>
<th>Multi-Part Items</th>
<th>Individually Scored Parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory of Measurement (ToM)</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Data Display (DaD)</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>Meta-Representational Competence (MRC)</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Conceptions of Statistics (CoS)</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Chance (Cha)</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>Models of Variability (MoV)</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Informal Inference (InI)</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

*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 deviances, and number of estimated parameters from the two models are given in Table 4.2.

Table 4.2. *Relative Model Fit between Unidimensional and Seven-dimensional*

<table>
<thead>
<tr>
<th></th>
<th>Deviance</th>
<th>Estimated Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidimensional</td>
<td>89435.54</td>
<td>363</td>
</tr>
<tr>
<td>Seven-dimensional</td>
<td>88236.23</td>
<td>390</td>
</tr>
<tr>
<td>Difference</td>
<td>1199.31</td>
<td>27</td>
</tr>
</tbody>
</table>

The deviance of the unidimensional analysis was 89,435.54, with a total of 363 estimated parameters. The deviance of the seven-dimensional analysis was 88,236.23, with a total of 390 estimated parameters. 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 cannot 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 ration test by two, and get a 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 significant difference corresponds to an important effect difference. To explore this, we looked at the correlations between the constructs/dimensions that were obtained from ConQuest when running the seven-dimensional MRCML model estimation. A matrix of these results is presented in Table 4.3, with the correlations displayed on the bottom left of the diagonal.

A matrix of the correlation estimates of the 7 dimensions is presented in Table 4.3.

Table 4.3. Correlations of the Seven Constructs/Dimensions

<table>
<thead>
<tr>
<th>Construct / Dimension</th>
<th>DaD</th>
<th>MRC</th>
<th>CoS</th>
<th>Cha</th>
<th>MoV</th>
<th>InI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRC</td>
<td>0.832</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoS</td>
<td>0.785</td>
<td>0.872</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cha</td>
<td>0.783</td>
<td>0.843</td>
<td>0.851</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MoV</td>
<td>0.807</td>
<td>0.917</td>
<td>0.893</td>
<td>0.902</td>
<td></td>
<td></td>
</tr>
<tr>
<td>InI</td>
<td>0.902</td>
<td>0.886</td>
<td>0.876</td>
<td>0.873</td>
<td>0.935</td>
<td></td>
</tr>
<tr>
<td>ToM</td>
<td>0.778</td>
<td>0.811</td>
<td>0.818</td>
<td>0.806</td>
<td>0.847</td>
<td>0.811</td>
</tr>
</tbody>
</table>

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 4.3 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. 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.

Similar to the analysis of unidimensional results, when analyzing the results of the multi-dimensional model we examine the Thurstonian thresholds instead of only looking at the item and step difficulties. The Thurstonian threshold is the location 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 = 0.5 line. These values tend to be more interpretable than the $\delta_i$ and $\tau_{ik}$ values because they identify levels where students are most likely to achieve specific scores (Kennedy, 2005). A seven-dimensional Wright Map for the ADMSR data is presented in Figure 4.2.

Insert Figure 4.2 here

Figure 4.2 consists of seven student ability distributions in the left columns and the seven dimensions of item threshold estimates are on the right side. Similar to the Wright Maps presented in Chapter 2, the thresholds in each dimension in Figure 4.2 are ordered by level so that each column shows all of the item thresholds for a given level on the construct map. Figure 4.2 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 4.2 is limited because it does not allow for comparisons of the ability distributions or the item threshold estimates across the dimensions.

**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. 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. This is necessary for identification purposes. The problem with this assumption in our case 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 examine whether certain levels of the construct maps are horizontally associated across the learning progression, as hypothesized, we need to use a technique to align the constructs/dimensions. One such alignment technique is Delta Dimensional Alignment (DDA; Schwartz & 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 ($\mu$) and standard deviations ($\sigma$) of the subsets of the items for each dimension ($d$), which are calculated from a unidimensional analysis (i.e.,

\[^2\text{Equivalently, the mean of item difficulties can be set to zero for every dimension.}\]
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 ($\mu_{uni}$) and standard deviation ($\sigma_{uni}$) 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 ($\sigma_{multi}$) 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:**

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

(4.3)

**Step parameters:**

$$\tau_{ikd\,(\text{transformed})} = \tau_{ikd\,(\text{multi})} \left( \frac{\sigma_{d\,(\text{uni})}}{\sigma_{d\,(\text{multi})}} \right)$$

(4.4)

Note that these transformations are for the $\delta_i$ and $\tau_{ik}$ 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.
Simulation of Delta Dimensional Alignment Technique

Before applying DDA to the data from the ADMSR project, we ran a simulation to see how well DDA aligns the dimensions under a variety of distributions for both the person ability and item difficulty parameters. We generated three-dimensional data sets using the Conquest software for each replication of the simulation. The generated data sets consisted of five items on the first dimension, and four items on the second and third dimensions. Every item was a polytomous item with four possible levels of response (0, 1, 2, or 3), and responses were generated for 1000 students of randomly distributed ability. The generation was replicated 30 times to produce 30 data sets. After the data was generated, each replication was analyzed using the three steps of the DDA method. There was, however, one important difference between the data as generated by Conquest and the data that was subsequently analyzed for each replication of the simulation. For each replication, the item response pattern for the last item on the first dimension (item 5) was duplicated to become the response pattern for the first item on the second dimension (item 6). The same was done for the last item on the second dimension (item 10) and the first item on the third dimension (item 11). This provided item response data for five items on each of the three dimensions, for a total of 15 items. Because the data was generated on multiple dimensions, and there exists no underlying theory as to how the dimensions are related, we created these links between the dimensions that can be examined after the DDA technique has been applied. If the DDA technique successfully aligns the dimensions, the items with the same response patterns should have equivalent step difficulty estimates after they are aligned on a single scale.

Table 4.4. Threshold Estimates Before Alignment for 30 Replications

<table>
<thead>
<tr>
<th>Item Step</th>
<th>Mean Difficulty Estimate</th>
<th>Difference of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 5, step 1</td>
<td>-0.777</td>
<td>1.270</td>
</tr>
<tr>
<td>Item 6, step 1</td>
<td>0.493</td>
<td></td>
</tr>
<tr>
<td>Item 5, step 2</td>
<td>-0.189</td>
<td>1.231</td>
</tr>
<tr>
<td>Item 6, step 2</td>
<td>1.042</td>
<td></td>
</tr>
<tr>
<td>Item 5, step 3</td>
<td>0.490</td>
<td></td>
</tr>
<tr>
<td>Item 6, step 3</td>
<td>1.678</td>
<td>1.188</td>
</tr>
<tr>
<td>Item 10, step 1</td>
<td>-0.871</td>
<td>1.418</td>
</tr>
<tr>
<td>Item 11, step 1</td>
<td>0.547</td>
<td></td>
</tr>
<tr>
<td>Item 10, step 2</td>
<td>-0.258</td>
<td>1.415</td>
</tr>
<tr>
<td>Item 11, step 2</td>
<td>1.157</td>
<td></td>
</tr>
<tr>
<td>Item 10, step 3</td>
<td>0.387</td>
<td>1.414</td>
</tr>
<tr>
<td>Item 11, step 3</td>
<td>1.801</td>
<td></td>
</tr>
</tbody>
</table>

The MRCML results from the 30 replications of the simulation, without applying the DDA technique, are shown in Table 4.4. As expected, there were clear differences between the threshold estimates for all steps in both sets of items. These differences are
attributable to the differences in the estimated person ability distributions for the different dimensions. Thus, even though the items have the same response pattern, different item step parameters are estimated.

Table 4.5. Threshold Estimates After DDA Technique for 30 Replications

<table>
<thead>
<tr>
<th>Item Step</th>
<th>Mean Difficulty Estimate</th>
<th>Difference of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 5, step 1</td>
<td>0.451</td>
<td>0.037</td>
</tr>
<tr>
<td>Item 6, step 1</td>
<td>0.488</td>
<td></td>
</tr>
<tr>
<td>Item 5, step 2</td>
<td>1.038</td>
<td>0.000</td>
</tr>
<tr>
<td>Item 6, step 2</td>
<td>1.038</td>
<td></td>
</tr>
<tr>
<td>Item 5, step 3</td>
<td>1.707</td>
<td>-0.038</td>
</tr>
<tr>
<td>Item 6, step 3</td>
<td>1.669</td>
<td></td>
</tr>
<tr>
<td>Item 10, step 1</td>
<td>-0.787</td>
<td>0.008</td>
</tr>
<tr>
<td>Item 11, step 1</td>
<td>-0.779</td>
<td></td>
</tr>
<tr>
<td>Item 10, step 2</td>
<td>-0.175</td>
<td>0.004</td>
</tr>
<tr>
<td>Item 11, step 2</td>
<td>-0.171</td>
<td></td>
</tr>
<tr>
<td>Item 10, step 3</td>
<td>0.469</td>
<td>0.000</td>
</tr>
<tr>
<td>Item 11, step 3</td>
<td>0.469</td>
<td></td>
</tr>
</tbody>
</table>

The MRCML results from the same 30 replications of the simulation after the application of the DDA method are shown in Table 4.5. By aligning the dimensions using the DDA technique, the differences between the threshold estimates have been reduced to values that are so small as to be below the level that they would be interpreted as significantly different. For the 30 replications, all six of the threshold comparisons for the two sets of items had a mean of the differences in the threshold estimates of less than 0.04.

A scatter plot of the estimates for items 5 and 6 prior to any alignment is given in Figure 4.3. A scatter plot for items 5 and 6 after the application of the DDA method is given in Figure 4.4. Two more scatter plots, for the pre-DDA and the post-DDA estimates for items 10 and 11, are given in Figure 4.5 and Figure 4.6. The estimates for both sets of comparisons after the application of the Delta Dimensional Alignment technique fall firmly of the identity line of $y = x$, indicating the consistency of the low differences between estimates for all 30 of the replications, regardless of the actual value of the threshold estimate. Figures 4.3 and 4.5 show that even though the estimated values of the thresholds varied between each replication (due to the random generation of item difficulties and student abilities), these estimates changed in unison between the items with the same response patterns.
Insert Figures 4.3 through 4.6 here

In this simulation, the DDA technique was successful in taking items from different dimensions, but with the same expected level of difficulty if they were placed on a single scale, and estimating them consistently together on that scale. Our results from the simulation give us confidence to apply the DDA technique to align the dimensions on the ADMSR data.

Examing 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 4.7 shows the results of the aligned multidimensional Wright Map with the thresholds of all 110 items, representing all seven dimensions.

Insert Figure 4.7 here

Similar to Figure 4.2, Figure 4.7 is ordered by level for each dimension and shows the rising threshold estimates across the dimension. Using the Wright Map in Figure 4.7, 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 4.1. The evidence supporting one 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.

Starting from the left hand side of Figure 4.1, 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 4.1 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. 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 for items that come from the same multi-part task called “Ruler.” After 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 4.8.

---------------------------------
Insert Figure 4.8 here
----------------------------------

Figure 4.8 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.

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. As described previously in Chapter 2, 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 supports 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 4.9 contains a close-up of these constructs on the Wright map, with the levels of the theoretical link highlighted.

By examining Figure 4.9 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 4.1 between MRC3-5 and DaD4-5 represents the 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 4.10.

---------------------------------
Insert Figure 4.10 here
----------------------------------

Figure 4.10 shows the item threshold estimates for DaD and MRC, and it includes a couple of red boxes that indicate the estimates that are a part of the bootstrapping section of Figure 4.1. 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 between 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 4.1 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. As described earlier in Chapter 2, 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 with 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 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 4.11.
While there exists a little bit of overlap between the threshold estimates, looking at Figure 4.11 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 analysis, we can examine the remaining theorized connections of the ADMSR learning progression represented in Figure 4.1. The results discussed above and these remaining results are summarized in Table 4.6. 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 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.
<table>
<thead>
<tr>
<th>Hypothesized Link</th>
<th>Observation</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ToM4 to DaD1</strong></td>
<td>Partially supported by data. When Ruler items are removed, data supports that ToM4 has similar difficulty as DaD1.</td>
<td>Examine Ruler items. Retain theorized connection</td>
</tr>
<tr>
<td><strong>ToM4 to MoV1</strong></td>
<td>Supported by data. When Ruler items are removed, data supports that ToM4 precedes MoV1</td>
<td>Examine Ruler items. Retain theorized connection</td>
</tr>
<tr>
<td><strong>ToM6 to CoS3</strong></td>
<td>ToM6 thresholds are less than ToM4 and ToM5. Data supports that ToM4-6 all precede CoS3.</td>
<td>Retain theorized connection</td>
</tr>
<tr>
<td><strong>ToM6 to MoV2</strong></td>
<td>ToM6 thresholds are less than ToM4 and ToM5. Data supports that ToM4-6 all precede MoV2</td>
<td>Retain theorized connection</td>
</tr>
<tr>
<td><strong>DaD4 to CoS3</strong></td>
<td>Partially supported by data. Similar threshold estimates for DaD4 and CoS3.</td>
<td>Retain theorized connection</td>
</tr>
<tr>
<td><strong>DaD5 to Cha4</strong></td>
<td>Data does not support connection. DaD5 appears more difficult than Cha4, but only two Cha4 thresholds.</td>
<td>Test on another sample to gather more Cha4 data.</td>
</tr>
<tr>
<td><strong>CoS3 to Cha3</strong></td>
<td>Data does not support connection. CoS3 is more difficult than Cha3.</td>
<td>Test on another sample before removing connection</td>
</tr>
<tr>
<td><strong>CoS3 to InI3</strong></td>
<td>Data does not support connection. CoS3 is more difficult than InI3.</td>
<td>Test on another sample before removing connection</td>
</tr>
<tr>
<td><strong>CoS4 to InI7</strong></td>
<td>Only one threshold estimate for InI7 and two estimates for CoS4. Not enough data for InI7 in sample.</td>
<td>Test higher performing sample to gather CoS4 and InI7 data.</td>
</tr>
<tr>
<td><strong>Cha6 to CoS4</strong></td>
<td>Only two CoS4 threshold estimates. Data supports connection, but not enough data to make conclusion.</td>
<td>Retain theorized connection. Examine more CoS4 items.</td>
</tr>
<tr>
<td><strong>Cha6 to InI5</strong></td>
<td>Data supports that Cha6 precedes InI5.</td>
<td>Retain theorized connection.</td>
</tr>
<tr>
<td><strong>MoV5 to InI5</strong></td>
<td>Only one threshold estimates for MoV5 in sample.</td>
<td>Test higher performing sample to gather MoV5 data.</td>
</tr>
</tbody>
</table>
Conclusions and Future Work

The development of the ADMSR learning progression relies on both unidimensional and multidimensional analysis of its seven constructs. In Chapter 2, we looked at the constructs individually, and the results from the Partial Credit analysis lead to refinements of the constructs, the scoring exemplars, and to the items. In Chapter 3, we continued to look at the constructs individually, but turned our attention to determining whether or not the students exposed to the ADMSR curriculum had gains in ability over the time of the intervention. Here, in Chapter 4, we shifted our focus to the learning progression as a whole and analyzed the seven ADMSR constructs together.

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 4.1. 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 after the results of only the samples examined herein. 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.
Figure 4.1 – A Map of the Theoretical ADMSR Learning Progression
Figure 4.2 – Seven Dimensional Wright Map of the ADMSR Learning Progression
Figure 4.3 – Threshold Estimates for Items 5 and 6 with No Alignment
Figure 4.4 – Threshold Estimates for Items 5 and 6 with Delta Dimensional Alignment
Figure 4.5 – Threshold Estimates for Items 10 and 11 with No Alignment
Figure 4.6 – Threshold Estimates for Items 10 and 11 with Delta Dimensional Alignment
Figure 4.7 – Aligned Wright Map of the ADMSR Learning Progression Using DDA
Figure 4.8 – Link Between ToM4 with DaD1 and MoV1
Figure 4.9 – Link Between ToM6 with CoS3 and MoV2
Figure 4.10 – Bootstrapping Between DaD and MRC
Figure 4.11 – Link Between Cha6 with CoS4
References


StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.


Appendix – Chapter 3 Stata/GLLAMM Commands

CoS Commands for Running Models 3.1 to 3.4

***Change names of items
rename battery item1
rename caffeine item2
rename corn item3
rename ballmedian item4
rename ballmode item5
rename ballmean item6
rename ball2 item7
rename ball3 item8
rename freethrow item9
rename range item10
rename max4mc item11
rename solarpanelmc item12
rename tallest1mc item13
rename tallest2mc item14
rename kayla item15
rename swimming1 item16
rename swimming2 item17
rename swimming3 item18
rename ten2 item19

***drop forms with too few items on CoS
drop if form==3
drop if form==5

***generate occasion id
gen occ_id = _n

***reshape data from wide to long form
reshape long item, i(occ_id) j(it)
drop if item==.

gen obs=_n

*** expand the items based on number of categories
expand 4 if it==1
expand 4 if it==2
expand 5 if it==3
expand 3 if it==4
expand 3 if it==5
expand 3 if it==6
expand 3 if it==7
expand 3 if it==8
expand 5 if it==9
expand 6 if it==10
expand 3 if it==11
expand 3 if it==12
expand 3 if it==13
expand 3 if it==14
expand 4 if it==15
expand 3 if it==16
expand 4 if it==17
expand 3 if it==18
expand 4 if it==19

***sort
sort occ_id item obs

***determine each possible score for each item, specify which category each ***person actually selected, and create item dummies by obs, sort: gen x = _n-1
gen chosen = item == x
tab it, gen(q)

***create variables for item steps corresponding to design matrix forvalues i=1/19{
  forvalues g=1/5{
    gen d`i'`g' = -1*q`i'*(x>=`g')
  }
  if `i'!=10 drop d`i'_5
  if `i'!=10 & `i'!=9 & `i'!=3 drop d`i'_4
  if `i'!=10 & `i'!=9 & `i'!=3 & `i'!=1 & `i'!=2 & `i'!=15 & `i'!=17 &
`i'!=19 drop d`i'_3
}

***PCM (3.1)
eq slope: x
gllamm x d1_1-d22_5, i(occ_id) eqs(slope) link(mlogit) expand(obs chosen o) adapt trace nocons
estimates store pcm

gllapred prob1, mu us(trait)
gllapred theta, u
estimates stat pcm

***FIXED EFFECT for prepost (3.2)
eq slope: x
eq f1: prepost
gllamm x d1_1-d19_3, i(id) eqs(slope) geqs(f1) link(mlogit) expand(obs chosen o) adapt trace nocons
estimates store timefixed

gllapred re_tf, u
gllapred theta_tf, fac

estimates stat timefixed

***RANDOM EFFECTS for prepost (3.3)
gen x_t = prepost * x
eq slope: x
eq slope2: x_t
eq f1: prepost
glamm x dl_1-dl9_3, i(id) nrf(2) eqs(slope slope2) geqs(f1)
link(mlogit) expand(obs chosen o) adapt trace nocons nip(4)
estimates store random

gllapred re_TR, u
gllapred theta_TR, fac

estimates stat random

***Classlevel Analysis (3.4)
gen x_t = prepost * x
eq slope: x
eq slope2: x_t
eq f1: prepost
glamm x dl_1-dl9_3, i(id teacher) nrf(2,2) eqs(slope slope2 slope slope2) geqs(f1) link(mlogit) expand(obs chosen o) adapt trace nocons nip(4)
estimates store classlevel

gllapred re_cl, u
gllapred theta_cl, fac

estimates stat classlevel
DaD Commands for Running Models 3.1 to 3.4

*** Change names of items
rename gottago1 item1
rename gottago2 item2
rename gottago3 item3
rename applause item4
rename bowling1 item5
rename bowlingq2 item6
rename crab1 item7
rename crab2 item8
rename crab3 item9
rename earthquake item10
rename headache item11
rename cherrytree item12
rename max5 item13
rename rocket1 item14
rename rocket2 item15
rename candle1 item16
rename candle2 item17
rename statecap1 item18
rename statecap2 item19
rename statue4 item20
rename statue6 item21
rename statue8 item22

*** Drop forms with too few DaD items
drop if form==2 & prepost==0
drop if form==5 & prepost==0

*** Generate occasion id
gen occ_id = _n

*** Reshape data from wide to long form
reshape long item, i(occ_id) j(it)
drop if item==.

gen obs=_n

*** Expand the items based on number of categories
expand 3 if it==1
expand 3 if it==2
expand 3 if it==3
expand 5 if it==4
expand 5 if it==5
expand 6 if it==6
expand 3 if it==7
expand 6 if it==8
expand 3 if it==9
expand 6 if it==10
expand 6 if it==11
expand 5 if it==12
expand 3 if it==13
expand 3 if it==14
expand 6 if it==15
expand 6 if it==16
expand 6 if it==17
expand 3 if it==18
expand 6 if it==19
expand 6 if it==20
expand 6 if it==21
expand 6 if it==22

*sort
sort obs it

***Identify each possible score for each item, specify which category each person actually selected, and create item dummies by obs, sort: gen x = _n-1
gen chosen = item == x
tab it, gen(q)

***create variables for item steps corresponding to design matrix
forvalues i=1/22 {
    forvalues g=1/5 {
        gen d`i'_`g' = -1*q`i'*(x>=`g')
    }
    if `i'!=6 & `i'!=8 & `i'!=10 & `i'!=11 & `i'!=15 & `i'!=16 & `i'!=17 & `i'!=19 & `i'!=20 & `i'!=21 & `i'!=22 drop d`i'_5
    if `i'!=6 & `i'!=8 & `i'!=10 & `i'!=11 & `i'!=15 & `i'!=16 & `i'!=17 & `i'!=19 & `i'!=20 & `i'!=21 & `i'!=22 & `i'!=4 & `i'!=5 & `i'!=12 drop d`i'_4
    if `i'!=6 & `i'!=8 & `i'!=10 & `i'!=11 & `i'!=15 & `i'!=16 & `i'!=17 & `i'!=19 & `i'!=20 & `i'!=21 & `i'!=4 & `i'!=5 & `i'!=12 drop d`i'_3
}

***PCM (3.1)
eq slope: x
gllamm x d1_1-d22_5, i(occ_id) eqs(slope) link(mlogit) expand(obs chosen o) adapt trace nocons
estimates store pcm
gllapred prob1, mu us(trait)
gllapred theta, u
estimates stat pcm
***FIXED EFFECT for prepost (3.2)
eq \text{slope}: x
\text{eq f1: prepost}

gllamm x d1_1-d22_5, i(id) eqs(slope) geqs(f1) link(mlogit) expand(obs chosen o) adapt trace nocons

estimates store timefixed

gllapred re_tf, u
gllapred theta_tf, fac

estimates stat timefixed

***RANDOM EFFECTS for prepost (3.3)
gen x_t = \text{prepost} * x
\text{eq slope}: x
\text{eq slope2: x_t}
\text{eq f1: prepost}

gllamm x d1_1-d22_5, i(id) nrf(2) eqs(slope slope2) geqs(f1) link(mlogit) expand(obs chosen o) adapt trace nocons nip(4)

estimates store random

gllapred re_TR, u
gllapred theta_TR, fac

estimates stat random

***Classlevel Analysis (3.4)
gen x_t = \text{prepost} * x
\text{eq slope}: x
\text{eq slope2: x_t}
\text{eq f1: prepost}

gllamm x d1_1-d22_5, i(id teacher) nrf(2,2) eqs(slope slope2 slope slope2) geqs(f1) link(mlogit) expand(obs chosen o) adapt trace nocons nip(4)

estimates store classlevel

gllapred re_cl, u
gllapred theta_cl, fac

estimates stat classlevel