A Construct Modeling Approach to Measuring Fidelity in Data Modeling Classes

By

Ryan Seth Jones

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

For the degree of

DOCTOR OF PHILOSOPHY

In

Learning, Teaching, and Diversity

May 2015

Nashville, Tennessee

Approved:

Richard Lehrer, Ph.D.
Ilana Horn, Ph.D.
Mark Lipsey, Ph.D.
Leona Schauble, Ph.D.
Mark Wilson, Ph.D.

For Jennifer.
Because you encourage me to be more than I am.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the efforts, wisdom, and generosity of many people. It is impossible to mention everyone, but I must name a few.

The people that make up the department of Teaching, Learning, and Diversity at Vanderbilt University have listened to, critiqued, shaped, and informed my thinking. Shelly Cotterman, Amy Holmes, Min-Joung Kim, Marta Kobiela, Eve Manz, Erin, Pfaff, Mayumi Shinohara, and Rob Rouse have generously read manuscripts, listened to talks, and helped me make sense of my data. My dissertation committee, Lani Horn, Mark Lipsey, Leona Schauble, and Mark Wilson, have patiently helped me frame my reading and writing about fidelity. And most importantly, my advisor Rich Lehrer has shaped my thinking, fed me, encouraged me, challenged me, and generously offered innovative ideas while letting me run with them. I cannot state enough how thankful I am for his gracious and wise guidance.

This work would not have been possible without my family's many sacrifices. Jacob Jones reminded me more than once that the data I create should be information to answer a question. Andrew Jones reminded me not to worry about the approval of my peers because he faithfully told me he was proud of me. Eleanor Jones has been a constant reminder that there are many things more important than my research. And my life partner, Jennifer Jones, has given of herself for my good for over a decade of my life. This is as much her accomplishment as it is mine.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A110685 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

A MEASUREMENT PERSPECTIVE ON FIDELITY

Introduction

Research that relates human learning and the conditions under which it develops is central to education. Collecting evidence of these relations is always a challenge, but is particularly difficult in large-scale studies of student learning in designed learning environments due to large numbers of participants, contextual variability across classrooms, and limited resources. While particular research designs, such as random assignment, can protect against many confounding factors that influence student learning, they do not ensure that the realized learning environments faithfully reflect the intentions of the design (Lipsey, 1993). If the designed learning environment relies on changes in practices, beliefs, perspectives, routines, or behaviors of participants, then a valid account of these changes is needed to make inferences about the relationship between them and student thinking.

At a smaller scale, design researchers make inferences about this relationship by looking for correspondences between particular qualities of classroom activity and evidence of student thinking manifested in discourse, action, or inscriptions. This painstaking work usually relies on video and/or audio recordings of classroom activity and interviews, student artifacts, and field notes to create detailed characterizations of the realized learning environments. Although these data records can include many units of time (days, weeks, months, or even years) they typically stretch across a much smaller number of participant units to make the analyses feasible. This methodology is central to design research because it allows researchers to test a particular

operationalization of their instructional theories, provides evidence about the relationship between researchers' instructional theories and learning theories, and can even produce new theoretical frameworks to model relations between instruction and learning (Cobb, Confrey, DiSessa, Lehrer, & Schauble, 2003; DiSessa & Cobb, 2009; Sandoval, 2014).

However, these methods are not helpful for questions about relations between instruction and learning when a designed instructional program is used on a large scale. Detailed qualitative analyses are much too time consuming for a large number of classrooms. In addition, qualitative methods typically relate instructional experiences with learning for a number of strategically selected cases. In large-scale studies, though, questions about aggregate relationships require measures of student learning and instructional contexts that can be used across a large number of participants and that are grounded in the theories developed through the qualitative studies.

It's not only pragmatics and limited resources that motivate the need for scaled measures, though. Even if infinite resources were available to qualitatively study these relations across large numbers of classes, there is additional value in going beyond counts or proportions of qualities to consider their measurement scale. A measurement scale uses the metaphor of linear space to represent the relative "distance" between qualitatively different categories. For example, a qualitative analysis could answer many questions about the difference between frozen water and boiling water. It could even help researchers understand qualitatively different states of water through narrative descriptions of ice, liquid water, and vapor. However, the qualitative analyses don't help to answer "how much" questions. Sure, boiling water is hotter than frozen water, but how much hotter? For this question the qualities of the theories about water must be interpreted as quantities in a measurement scale. This logic holds true for measures of classroom instruction as well. When qualitatively different categories of instruction are used to create a

measurement scale we can then ask "how much more different?" The answer to this question will likely have theoretical implications for our understanding of the phenomenon, and also practical consequences for the strategies we deploy to create change, such as professional development.

In evaluation research, the measures of classroom instruction are often referred to as "fidelity" measures because the different numerical scales are typically interpreted in reference to researchers' idealized instructional theories. Classrooms with higher scores on the scale indicate realized learning environments that are more faithful to the intent of the designed program than classrooms that are lower on the scale. Recent reviews agree that there has historically been little consistency in conceptual frameworks, methods, or tools for this type of measurement (e.g. O'Donnell, 2008). In response, there is now a growing consensus on important steps researchers should take when developing a fidelity measure. Since it is clearly important for a fidelity measure to account for a program's most important tasks and materials, many of these steps emphasize the widespread acknowledgement that essential program components should be explicitly represented in the measure. Often ignored, though, is that a list of visible program components is but a particular operationalization of an underlying instructional theory. This theory could be operationalized differently, or the visible components might be used in a manner that does not manifest the underlying instructional theories. With this in mind, fidelity measures, like any measures, should also articulate relationships between the visible components to be observed and the underlying theories that motivated attention to them.

This dissertation is my attempt to contribute to this challenge in the context of a large-scale study of an approach to learning called Data Modeling. Data Modeling is designed to support middle school students in developing competencies in a related set of practices including

posing a question, deciding upon appropriate attributes and measures that address the question, designing investigations, structuring resulting data as distribution, developing statistics to measure distributional characteristics, and inventing and revising models of chance to guide inference (Lehrer & Romberg, 1996; Lehrer, Schauble, & Kim, 2007). For STEM professionals these practices serve an epistemic purpose since they are the means by which disciplines construct meaning from data. Since it is our goal for students to use these ideas to construct meaning from data for themselves we refer to them as epistemic practices. In order to study this approach to learning it is critical to measure the extent to which students are supported during instruction to participate in these epistemic practices. This need motivates the following questions in this dissertation:

Research Question 1: Can we validly and reliably measure the extent to which students are supported to represent, measure, and model variability in Data Modeling classrooms?

Research Question 2: What do we learn about differences in classroom instruction with such a measurement system?

In this chapter I argue that relationships between underlying theoretical constructs and observable evidence are only implicitly addressed in the growing consensus on methods for developing fidelity measures and I describe three approaches researchers have taken when representing program theory in fidelity measures. In chapter two I lay out a vision for construct driven fidelity measurement that explicitly represents the underlying theoretical constructs of a program theory and the observable variables taken as evidence of the constructs. I end chapter two by describing Wilson's four building blocks of measurement (Wilson, 2004) and present it as a useful framework for supporting the development of construct driven fidelity measures. In

chapter three I describe my work in the first two blocks of this framework: representing the construct of fidelity in Data Modeling classes and operationalizing this construct in the form of a classroom observation system. In chapter four I describe my work in the last two blocks of this framework: creating an outcome space for the observation data that is related to the fidelity construct and modeling the observation data. These two chapters constitute what Mislevy (2007) calls a validity argument for this measurement system. Since this project is a test of the Data Modeling instructional theories, I have attempted to build a measurement system that indexes the extent to which these theories were realized in classroom interactions.

Program Theory as a Foundation for Fidelity Measurement

To address the wide variability in fidelity measurement, recent literature reviews have worked to establish common measurement practices by proposing procedures for creating fidelity measures (e.g. O'Donnell, 2008; Hagermoser-Sanetti & Kratochwill, 2009; Nelson et al., 2012). However, each measurement procedure, like any observation, is motivated by theoretical assumptions about teaching, learning, and the conceptual work of measurement. These theories guide observation schemes, and the interpretation of data. In the words of the NRC Report, *Knowing What Students Know*:

> *"Data do not provide their own meaning; their value as evidence can arise only through some interpretational framework. What a person perceives visually, for example, depends not only on the data she receives as photons of light striking her retinas, but also on what she thinks she might see." (Pellegrino et al., 2001, p. 43)*

In evaluation research, the researcher's conceptualization of the intervention being studied guides observation, what is often termed program theory (e.g. O'Donnell, 2008; Lipsey, 1993; Bickman, 1987). Although the term "curricula" typically conjures images of materials and teacher guides, a designed learning environment "is as much a conceptual construction as it is an

organizational one" (Lipsey et al., 1985, p. 10). The materials, tasks, and protocols in the curriculum design are motivated, if even implicitly, by a theory of teaching and learning.

Since the purpose of fidelity measurement is to measure the extent to which a designed program was realized as researchers had intended, a clear representation of one's program theory is a required antecedent to any meaningful fidelity definition. In fact, Lipsey (1993) argues that researchers must be theory oriented for experimental research to produce practical knowledge about mechanisms of change. Chen & Rossi (1980) argued over 30 years ago "Events, processes, and outcomes are not obvious; they can only be discerned with appropriate conceptual schemes that provide guides for observation" (p. 110). Without these, causal relationships between interventions and desired outcomes are little more than "buttons" (Lipsey, 1993, p. 34) to produce desired effects. A-priori representations of program theories are critical tools for describing the integrity of the realized learning environments, differentiating from business as usual learning environments, improving program implementation supports (such as professional development), and relating intervention mechanisms with learning outcomes (Cordray & Pion, 2006).

Chen & Rossi (1980) argue that the lack of a program theory is an early indication that a program is unlikely to be effective, and Lipsey et al. (1985) argue that "if there is not at least an implicit program model…then there is no program to evaluate" (p. 10). The program theory is separate in nature from what is realized in practice, and the relationship between the two gives conceptual meaning to the fidelity scale. Cordray & Pion (2006) use the term "treatment integrity," more common, however, is the term "fidelity of implementation" (O'Donnell, 2008).

In spite of the foundational nature of these representations, relations between program theory and the operationalized measures are usually poorly articulated in practice (Cordray &

Pion, 2006; Hagermoser-Sanetti & Kratochwill, 2009). In fact, Donalson & Lipsey (2006) argue, "what is meant by 'theory' encompasses a confusing mix of concepts related to evaluators notions about how evaluation should be practiced" (p.57). For this reason there is very little guidance on what constitutes a sufficient representation of a program theory. In fact, this issue is usually only addressed implicitly in the literature on developing fidelity measures.

This has led some to believe that a fidelity framing must assume a literal replication theory of program use. Some criticize fidelity measurers for conceptualizing program use as a straightforward act of implementing program components (e.g. Penuel, Phillips, & Harris, 2014). This is not surprising. Underrepresented program theories along with constructs borrowed from medical research, such as dosage, have given this impression. I'd like to suggest that this is too narrow an image of the work of measuring treatment fidelity. It is possible to conceptualize a program theory that views program use as an act of using tools, concepts, and practices to bring about particular instructional theories in relation to the needs of students in a local context. Under this type of program theory measuring fidelity is anything but straightforward. A more explicit conversation about appropriate representations of program theory is needed to differentiate between the different types of learning theories being tested in evaluation studies.

To begin this conversation I suggest that there have been three perspectives on the representational challenge of articulating program theory in a fidelity measure: 1) representing visible program structures, 2) representing visible program processes, and 3) representing underlying theoretical constructs. These perspectives are not mutually exclusive. In fact, they are better thought of as different aspects of a program theory representation. Another way of thinking about these categories is to imagine that researchers could represent 1) What

participants should do, 2) How participants should accomplish what they do, and 3) Why participants should engage in these particular forms of what and how.

Representing Program Theory as Program Structure

By far, the most common aspect of program theory explicitly represented in fidelity measures is the set of observable components of an intervention. Measures very often rely heavily on counts or proportions of visible pieces of the curriculum important for its success, often termed critical components or active ingredients (O'Donnell, 2008; Mowbray, 2003). Articulating the *critical* components is important because it is impossible for a measurement scheme to account for every facet of a designed intervention. So, the parts of the program that are central to its success should be prioritized over less important ones. Lists of critical components are used to define optimal, sufficient, and insufficient implementation by conceptualizing fidelity indices as a proportion of components observed with implicit reference to medical constructs such as dosage or adherence (Dane & Schneider, 1998).

In addition to the critical components, relationships among components are often

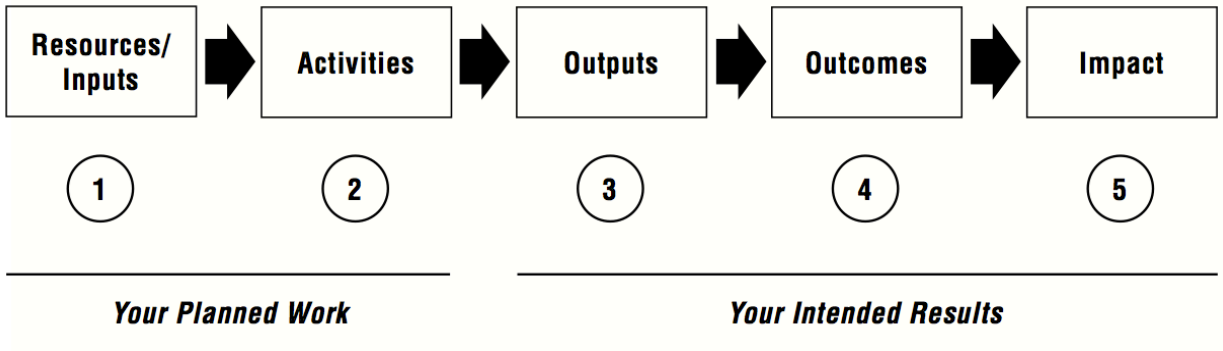


Figure 1. Example of general logic model structure from Kellogg Foundation

represented (e.g. Mowbray et al., 2000, Lipsey, 1993). Once components are listed, the "interrelationships among the inputs" (Lipsey, 1993, p. 36) describe how the components are to interact with each other. One common representation for this is the logic model (See figure 1 for a general example of logic model structure). The logic model arranges the critical components to communicate ordinal relationships, and the outcomes that they are expected to produce. For example, which components should come first? Which should co-occur? What changes in participants are expected if the components are observed?

Critical components and their relations are commonly described as program *structure* (Mowbray et al., 2000). The structure is seen as "the framework for service delivery" of an intervention (Mowbray et al., p. 318). Researchers either draw upon specific and scripted program models that have shown evidence of effectiveness, gather expert opinions from either content experts or from literature, or use qualitative research and site visits to determine the structure of the intervention (Mowbray et al., 2000, O'Donnell, 2008, Donaldson & Lipsey, 2006). For example, O'Donnell (2008) cites a five-step process developed by Mills & Ragan (2000) for identifying critical components:

> *(a) Identify the innovation components—participant activities, behaviors of the implementer, materials—by reviewing the program materials and consulting the program developer; (b) identify additional components and variations by interviewing past implementers to ascertain ideal use and unacceptable use for each component; (c) refine the program components by going back to the developer and clarifying with him or her user discrepancies regarding which of the observed components is the most important; (d) finalize the innovation*

components by constructing a component checklist and a set of variations within each before piloting; and (e) collect data either in writing (i.e., through questionnaires), classroom observation, or interview. (p. 49)

Notice that this process guides researchers in explicitly representing what participants should do but is much less explicit about representing the "how" and "why." The focus is on identifying discrete pieces of the designed program, and on documenting their presence in research settings. Components might be differentially valued if the program developer values some over others, but the rationale behind the differential values may not be clearly represented (step C from Mills & Ragan). Although the program developer's values are likely motivated by the theories that influenced the program design, explicit relations to that theory (the why) are not communicated in this framework.

This emphasis on the pieces of the intervention is seen in practice quite often. For a study of the 4th grade computer based math program, *Odyssey Math*, Wijkumar et al. (2009) described the program theory in terms of visible components in the classroom, such as student use of headphones and the visible posting of learning objectives. Fuchs et al. (2001) articulated the program theory of the *Peer-Assisted Learning Strategies* program by referring to structural components, such as the presence of curriculum folders for students and appropriate teacher feedback to students. Chard et al. (2008) used a checklist of lesson components, such as number of lessons completed, to conduct classroom observations in an evaluation of *Early Learning in Mathematics (ELM)*. They defined fidelity as the proportion of components checked across three observations. Crawford et al. (2012) explicitly articulated program structure for the web-based supplement, *HELP Math.* They defined the program structure as "observable behaviors and extant data such as frequency and intensity of contacts and evidence of procedural guidelines."

(p. 1). They described this aspect of the program theory as objective and visible. For a study of fidelity to a variety of high school math textbooks, Mcnaught, Tarr, & Sears (2010) referred to the program structure as "content fidelity," and they articulated this structure based on the written directions in secondary math textbooks. Ertle et. al (2006) referred to the program structure in the curriculum *Big Math for Little Kids* as "by the letter" implementation.

The critical components, although visible descriptions, still had to be articulated in a way that explicitly described the visible evidence of them in a classroom at a particular point in time. This work is often referred to as operationalization, and for program structure primarily includes decisions about measurement instruments and observation location, timing, and frequency. Of the conceptual papers I reviewed there was very little guidance as to what instruments measurement systems should make use of to account for program structure. In fact, the choice of measurement instruments is highly variable (O'Donnell, 2008), and there appear to be very few guiding principles in their selection. Researchers operationalize program structure as survey questions, observation protocols, student questionnaires, and interviews. There is some evidence that observations often suggest lower fidelity than self report (O'Donnell, 2008 from Emshoff et al, 1987), so researchers often employ multiple methods in order to triangulate data.

Some of the fidelity studies I reviewed only briefly discussed their operationalization of the visible program components represented in their program theory (Fuchs, et al., 2002; Wijkumar et al., 2009). They turned the critical components into observational checklists, and they selected the number of observations to be conducted. However, they did not provide a justification for the structural elements included in the protocols or for the number of observations.

Representing Program Theory as Program Processes

Standards for how components should be used are often referred to as program processes. Mowbray et al. (2000) and O'Donnell (2008) direct researchers to distinguish between "fidelity to structure" (did participants use the components?) and "fidelity to process" (did participants use them as they are intended to be used?). For example, a program component might be to post the lesson objectives each day. But, even if the objectives are posted they might be placed in an obscure location or printed in a manner that makes them difficult to read by students. Fidelity to structure is grounded in the critical component lists. Higher fidelity is differentiated from lower fidelity based on the amount of the structure visible in practice (i.e. how many days were the objectives posted?). In contrast, fidelity to process is grounded in the ways in which the components should be used (i.e. were the lesson objectives posted in a way that clearly communicated to the students?). This often evokes slippery theoretical constructs such as "quality." In fact, some researchers argue that structure can be observed objectively, but processes are always subjective judgments. For these reasons, fidelity to process is usually more difficult to measure reliably (O'Donnell, 2008).

Representations of program theories that include processes also challenge some of the conceptualizations of fidelity implicitly taken from medical trials. For example, the meaning of a construct such as "dosage" is less clear when one begins to account for quality of use (e.g. What "dose" of the posted objectives did students receive if they were posted every day, but in an obscure location?). Program theories that include processes also allow for more flexible fidelity definitions, especially if quality is conceptualized as responsiveness to student needs in a given context.

In addition to representing structure, Crawford et al. (2012) described separately intended teacher processes for the online supplement, *Help Math*. They considered this to be a more subjective theory of change. They listed teacher communication, classroom management, and problem solving as the three processes critical to the computer based math intervention they studied. Ertle et al. (2006) make a similar distinction by describing their program theory in terms of "by the letter implementation" (p. 9) and the "spirit of the implementation" (p.9). The "spirit" theory relied on general notions of quality drawn from math education literature, and the distinction between their program specific process theories and more general process theories was not clear. Doabler et al. (2012) took a similar approach to articulating their process theories by looking into the "converging knowledge base of effective mathematics instruction" (p. 3). They indexed behaviors such as teacher models, student mistakes, and teacher feedback. These fidelity measures, while differing in terms and definitions, all articulated program theory processes using general constructs found in math education literature.

In contrast, Clements et al. (2011) articulated specific teacher practices that are central to the preschool intervention they studied. They refer to this as evidence of "deep change" that "goes beyond surface structures and procedures…" (p. 137, from Coburn, 2003, p. 4). For example, a sample item that is scored on a five point Likert scale in the "Organization, Teaching Approaches, and Interaction" section is "The teacher conducted the activity as written in the curriculum or made positive adaptations to it (not changes that violated the spirit of the core mathematical activity)." The focus is on curriculum specific process, but these authors still draw on vague conceptions of quality, such as "positive adaptations." Munter et al. (2014) and Leiber et al. (2012), describe process in program theories as evidence of the "quality of delivery" (Munter et al., in press, p. 18). Although Munter et al. also looked to the math education

literature to identify aspects of quality, they linked these to the specific structures of the math tutoring intervention they studied, *Math Recovery.* For example, wait time and tasks in children's zone of proximal development are general theoretical constructs they drew from the literature, but they drew relationships between these general ideas to specific processes in the *Math Recovery* protocol. Similarly, McNaught et al. (2010) contrasted their structure, what they termed content fidelity, with processes called "presentation fidelity." However, they did not look to the literature, but rather to directions in the teacher editions of the secondary textbooks they studied.

I found one study that was unique in orientation and method in their articulations of processes in the program theory. Abry et al. (2011) studied the effectiveness of using *Responsive Classroom* practices in math classrooms. In this case, the program theory was articulated in terms of only processes, and not structures. This was because the *Responsive Classroom* approach seeks to modify the processes by which teachers are engaging students around current curricular structures. The program theory lists practices such as "positive teacher language", "guided discovery", and "academic choice." These practices were defined with one sentence. For example, academic choice was defined as "structured and monitored opportunities for students to choose, reflect on, and share work options based on individual interests."

When representations of program theory included descriptions of processes then the work of operationalization was more difficult because there is only implicit reference to theoretical constructs. In many cases the "why" still remains tacit. For example, Dane & Schneider (1998), name "quality of delivery" as a general dimension of fidelity to process that should always be measured, but its operationalization is not clear. They describe quality of delivery as "a measure of qualitative aspects of program delivery that are not directly related to the implementation of

prescribed content, such as implementer enthusiasm, leader preparedness, global estimates of session effectiveness, and leader attitudes toward program" (p. 45). This definition draws upon theoretical constructs of "enthusiasm" and "preparedness" and "attitudes" without articulating the meaning of these terms. These terms and could be conceptualized in many different ways.

While structural aspects of the theories can be turned into checklists, the processes often cannot. Instead, they are often operationalized as Likert type rating scales. In fact, the rating scales were sometimes the only representation of an underlying construct's meaning. Since O'Donnell (2008) directed researchers to use operational definitions to communicate to teachers what the constructs mean researchers often reference the operationalized instruments in an appendix as the definition.

I found that operationalizations of math intervention program theories articulated in terms of structure and process were more complicated, and many researchers described the process as a challenge because notions of quality were often difficult to operationally define and score reliably. It was common to employ multiple instruments including live observations, surveys, logs, video coding, and transcript coding. While many of these studies discussed the desire to triangulate data, none provided a rationale for which source was privileged or an explicit description of the object being triangulated. Moreover, although the instruments often provided contradictory information, they did not describe which account was more valid. In fact, Clements et al. (2011) and Ertle et al. (2006) used previously developed measures of general instructional quality in addition to the intervention specific measures. Although they did not explicitly discuss the rationale for the choice, it does suggest that there was an implicit theoretical alignment between the measures that was not articulated. Ertle et al. carefully selected 4 curriculum activities with the intent of observing a wide variety of concepts and practices. Munter et al.

(2014) chunked their video into instructional moments and coded each of these with the fidelity scheme. The instructional moments selected were designed to show the teacher responding to student thinking with tasks and questions. Crawford et al. (2012) used a mix of "formal" observations in which the teacher was notified ahead of time and "informal" observations that were conducted by surprise. This suggests an implicit theory that sees classroom interactions during planned observations as different than those during surprise ones. Brown et al. (2009) coded transcripts by "shifts in activity," and then scored the fidelity variables for each section. The shifts were chosen so that each activity chuck could be characterized by the two "opportunities to learn" categories for students.

Representing Program Theory as Underlying Theoretical Constructs

In addition to representing program structure and process, a small number of studies represented underlying theoretical constructs in their program theory. These studies attempted to explicate the theoretical commitments that motivated the structures and processes (Nelson, et al., 2012). From this perspective, specific tasks, materials, and protocols of each intervention (the what and how) are but visible manifestations of underlying instructional theories (the why). These studies emphasize that there is an important distinction between underlying constructs, and the material form motivated by the constructs.

Nelson et al. (2012) argue that a program theory should be first articulated "in conceptual terms by describing the constructs that underlie activities and resources and effect outcomes" (p. 6). The authors call this type of theory a program's "change model." Nelson and his colleagues provided figure 2 as an example of a change model from the LINCS project (Swafford, Jones, & Thornton, 1997). This representation is termed a "theory approach model" by the W. K. Kellogg Foundation's logic model development guide (W. K. Kellogg Foundation, 2004). This model

names the assumptions that led the program designer to structure the intervention as he or she did. The rationale for these representations is based on Weiss's (1998) idea that "A program is a theory, and an evaluation is its test. In order to organize the evaluation to provide a responsible test, the evaluator needs to understand the theoretical premises on which the program is based" (p. 55). The change model names the underlying theoretical constructs that motivate the intervention structure and processes.

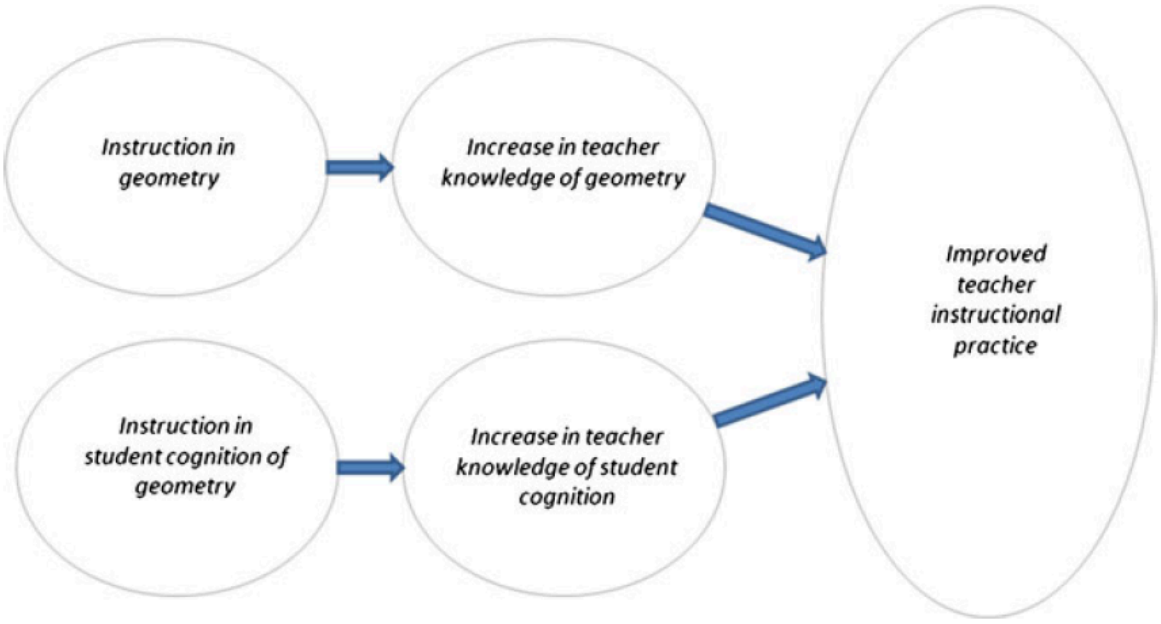


Figure 2. Example of change model from LINCS project

Nelson et al. (2012) go on to justify the need for articulations of underlying constructs in five parts. First, abstract constructs provide the opportunity to inform theory outside of the specific context of an intervention's evaluation. If the program theory is articulated in terms of the underlying constructs then data can be used to inform more general questions than those related to the specific realization of the intervention at hand. Descriptions of the conceptual

organization of a program are needed to inform more general questions around the theoretical orientation of the design. Second, change models can support the development of a richer network of causal connections due to the specificity of the constructs. Third, the change model informs what is to be measured and how. As opposed to looking for every resemblance of an intervention, or a list of components that may or may not be evidence of the underlying constructs of interest, measures informed by theoretical constructs can be deliberately built to make inferences about the integrity of the instructional theory realized in classroom practice. Fourth, the change model guides the analysis of the measurement instruments. This is a particularly timely goal given the lack of validity evidence in most fidelity tools. Threats to validity most often come from two sources: 1) construct underrepresentation, and 2) construct irrelevant variance (Messick, 1994). When visible indicators underrepresent instantiations of the construct, then fluctuations in the construct can go unseen. On the other hand, if indicators are not, in fact, related to the construct, then variation in measures can occur even though there is no true variation in the constructs of interest.

I found only two fidelity measurement systems that explicitly named theoretical constructs. Although not articulated in a change model, Munter et al. (in press) described a construct of "positioning students as sense makers and their own mathematical authorities" (p. 16). So, fidelity of the intervention was conceptualized as the extent to which tutors positioned students in this manner. Brown et al. (2009) focused their attention on the "opportunity to learn" available to students through instruction. They conceptualized this opportunity using two constructs, opportunity to reason about mathematics and opportunity to communicate about mathematics. In addition, they explicitly articulated theory related to "implementation." They drew on Remillard's (2004) ideas about the socio-historical production of curriculum materials.

Figure 3 is their model of implementation in which they view the teacher and the materials in interaction with the students and external forces.

Abry et al. (2011) was the only study found that explicitly named Nelson's framework. However, their change model did not actually name latent constructs. They named "use of Responsive Classroom practices" as the construct of interest, and they listed the practices they looked for without providing a description of the constructs motivating the processes. I included them here because of their explicit reference to Nelson's framework, but in practice they actually named processes without explicitly naming the theoretical constructs motivating them.

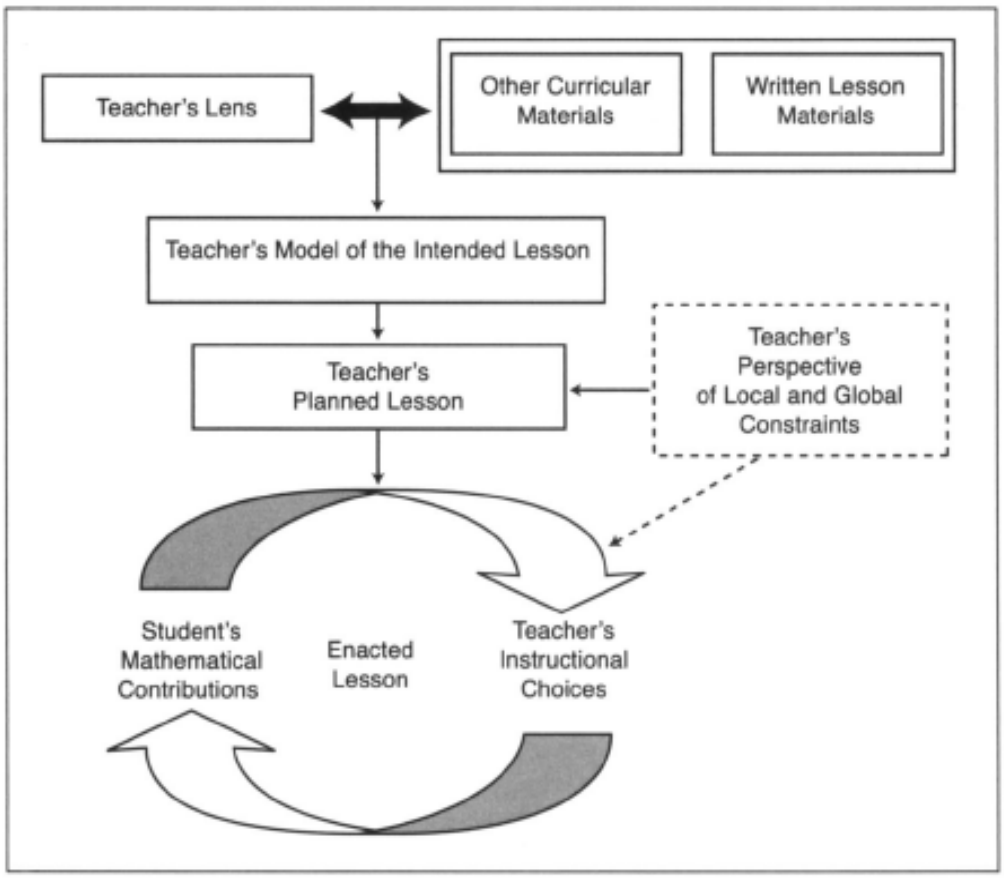


Figure 3. Brown's conceptualization of curriculum use.

Nelson and his colleagues describe operationalization as the work of transforming a change model into a logic model. While the change model names the constructs of interest and their relationships, the logic model describes the visible structures and processes these theories take in the intervention. These authors dissect the main constructs into sub-components, and then generate facets, or the "specific behaviors, events or resources that constitute the implementation of a subcomponent" (p. 12). Using these facets the authors describe a process of defining the indicators by creating observable variables and indices that constitute a measurement instrument used to account for the indicators. The logic model allows one to make inferences into the meanings of the underlying constructs based on their material form; they do not provide explicit articulations of their meanings.

For Munter et al. (in press), operationalization was an iterative process of defining scoring rubrics, piloting, and reviewing with Math Recovery experts. They note the program theory as the lens by which critical components of the curriculum were identified in curriculum materials. However, the authors were not the original curriculum developers, so the task of operationalization was challenging and even included attending national Math Recovery conferences and working directly with curriculum developers. In addition, they categorized operationalized variables as "unique and essential", "essential but not unique", or "prohibited" (Waltz et al., 1993). Operationalizing the construct of "positioning students as sense makers" was challenging for these authors, took multiple variables, and it was often easier to operationally define what should *not* be done than what *should* be done. They attributed this to the lack of shared understanding of the program theory among expert users.

Brown et al. (2009) used a similar strategy to operationalize their "opportunities to learn" constructs. They looked through curriculum materials and defined nine codes that informed this

construct. For example, they noted when students were given opportunities to compare and make connections among multiple representations. They also segmented classroom transcripts into segments of activity that provided opportunities to see the codes.

The very few studies that name the underlying constructs when representing program theory describe them at a very general level. This still leaves many fidelity measures susceptible to construct underrepresentation and construct irrelevant variance, because this level of generality does not support explicit relations between the constructs and the observable components of the fidelity measure. In addition, it does not encourage opportunities for researchers to discuss alternative conceptualizations of phenomena, such as teacher quality.

Summary

Almost all fidelity measurement systems reviewed here represented program theories exclusively in terms of program structure and/or processes. Those that only represent structure allow for very little flexibility in curriculum use since any deviation from the structure is evidence of infidelity. This has led to the impression outside of evaluation research that fidelity frameworks always theorize program use as the implementation of static objects. Some fidelity researchers perpetuate this idea by arguing that fidelity and adaptation are separate constructs requiring different measurement instruments (O'Connor, 2008).

Researchers that account for processes allow for more flexibility, but implicitly draw upon theoretical constructs to account for the quality of the use of program structure. However, observers often have different conceptualizations of these constructs, even among expert program users, and without explicit articulations to guide their disagreements, program developers were called upon to resolve differences in an ad hoc manner. Even this tactic, though, didn't always work. Munter et al. (in press) found that expert users of the *Math Recovery*

tutoring program differed significantly on their ideas about teacher practice. Since researchers rarely represent the meaning of the theoretical constructs it was usually impossible to compare different conceptualizations of identically named constructs, such as quality.

A small number of measurement systems named the underlying constructs that motivated visible components. These represented the what, how, and why for the programs they studied. These articulations, though, were usually at a very general grain size, which still left ambiguities about theory-evidence relations in the measurement scales. These ambiguities, which were present in all the studies I reviewed, resulted in scales that were not easily interpretable in terms of the program theory. Is a score that is twice that of another score representative of an enactment that is twice as aligned with the designer's intentions?

To illustrate this limitation consider again temperature measurement. The quantities that make up the scales are grounded in the experiences that contributed to our qualitative understanding of temperature. Ice is not simply *colder* than boiling water; it is 100 degrees Celsius colder. On the different scales there are reference points that we can interpret in terms of our qualitative understandings of the phenomenon. This is not true for these fidelity measures. Although one classroom might have a score that is twice the score of another classroom it is not clear that this indicates twice as much alignment. In addition, there were no reference points on the scales that related the quantities to qualitatively different theoretical states of fidelity. For this reason, fidelity measures are typically interpreted as having only ordinal scales.

CHAPTER II

CONSTRUCT DRIVEN FIDELITY MEASUREMENT

Introduction

A more explicit treatment of the theories motivating fidelity measures is needed to create measurement scales that help address the "how much?" question. In this chapter I describe a theoretical perspective for fidelity measurement systems called construct driven measurement as a good starting place for this type of measurement work (Messick, 1980, 1994). I also discuss how this perspective builds onto the work done by researchers that represented structure, process, and theoretical constructs. I should make clear, though, that construct driven measurement is not a new idea. It has been discussed at length under various terms in the development of student assessments. This is not to say that all written student assessments align with this perspective in practice, but that there has been much work recently to bring contemporary learning theories into contact with contemporary measurement theories (e.g. Pellegrino, et al. 2001).

Although student assessments are different in many ways from fidelity measures, they are relevant and important to consider for three reasons. First, like fidelity measures, the theoretical foundations of student assessments are rarely articulated and discussed explicitly. Instead, issues of implementation and score interpretation dominate the discourse (Pellegrino et al., 2001). Second, because the theoretical underpinnings have remained implicit, they are rarely scrutinized or revised. Mislevy (1993) has argued "it is only a slight exaggeration to describe the test theory that dominates educational assessment as the application of 20th century statistics to 19th century

psychology" (p. 19). Chen & Rossi (1980) characterized the program theories behind educational interventions similarly. They said that that all programs are motivated by some theory, but that often times the theories are based on "conventional, commonsense understandings" (p. 110) of social phenomena that do not align with contemporary research. However, many of the questions student assessments and fidelity measures are used to answer are related to fundamental conceptualizations of teaching and learning. What is important to teach? How should we teach it? Like student assessments, the fidelity measures designed to characterize the "causes" of the assessment results often leave the underlying theories and assumptions implicit. Third, fidelity measures are inextricably linked to student assessments in evaluation research. Fidelity measures relate classroom interactions with student learning. If the theoretical assumptions of one or both of these types of measures are not made explicit, then we run the risk of chasing co-variation between them without contributing to our theoretical understanding of the complex relationships between teaching and learning.

Program Theory Relating Latent Constructs to Visible Structure and Processes

Conceptualizations of learning, teaching, and as a result, fidelity will be unique to each intervention design. For some, the curriculum guide might be conceptualized as a straightforward representation of practice for teachers to follow, so fidelity might be thought of as the extent to which the teacher literally reproduces the written descriptions of the curriculum design (Remillard, 2006). How much of the curriculum was replicated? What dose did participants receive? In contrast, curriculum guides might be conceptualized as socio-cultural artifacts to support teachers to recontextualize a theory of action to meet the needs of students in a local context. In this case, fidelity would take on a very different meaning. There can be no literal replication, but there are more and less productive uses of the materials from the developer's

perspective. Are materials supporting classroom interactions as we intended? Are students discussing the concepts we value? It is critical that these theories be made explicit so that the conceptualization of the object being measured is communicated.

Non-scripted interventions, such as the *Math Recovery* tutoring program, highlight the need for more than articulations of structure and process when representing a program theory. Programs that are responsive to local contexts can take shape in many different ways while remaining faithful to the intentions of the curriculum design. Since the structure of the program is dependent on local context, it is dynamic and providing a checklist of critical components and processes becomes much more difficult due to variation in faithful implementation. In addition, there is sometimes no clear delineation between the structure of the program and the process by which participants should enact it. For example, if a critical component is to hold a whole class discussion in which a teacher identifies and responds to student thinking in real time then determining whether the component was seen cannot be done without an assessment of the process. In this case the structural component *is* the process.

To further illustrate the distinction between visible components and motivating theory, consider recent efforts to measure "positive infidelity" (Hulleman & Cordray, 2009; Munter, 2010) or "positive adaptations" (Clements et al., 2011). Both positive infidelity and positive adaptations describe alterations to the intervention design, otherwise considered infidelity, that actually accomplish the goals of the intervention better than the original design. This term makes clear that the visible instantiations of the curriculum are not sufficient descriptions of program theory. Program components are motivated by invisible theoretical constructs, and as the idea of positive infidelity demonstrates, it is possible to modify the components to better realize the constructs in action. Without the distinction between motivating latent constructs and visible

components, this concept would be senseless. But when the motivating theories are acknowledged to be distinct from their material instantiations, then one can be faithful to the invisible theory while simultaneously unfaithful to the intervention protocol.

There is a more fundamental limitation with articulations of program theories composed of only visible structure and process. The underlying motivating theories remain implicit at best, and invisible at worst. So, although the research design might seek to infer relationships between visible aspects of the intervention (materials, protocols, etc.) and visible aspects of the outcomes (test scores or survey responses), it remains unknown if the visible evidence supports the latent theories motivating the measurement instruments. Identical observable behaviors can be prompted by very different motivations. Program theories that are articulated in only observable terms, therefore, produce instruments that do not explicate the rationale for looking for the observable component, and therefore may provide invalid interpretations of empirical data.

So, what is a sufficient articulation of an underlying theory? Recently, developers of student assessments have found it productive to articulate their theories as one or more latent constructs (for an example see Lehrer, Kim, et al., in press). Descriptions of the constructs should be specific enough to communicate meaning in a particular context. They should include a "coherent and substantive definition for the content of the construct" (Wilson 2006, p. 26). However, the descriptions must be general enough to be re-contextualized in new settings. So, the articulations cannot be solely in terms of visible curriculum components because these do not provide guidance for finding evidence of the construct in a new setting. The representation of a construct should describe a trajectory along which participants may be located. Remember, fidelity measures are designed to describe realized classroom interactions in relation to a

theoretical ideal, so the descriptions can describe the construct along a trajectory to give meaning to "more" and "less" ideal.

Of course, using one or more latent constructs to represent program theory will significantly reduce the complexity of it. As a result, these articulations will always simplify the theory by ignoring many aspects of classroom phenomena. However, this reduction is not haphazard. The purpose is to amplify the most relevant aspects of the theory by reducing those that are peripheral (Latour, 1999). Latent constructs were named in Nelson's change models, but the meanings of the constructs were not necessarily articulated at a level of specificity I am describing here and they were not represented as a trajectory. This is problematic because researchers in education often use different terms to reference the same construct, or alternatively, the same term with different conceptualizations. But if researchers do not articulate the meaning of the latent constructs underlying an intervention then the nature of the discrepancies can remain hidden.

Generating Observable Evidence of Program Theory

An understanding of how a program theory is manifested and where clear evidence of it might be found is necessary to index it. Observable evidence of the latent constructs in a program theory is not equally visible at all times, and it can be displayed in many different ways. Program users' assumptions and theories related to teaching and learning always serve as an interpretive framework for curriculum materials (Remillard, 2005), and materials are modified and used in a variety of ways, which may or may not meet the goals of the program theory. Spillane (2004) put it nicely by stating "When locals understand ideas about revising their practice from policies that are inconsistent with or that fall short of those intended by policy makers, they unwittingly and unknowingly undermine the local implementation of these

policies" (p. 169). Because of this, observable variables can be seen in many different configurations that may be more or less faithful to the latent constructs motivating their original design.

From a construct driven measurement perspective, operationalization is the work of describing observable scenarios in which relevant latent constructs are made manifest (Wilson, 2004). This includes the physical location where the construct is likely to be realized in observable action, the variables one should account for to characterize the realization of the construct, and an outcome space of meaningful combinations of the variables that might be observed (Wilson, 2004). Fidelity measurement systems should communicate the inferences made from the outcome space by demarcating the regions that indicate qualitatively different classroom interactions. If articulations of latent constructs in the program theory describe "more" and "less" fidelity, the qualitatively different outcomes from observational variables can be mapped onto them so that the quantities are ordered in terms of increasing fidelity.

At first glance, this appears very similar to a framework developed by Hall & Loucks (1977). They argued for researchers to think developmentally about treatment implementation, and to articulate "Levels of Use" (LoU) as a foundation for fidelity measures. As the name suggests, the levels described the "uses" of the intervention, or how the different core components and processes were implemented. However, it did not represent the underlying latent constructs that the different uses provided evidence about. LoUs are, however, helpful in thinking about how one might operationalize a theoretical construct that has been articulated along a trajectory. The outcome space of the variables can be ordered to describe different levels of use, and can be coordinated with representations of the motivating latent constructs. I will

provide an example of this kind of representation when I describe the Data Modeling measurement system in chapter three.

The Relationship Between Observation and Theory

With the program theory represented as a trajectory of fidelity, and the visible evidence and expected states articulated, the relationship between the two must be described. This is the crux of measurement work, and it highlights that measurement is fundamentally the modeling of relationships between latent constructs and observable evidence of them (van Fraassen, 2008). Researchers reduce the world's complexity and impose theoretically motivated coding schemes to generate measures. But these measures are used to make inferences about the motivating theories. This modeling work is almost never explicitly discussed in contemporary fidelity methods, which is why the fidelity scales are typically unable to address the "how much more?" question. Measures inform our understanding of our theories, and our theories determine if our measures validly account for relevant aspects of a phenomenon. Since the measures stand for the theory, one must interrogate them for both construct underrepresentation and construct irrelevant variance (Messick, 1997). So, do changes in measure faithfully reflect changes in the latent construct? What do the changes say about the construct?

Even if observable variables are motivated by well-articulated program theory, researchers should still interrogate each variable to see if they "behave" as expected. To illustrate, consider for a moment a more common measurement task. If an outdoor thermometer reads 95 degrees while you are standing in snow on a winter day then you will immediately recognize that the measurement tool is not behaving as you expected. This is because we constantly look, although often implicitly, for a correspondence between measures and the phenomena they are designed to index. Implicit theories about the relationship between

temperature and snow would tell you that the measure produced by the tool and your theoretical understanding of temperature are incongruent, leading you to distrust the measure. Likewise, measures produced by classroom operationalizations must be related to the theoretical constructs that motivated them to determine if they are "working."

However, the correspondence has influence in both directions. The latent construct serves as the motivation for the observable variables, but the observable variables can also change our understanding of the construct. Researchers are often surprised to find that their variables did not function as they expected. Sometimes this is evidence of poor operationalization, but other times it leads to a reconceptualization of the theoretical construct. Establishing such a correspondence requires a model of the expected relationship between evidence and theory, often called a measurement model (Wilson, 2004). The measurement models supporting validity arguments in fidelity measures are not explicitly discussed in contemporary methods. While operationalization should explicitly describe how the latent construct motivates observable variables, a measurement model should explain how the observable evidence informs inferences about the construct. Remember, the latent construct is not only an initial motivation for variables, but is the articulation of the phenomenon being measured. However, the data alone does not provide inference back to the construct map. The measurement model guides this work. Figure 4 represents the three main parts of fidelity measurement work from a construct driven perspective: the structure of the program theory, the structure of the observable evidence, and the relationship between the two.

Not all measurement shares a construct driven perspective. Messick (1997) calls measures that are only interested in visible products and behaviors "task driven." For example, figure skating judges in the Olympic games measure only the production of an act in one

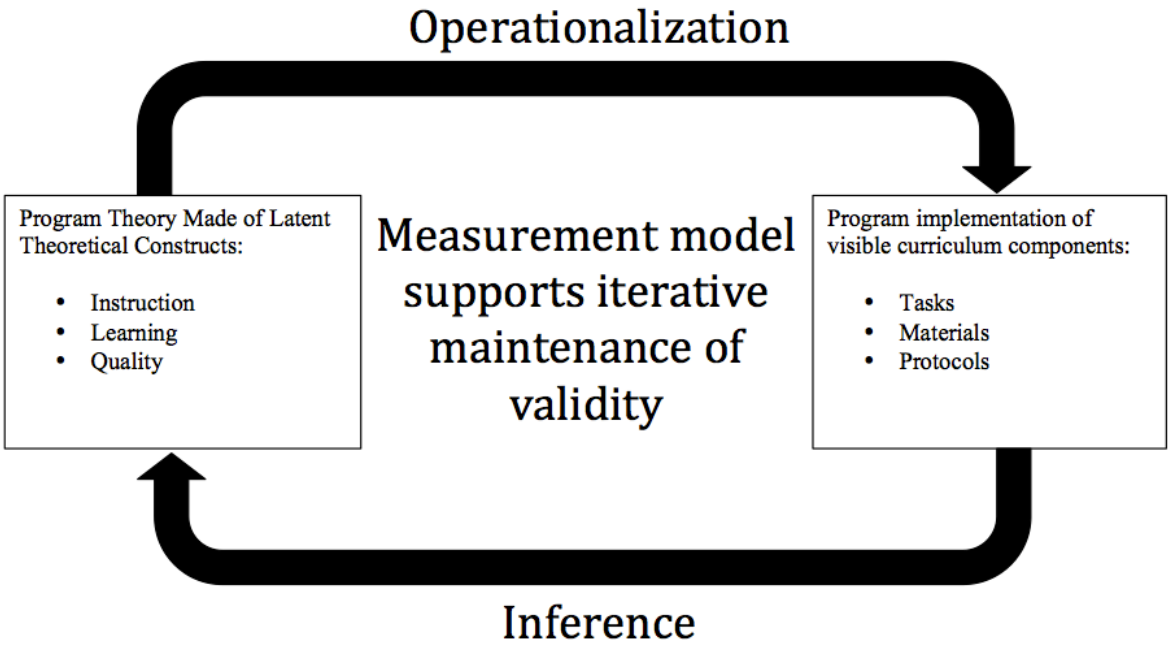


Figure 4. Construct driven fidelity measure.

moment. Although they might acknowledge that the act sometimes does not accurately reflect the person's ability, the latent construct of "ability" is not the ultimate focus of the measures. The object of measurement is the routine performed in a particular moment. However, fidelity measurers have a very different goal. Researchers want to characterize program implementation in relation to a theoretical ideal using only a sample of classroom interactions. So, fidelity measurement must be designed to make inferences to latent theory using a sample of observable evidence.

Another motivation for a construct driven approach is that fidelity studies should contribute to the field's understanding of theoretical constructs related to teaching and learning (Cordray & Pion, 2006). A task driven measurement approach does not allow for these types of claims since descriptions of the object of measurement are solely in terms of the visible manifestations in a specific context. It does not provide tools for making claims about theory. For example, two designed learning environments could be based on the same latent theories of learning but have very different visible designs. It is important to delineate between claims about

the theory, and claims about the designs. Alternatively, designs may be based on very different theories. In this case, the question is not which design is better at realizing the theory, but which theory should be preferred. To make these distinctions visible one needs articulations of the latent constructs motivating the visible instantiation. Why is a particular task important? What types of learning do each task support? What alternative material form could theories have taken, and why did they take their current form?

Although I see the Nelson et al. (2012) discussion on validity as compatible with the construct driven measurement perspective, there is an important distinction. The modeling activity for Nelson et al. (2012) is to "completely characterize the intervention" in the change and logic models (p. 11). However, from a construct driven perspective measurement is a modeling activity that reduces the complexity of the world into material and social structures that highlight relevant attributes of the target domain (Pickering, 1995; Kuhn, 1996). The model, therefore, will never completely characterize the intervention, but should highlight relevant components in a manageable way. In fact, Pickering (1995) argues that the world and the machines (including measures) we build to engage with the world have agency themselves. For fidelity measurement, effects of local contexts on the use of curriculum materials are examples of such agency. For example, the very definition of fidelity can change when the needs or goals of the local context change (Mowbray et al., 2003).

At first glance this may come across as a head-in-the-clouds distinction, but my purpose is pragmatic. Construct driven measures treat validity not as something that is established once and for all, but something that is maintained through iterative mapping between the constructs of interest and the observable attributes taken as evidence (Wilson, 2004). So, the relationship is never stable, and never completely characterizes an intervention. Rather, it is a tenuous

relationship where the construct influences our operationalization, but our data can also change our conceptualizations of our construct through a measurement model. So, if the model is what upholds the tenuous relationship, the importance of explicit attention to it cannot be over stated.

A Tool for Developing Construct Driven Fidelity Measures

In this chapter I have described a construct driven conceptualization of fidelity measurement. However, I have suggested no practical tools to carry out this work. Although general notions of fidelity and common measurement schemes are unlikely due to the uniqueness of each intervention, it is important to work towards a common framework and set of conceptual tools for carrying out this challenging work. Here I briefly describe a framework that has been used to construct measures of student thinking, Wilson's (2004) Four Building Blocks of Measurement.

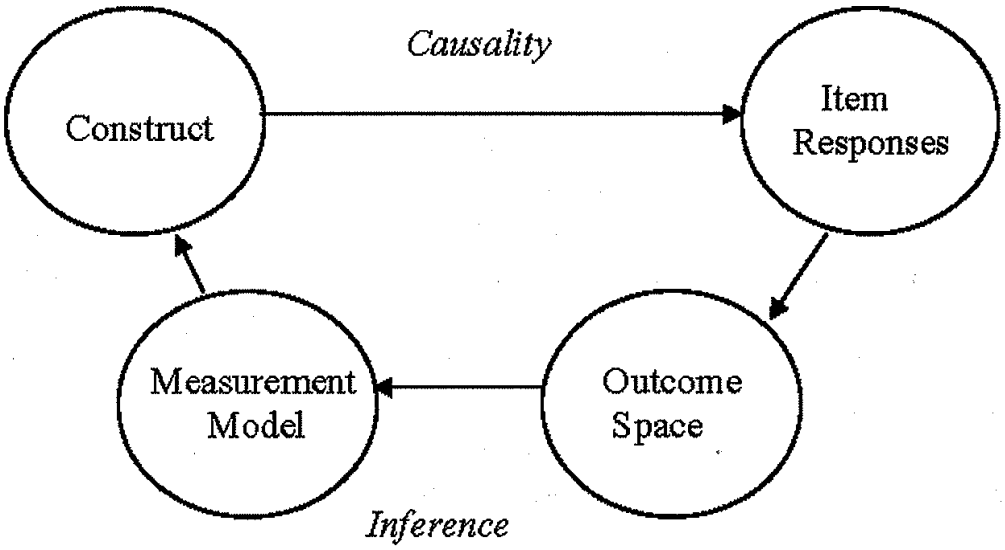


Figure 5. Wilson's four building blocks of measurement.

As figure 5 shows, this framework provides tools that guide researchers through four steps: (1) articulating latent constructs, (2) designing observations to index visible evidence of the construct, (3) considering the meaningful differentiations in observation outcomes, and (4)

modeling the measures to support inferences about the constructs. Notice that figure 4 is an image of a method for carrying out the iterative work seen in figure 4. This framework is well suited for developing fidelity measures since it does not assume common constructs and theories, but instead supports researchers in articulating the latent constructs unique to each program theory and relating them to the observable evidence used to make inferences about the theory.

However, this framework has yet to be used to develop a fidelity measure. As a result, there are some concepts that have a shared meaning in the world of student assessments, but do not in the context of fidelity measurement. In this section I briefly describe each of the four blocks while recontextualizing their meaning for fidelity measurement. In addition, I describe how this framework might mesh with the current tools supporting the development of fidelity methods, such as change models and logic models.

Block 1: Construct Map

The intention of the first block is to make explicit the answer to an all to often-ignored question: *What am I trying to measure?* This framework provides the construct *map* as a tool for articulating the latent constructs motivating one's theory. This tool can be used to represent constructs in terms of both the different states relevant to the program theory, and the order of the states in relation to fidelity to the program theory. In fidelity measurement this serves as the foundation for defining the relationship between ideal and realized implementation. The highest level of a construct map serves as the ideal, while the other levels serve as likely alternatives ordered in increasing similarity to the ideal. The levels might also be thought of as stepping stones towards the ideal. This also provides an initial theoretical grounding for one's scale. This tool makes visible the way one conceptualizes their construct, both for themselves and for others.

It provides a common framework for articulating latent constructs in a specific enough language to have meaning, while also being general enough to be meaningful in other contexts.

The construct map would be a productive contribution to Nelson's change model. While the change model names the latent constructs of interest, the construct map would provide a tool to communicate their meaning. Of course, one's program theory might be composed of multiple constructs. In this case, separate construct maps serve as articulations of each construct. Additionally, relations between constructs might be developed and articulated. On the other hand, one particular construct might be prioritized to create a manageable measure. These decisions will vary across different designed programs, but this step of measurement development would make explicit the rationale for these kids of choices.

Block 2: Item Design

Once researchers articulate relevant constructs, the next block has them consider observations that would generate knowledge about the theory. However, observations consist of more than looking for, attending to, recording, and summarizing visible attributes. This is a "special sort of observation that is generally termed an *item*" (Wilson, 2004, p. 41). The term "item" has a shared meaning when developing achievement, attitude, or survey measures. In these contexts, researchers often conduct an observation by designing a written prompt (such as a test question) with instruments to record the participants' response (such as pencil and paper). However, this type of observation will likely be insufficient for fidelity measures since evidence for program theory constructs is typically found during classroom interaction. The specific details of each intervention could require very different observable evidence so the observation design must be articulated in relation to the construct map to communicate the theoretical rationale.

Although the specific details of item design will vary for study to study, the items must justify the relationship between the observations and the program theory described in the construct map. It is thinking carefully about item design that makes clear that measurement systems don't *collect* data, but rather *create* data. Measurers must justify the mechanisms designed to create the data. There are four elements of an item design that should be considered in the context of fidelity.

- **Observation Location:** For fidelity measurement it is unlikely that a researcher will need to design a task to manifest the construct since the construct will be manifested during the ongoing classroom instruction. However, researchers will need to consider which moments of classroom interaction will provide the most relevant information about the underlying program theory. This is not only a choice about physical location, but also the location within the program trajectory (i.e. which units and activities?). Early in the process this will likely involve more open-ended observations and considerations of tradeoffs between alternative units, days, or times for observations. The question that must be asked is "where will we see evidence of the constructs in our program theory?" It is important that this question is given careful thought, and that the answers to the question are grounded in the program theory because decisions about observation location both provide for and eliminate particular kinds of evidence.
- **Observation Evidence:** Once observation locations are chosen researchers must decide what they will count as evidence of the program theory's constructs during the observations. This might be thought of as the observable variables, indicators, or scoring guides. The question at this stage is "what are the observable actions that will provide evidence about the construct?" There are a wide variety of choices that can be made here.

Researchers might look for counts of behaviors (such as student comments), duration of behaviors, or ratings of behaviors. However, these choices all should be justified by the extent to which they provide relevant evidence about the construct map.

- **Observation Instrumentation:** The first two decisions might give the reader the impression that item design in fidelity contexts assumes live observations, but this need not be the case. The final question of an item design that must be considered is how to instrument the data-generating framework (Observation location and evidence) to create records of the observations. This could be done in a live observation, but even then researchers must decide how the observers will see, hear, and record the desired information. On the other hand, live observations are not the only strategy. Researchers could video record classroom instruction, or they might ask the students and teacher what happened during instruction after the fact. Clearly these choices will have implications for the validity of the inferences supported by the data. Some of the concerns will be theoretically grounded, "Do teachers really 'see' their instruction in a way that will allow them to tell us what happened after the fact?" Others are more practical, "can observers reliably record all of the variables in the system during live observations?" Both considerations must be justified to create items that can provide trustworthy information about he program theory.

Block 3: Specifying Outcome Space

Even in relatively simple measurement schemes there are likely a number of different data profiles that the items might produce. For example, consider a hypothetical fidelity measure with three variables scored on a 5-point Likert scale. Even in this overly simplistic example there are 125 possible combinations for a given observation. Which combinations count as "more"

faithful? Which fall in the intermediate states? Which counts as the least faithful version of the theory? In this step of development researchers have to provide a rationale for which profiles correspond to each of the construct map levels. This provides visible instantiations of each theoretical level articulated in the construct map, thus linking the item responses to the theory they are designed to inform. The qualitative distinctions described in these categories provide a rationale for the inferences they suggest about the construct map.

It is important to note that not all items will inform every level of a construct map. For example, items looking for the presence of curricular materials might help a measurer to see that at least low levels of the construct are observed (the materials are at least present in the class!), but it likely won't provide much information about higher levels. It is important during this stage of development that researchers make sure their outcome space provides information about the entire span of the construct map. If one finds that the items only inform the lower or higher levels then more work is needed to design alternative items that will index the others.

Block 4: Measurement Model

Different measurement models can be specified to make inferences about the program theory, but they should allow for two types of inferences. 1) They should support inferences regarding the location of both item responses and respondents on the construct map. For fidelity this provides information about the relationship between observable variables, classrooms, and the conceptualization of fidelity articulated in the construct map. 2) The model should enable meaningful interpretations of distance between different item responses and particular respondents (Wilson, 2004). These two principles support the development of a scale that can be interpreted in terms of the construct map so that the distance between respondents can be characterized by qualitatively different program implementation.

CHAPTER III

THEORIZATION AND METHODS FOR MEASURING FIDELITY IN DATA MODELING CLASSES

Data Modeling Instructional Sequence

Students are often exposed to data displays and statistics as procedures to do with numbers. Rarely are they supported to see these as tools to give data meaning and to communicate that meaning to readers. In contrast, the *Data Modeling* instructional sequence engages students with the concepts of data display, statistics, chance, modeling, and inference as tools to answer two seemingly simple questions: 1) what is the length of our teacher's arm-span? And 2) How precise were we as a group of measurers? With these questions as the driving motivation, the Data Modeling materials support teachers to facilitate the development of three epistemic practices in their classrooms: representing variability, measuring variability, and modeling variability. The practices are epistemic because they provide the means by which the original question (about the teacher's arm span) is answered. As the students develop competencies in these practices they use them to generate new concepts, which in turn provide new epistemic tools that further develop the practices. This observation measure is designed to index the extent to which teachers supported students to productively engage with these concepts and epistemic practices.

For these concepts and practices to hold epistemic value it is important to first engage students in a context in which the tools are needed to tell them something about the world. This is the rationale for the first box in figure 6, the conceptual logic model for the Data Modeling

design. In Data Modeling, students begin the first unit by independently measuring their teacher's arm span with a 15 cm ruler. The multiple iterations needed to span the distance produces significant variability in the measurements, and this is the first indication to students that their driving question is not as straightforward as some originally thought. How can we know the true length if we all got different measurements? The teacher's job is to use students' intuitions about the hidden structures in the data to provoke the need for data displays that highlight the structure. For example, students notice that some measurements are similar, but others are very different. So, the students next invent strategies for displaying the data to make a pattern or trend visible at-a-glance. Giving students opportunities to invent their own data display allows them to grapple with the choices one must make and the epistemic consequences of the choices. This is the rationale for the second box in figure 6.

However, methods are best understood when brought into contact, challenged, and refined (Ford, 2010). This is the rationale for sharing and comparing in figure 6. With this in mind, the teacher uses the student-invented displays as resources to facilitate a whole class discussion we call "display reviews" to promote concepts of order, class, scale, and count (Lehrer, Kim, & Schauble, 2007). In contrast to typical instruction, these concepts are not treated as abstract ideas or rules to create a "correct" display, but as epistemic tools to help readers "see" the true length of the teacher's arm-span, which the final box in figure 6 represents.

The whole class conversations about the invented displays are designed to bring out two additional ideas, the data has a signal to the true length of the teacher's arm-span while also having noise from the mistakes students made while measuring (Petrosino, Lehrer, & Schauble, 2003). Students can readily observe that each measurement is not equally likely to be the true length. For example, the measurement that 5 people obtained is more likely the true length than

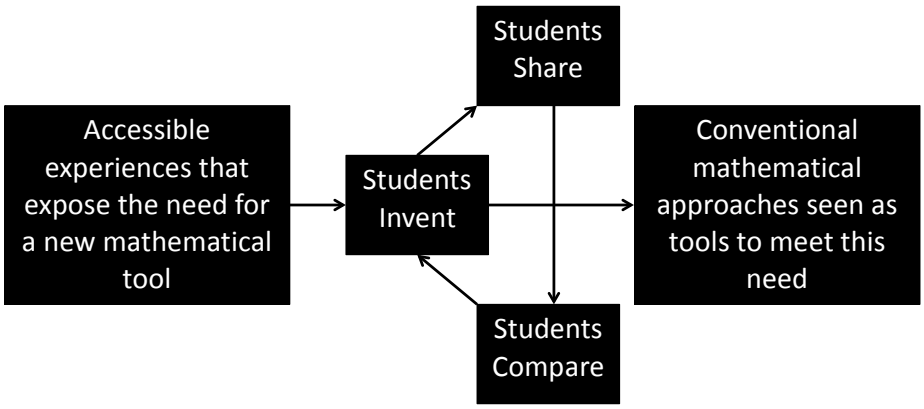


Figure 6: Conceptual logic model for Data Modeling design.

extreme measures that are rarely observed. In other words, the aggregate shape of the data suggests a signal to the true length. However, there is noise around the signal because of the variability in the data. This means there is uncertainty when students (or any measurers!) estimate the true length.

The second unit is designed to support students to formalize their ideas about signal in the form of a measure. The teacher asks students to invent a replicable method, what adults call an algorithm, which generates a number that represents the "best guess" of the true length of the arm-span. The teacher uses the student-invented methods, which typically value either agreement or center, to facilitate a whole class discussion that we call a "measure review." This conversation is similar to the one around displays, but here the purpose is for students to think about concepts of replicability, generalizability, and ultimately the rationale for the canonical methods of mean, median, and mode.

Once students have had an opportunity to think about signal the instructional sequence focuses on variability. Students measure the teacher's arm-span again, but this time with a meter-stick. The more precise tool produces data with a very similar center (signal), but less variability

(noise). The teacher then asks students which tool produced a more precise group of measurements, and they discuss the concept of variability using informal intuitions, such as the "clumpiness" of the data. This helps students develop initial conceptions of variability as an attribute worth noticing, and students use this idea as a resource to invent shareable methods for measuring the "precision" or the "tendency of the measures to agree" of a data set. The teacher facilitates a measure review around these measures to again discuss replicability and generalizability, but now with the intent of understanding the rationale behind the canonical methods of range, interquartile range (IQR), and deviation based variability statistics (Lehrer & Kim, 2009).

After exploring the meaning of their newfound display and statistical concepts in a new data context (manufacturing process), students explore the relationship between theoretical and empirical probability as measures of chance. They use the computer software *Tinkerplots* to run simulations designed to develop a conceptualization of theoretical probability as a trend across many repeated trials of an event. Students also investigate the relationship between sample size and sample-to-sample variability of a statistic by constructing empirical sampling distributions. These activities extend the practices of data representation and statistical measurement to develop the notion of sampling and sample statistics.

In the final two units students use display, statistics, and probability concepts to build chance models that simulate the measuring process they experienced at the beginning of the first unit. Students share and compare their invented models in a "model review" that is designed to support concepts of model intelligibility and model fit. These whole class discussions often lead students to modify their model so that it more faithfully reflects their measurement experience, and so that it produces data that is more similar to the empirical data. After modifying their

models the students use them to generate simulated data to estimate the shape of the sampling distributions of both the median and the IQR. The students integrate all of these ideas to make informal inferences about new data by asking: did these data come from the same population as our measurements? These discussions are designed to support students to consider boundaries in the sampling distributions that distinguish "real" differences from those that they would expect to observe just by chance. While these regions are usually informal, and rarely resemble conventional boundaries (i.e. $p < .05$), the curriculum is designed to provide access to the epistemic concepts and practices that produced such conventions.

In this section I describe how I have used the first two blocks of Wilson (2004) framework to design an observational measurement system that indexes the extent to which classroom interactions resemble the types of interactions described in the previous paragraphs. First I describe the ways in which previous design studies and qualitative research in *Data Modeling* classes influenced my conceptualization of fidelity measurement in this context. Then I explain my thinking during the first two of Wilson (2004) four building blocks and how they informed the development of the *Data Modeling* fidelity observation. I also describe the nature of the data collected during the first year of the larger study. This will set up the last section in which I describe work in the third and fourth blocks to score and model this data for additional validity evidence.

Theorizing Fidelity from Data Modeling Design Studies

There is a recurring structure in this instructional sequence. The teacher introduces a scenario that is intended to provoke the need for a new mathematical tool, students invent strategies to meet the new need, and then the teacher leads a discussion of the strategies (figure 6). Take for example a common measure of center, the median. This is a common student

invention during unit 2 since students often conceptualize it as the middle of all the numbers. As students share this "invention" it is typically given an unconventional term, such as "James's best guess method," but the teacher supports students to formalize the concept into a procedure that could be readily followed by other students. This often leads students to refine idiosyncrasies in their invention, which supports an understanding of reliable procedures to obtain the statistic. In addition to the conceptual and procedural knowledge, the invention and critique are designed to support students in approximating accessible aspects of professional statistical practice. So, they consider measurement properties of invented statistics such as correspondence to target phenomenon and the meaning of the measure's scale. These procedures, concepts, and practices are leveraged when students compare alternative statistics designed with similar goals in mind, such as mean, median, and mode.

While this is the theoretical foundation for much of the curricular design, these principles were operationalized differently in each unit to support specific procedures, concepts, and practices. For example, in the first unit students independently measure the length of a common object and are often surprised to find great variability in the measurements. It is in the face of this variability they are asked to look for patterns or trends in the data that would help them estimate the "true" length of the object. Next, they invent ways to display the data that would lead others to see "at-a-glance" the trend they noticed. Last, the students discuss the different choices made to create the displays, and the effect of these choices on what one could see (or not see) about the data. On the other hand, the second unit is designed to support students to develop an understanding of measures of center. With this in mind the tasks and teacher strategies are aimed at helping kids value the center of a distribution as an indicator of signal, and to formalize

this concept in various statistics (i.e. mean, median, and mode) Appendix A contains unit specific logic models for each of the observed units.

This logic model is very misleading because the arrows suggest that the descriptions in the boxes are carried out in a straightforward manner. In fact, the work of orchestrating classroom instruction to bring about these opportunities takes the hand of a highly trained and skilled teacher. Qualitative analyses during earlier design studies generated knowledge about particular design features and fruitful interactional practices that were critical in supporting the target procedures, concepts, and practices (Petrosino, Lehrer, & Schauble, 2003; Lehrer, Kim, & Schauble 2007; Lehrer & Kim, 2009; Lehrer, Kim, & Jones, 2011). The knowledge generated from this work served as the primary motivation for our fidelity measure. I do not have room to give a comprehensive account of the knowledge generated during the design studies, so here I will address the three primary areas in which previous theory influenced the development of the fidelity measure.

Fidelity Measurment Principle 1: Students must have opportunities to invent diverse mathematical solutions to solve problems that arise from the original task.

The Data Modeling curricular materials and professional development are designed to support the generation of a variety of student inventions in accessible contexts. For example, we recommend using data from a repeated measure context (where students independently measure the same object) to provide interpretable conceptions of signal and noise for students (Petrosino, Lehrer, & Schauble, 2003). Within these contexts teachers can use the variability in invented methods to support students to develop procedures, concepts, and practices related to data and statistics. Lehrer & Kim (2009) described a teacher that leveraged the diversity in student-invented measures of variability, and student critique of these inventions, to support students to

understand a statistic as a measure that produces a quantity that should vary with the state of the distributional quality being measured. This helped students to develop a professional vision towards statistics (Goodwin, 1994). The students' conceptions of variability changed from informal and idiosyncratic ideas to shareable concepts instantiated in their invented-statistics. For example, one student, Shakira, invented a statistic that intuitively valued agreement around the center. In short, she counted the number of data points that were "closest" to the median and to multiple modes. However, the measure itself was composed of idiosyncrasies that made is unreplicable. As her method came into contact with others, and as her peers pushed her to justify her choices, she developed a more explicit structure of variability through its measure. The student-invented measures provided a context in which her and the rest of the class could consider the merits of their choices in relation to the goal of measuring variability. In addition, they developed disciplinary values for measure such as generalizability, replicability, and correspondence to the phenomenon of interest (in this case, variability). This was a development we saw in other *Data Modeling* classes as well (Jones, Lehrer, & Kim, 2012). This principle led us to conceptualize and measure fidelity in a way that accounted for the use of curricular materials and tasks that support the production of variability in student-invented methods. It also led us to document the specific invention types that are fruitful for conceptual growth.

Fidelity Measurement Principle 2: Teachers must facilitate whole class discussions for students to share and compare their invented methods.

Teachers are critical in orchestrating classroom interactions that exploit the resources in the invented methods and make important mathematical ideas visible. Although the tasks support the production of a variety of invented methods, teachers are responsible for identifying the methods that have potential to provoke conversation about particular ideas, and to make

connections among different ideas. Teachers must be able to recognize and sequence worthwhile student inventions, and then support students to connect these ways of thinking to conventional mathematical tools (Stein, Engle, Smith, & Hughes, 2008). This conversation is intended to be an approximation to the professional statistical practice of negotiating the value of novel techniques. With this in mind, the teacher also has the responsibility to support students in developing goals, values, and discourse norms that are productive in collective activities that resemble disciplinary ways of generating and revising knowledge (Forman & Ford, 2006; Forman & Ford, 2013; Horn, 2008). Through the design studies we developed material, conceptual, and pedagogical resources to support teachers to carry out this challenging work (e.g. Lehrer & Schauble, 2007). For example, we provide teachers with examples of the types of discussion questions that can help students consider the conceptual aspects of the methods and also different ways of thinking that these questions will likely evoke from students. We also support teachers to develop strategies for comparing and contrasting the mathematical differences in the invented methods. However, since these teaching strategies rely on student contributions to create a productive discussion the nature of the use of student-invented methods is co-constructed by teachers and students. Teachers are responsible for facilitating interaction, but students must share their thinking and engaging with other students. For these reasons we defined the fidelity measure to account for important teacher discourse moves as well as student contributions during whole class conversations.

Fidelity Measurement Principle 3: Teachers should support students to consider particular mathematical ideas during a whole class discussion.

The instructional design is built to support the development of particular mathematical procedures, concepts, and practices. Decisions about discussion questions, comparisons, and

critique should be made in relation to particular concepts. In many cases, the concepts are unit specific. In unit one the conversations should highlight design principles for data displays, such as scale, order, and grouping. In unit three, though, the conversations should focus on measuring variability. It is important that teachers understand the target concepts, and that they know how to identify them in the various ways students might express them. For example, Lehrer, Kim, & Jones (2011) described a teacher that modified the visible data into a new, imagined distribution to provoke a closer examination of the range as a measure of variability. This teacher understood the merits of the range, and he identified that the students were interrogating it as a measure by asking what would happen to the range under particular data transformations. The teacher noticed this as an opportunity to further develop students' thinking by suggesting data in which extreme values were far apart, but the rest of the data was much less dispersed. Students then were able to think about and discuss the implications of a statistic that only attends to extreme values. Student generated statistics and displays are messy, but they often share conceptual similarities with canonical methods that can be leveraged to support students to engage with the ideas in ways consistent with disciplinary engagement. Lehrer, Kim, & Jones (2011) also described instances in which students identified relationships between student-inventions and visible or imagined distributions to examine the generalizability of particular methods. So, the measurement system also takes into account the extent to which unit specific concepts are explicitly discussed, either by teachers or by students. The measure also accounts for disciplinary practices such as considering the general use of a statistic by imagining new distributions.

The careful qualitative analyses during the iterative design studies were invaluable in the development of this fidelity measure. This measure is rooted in the theories generated over many years of close study. Although the measure cannot index these theories at same grain size as the

qualitative descriptions, it is designed to inform the most relevant parts of the program theory. This led us to define fidelity in *Data Modeling* as:

> *Fidelity Definition: The extent to which student invented methods are used as instructional resources during classroom interactions to support students to represent, measure, and model variability.*

Construct Representation

After defining fidelity to *Data Modeling* I represented this construct as a linear continuum using a construct map (figure 7). A construct map is a representational tool that supports an image of the fidelity construct that is general enough to apply to multiple units, but also specific enough to give meaning to phrases such as "support the development of" found in the fidelity definition. Additionally, it describes a trajectory along which classroom instruction might lie, with the higher levels indicating more fidelity to the program theory. In this map I describe 5 qualitatively different characterizations for classes using student-invented methods. Since we are relying on classroom interactions as an indicator of this construct I also provide general descriptions of the types of interactions we would count as evidence of each level.

I conceptualized this construct as a trait of classroom instructional practice. Since teacher practice is interactional we committed to measure it during interactions. This measure should not be viewed as a description of teacher quality. Teachers are not alone in this practice since students' contributions are a necessary requirement for instructional interactions to take place. Just like many curricula that are designed to be responsive to student thinking, it is impossible for a teacher to faithfully carry out instruction absent the necessary student productions and participation (Cohen, 2011). However, students are not trained during summer professional development workshops like teachers are. This is why the tasks in this curriculum are designed to provoke and elicit particular types of participation from students. A teacher's practice is

socially situated in relation to particular students, schools, and educational infrastructures (like this project), which all influence practice. This measure is designed to account for the interactional practices that occurred in particular settings. Teacher quality, no doubt, is a significant influence on the instructional quality of a classroom, but it is not the only influence. This measure describes the extent to which students experienced classroom interactions that resembled the instructional principles Data Modeling supports, but is not able to disentangle the various factors that contributed to the instructional quality.

Level	Description	Observable Classroom Interaction
5	Student invented methods are seen as a resource to communicate different mathematical strategies in order to synthesize specific mathematical ideas into systems of epistemic meaning.	Conversation includes the description of level four, but goes on to make connections among the different mathematical ideas in terms of the epistemic work they do. For example, students might discuss what different display strategies show and hide about the data.
4	Student invented methods are seen as a resource to communicate different mathematical ideas and to begin to use the ideas as epistemic tools	The conversation focuses on important mathematical concepts and students begin to use the concepts as epistemic tools. For example, students might talk about the order of the data while discussing what ordering decisions show about the data.
3	Student invented methods are seen as an instructional resource to promote key concepts.	The class discusses important ideas about the invented methods, but does not treat them as epistemic tools. They may talk about them procedurally, or use them as labels to describe other invented methods.
2	Student invented method are seen as an instructional resource to support student discourse.	Methods are invented and presented, but the mathematical components of them are not highlighted or discussed.
1	Student invented methods are seen as an instructional resource.	Teachers make use of tasks in curriculum that support student invention and thinking.

Figure 7: Fidelity construct map

This construct map describes five theoretical roles for student-invented methods, and orders them in increasing similarity to our ideal. The first level describes a state in which teachers use Data Modeling tasks and materials to support the generation of student-invented methods. At the second level, in addition to using tasks and tools to generate invented methods, teachers provide opportunities for students to talk about their invented methods. Next, student-invented methods are used as a resource for supporting a discussion of worthwhile mathematical concepts. The fourth level describes a state in which classes begin to discuss the mathematical concepts as epistemic tools. At the highest-level classes talk about the concepts as epistemic tools, consider the tradeoffs between different key concepts, and use these concepts to approximate disciplinary practices such as considering the generalization of an idea or method.

Notice that this map is intentionally cast at a high level of generality so it can describe fidelity across the curriculum units. Core mathematical concepts and practices vary by unit, so I also articulated construct maps for the specific roles of student-invented methods in each unit. These retain much of the same language, but I use them to represent the specific mathematical concepts I account for in each unit. Appendix B contains unit specific construct maps for Units 1-5.

Also notice the nature of the descriptions in the middle column of the construct map. These are theoretical states that cannot be directly observed. In the far right column I have described observable classroom interactions that I would count as evidence of the different theoretical states of the construct. These descriptions are similar to the Hall & Loucks (1978) "Levels of Use." So, this is different than a rubric or a scoring protocol because the descriptions of observable evidence are mapped onto theoretically different states of the underlying construct. The rationales for the observable states are grounded in the levels of the latent construct.

An illustration is needed to better communicate the meaning of each level and the differences between the levels. I am going to build this illustration by contrasting unit 1 whole class discussions from two different classes. The teacher in the first class is Ms. West. She participated in both years of this project and served as a case study for Mayumi Shinohara's research (Shinohara & Lehrer, 2013). Mayumi collected video records during all Data Modeling instruction in Ms. West's classroom. For this illustration I have selected excerpts from the class's unit 1 display review from the first year of the study. The teacher in the second class is Mr. North. Mr. North participated in previous Data Modeling deign studies and Min-Joung Kim collected video records of all Data Modeling instruction during the 2009-2010 school year. I selected these excerpts from the class's unit 1 display review.

Both of these classes provide evidence of the first level of the construct map. The students had opportunities to invent data displays during previous days of instruction and the wide variability in student products suggests that neither teacher overly structured the invention task to replicate conventional data displays. In addition, both classes used data that the students collected in a repeated measure context. The excerpts below illustrate that invention is not sufficient. The discussions in these classes differed in terms of the mathematical content and the epistemic nature of the students' comments. I have selected excerpts from the transcripts of the whole class discussions to illustrate these differences and to relate them to the construct map.

Ms. West

The whole class discussion below occurred two days after students invented their data displays. During the previous day of instruction the students went though a "gallery walk" where they spent time at each invented display, and each student took notes on things that were similar and different to their own display. Ms. West selected three invented displays for the class to

discuss during this conversation. I have selected transcript from two moments in class that illustrate the nature of the conversation in this class. The two sections are time stamped to show when they occurred in the class.

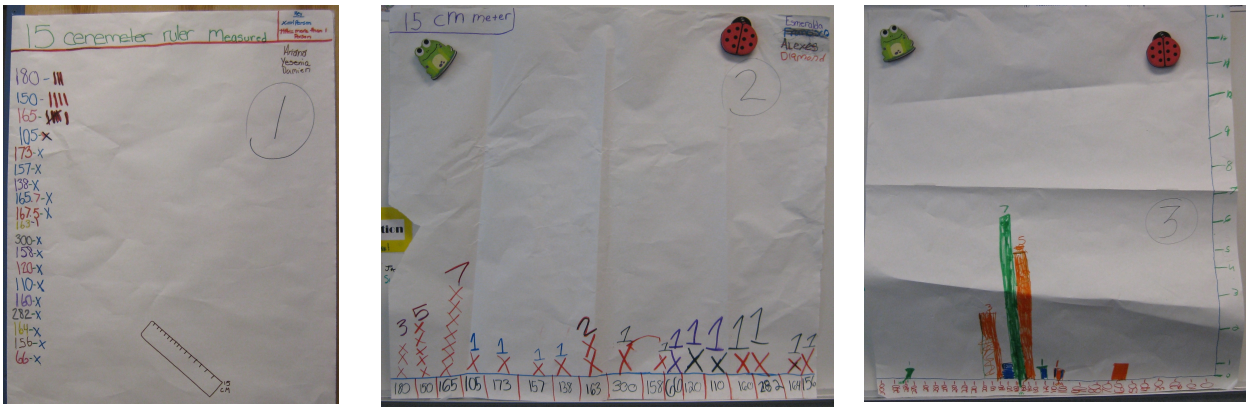


Figure 8: Three displays selected for whole class discussion by Ms. West

[00:03:22.07]
Ms. West: So, between graph one and graph two. What are some things that they have that are similar? Some things they have in common? Juan, what is one of them?
Juan: Um, they both use Xs.
Ms. West: They both use Xs. What else do they have the same? Othello?
Othello: Um, they both have how many there is on each number.
Ms. West: So they both say how many on each number. So, how many people got 180, that one says three this one says three. Athenia, what else do they say?
Athenia: They are both in the same order.
Ms. West: They are both in the same order. Order means how they are arranged, so they are both in the same order.
Student: No, that one has 66 in the middle, and the other doesn't.
Ms. West: For the most part they are both in the same order. Andrew, what else?
Andrew: It shows all the people's measurements.
Ms. West: OK, so on this one it did show all the people's measurements. What else? How are they the same? Othello.
Othello: They both have titles.
Ms. West: OK, they both have titles. Andrew, what else?
Andrew: They are both doing the 15 cm ruler.
Ms. West: So they're both using the 15 cm ruler. I have a question. If I used a 15 cm ruler and this graph was a meter stick, so you think they would both look the same?
Students: No
[00:05:23.07]

The three data displays in figure 8 were hanging at the front of the class and were numbered 1, 2, and 3 from left to right. During this segment of talk the class focused on a

comparison between #1 and #2. Notice that Ms. West was attempting to support the students to compare similarities between the two. She asked "What do they have in common?" and students responded with similarities between the two. Ms. West was making an attempt to support students to talk about the inventions by selecting inventions to serve as the focus of the conversation and by using a question similar to one in the curriculum guide. However, the class primarily discussed similarities between the two that were not related to the display concepts of scale, order, and grouping, even though the curriculum guide suggested questions to highlight these ideas. Instead, they focused on more superficial aspects of the displays like titles. Although one student did bring up the idea of order, and another challenged her statement, the teacher quickly moved on to another student. In this short segment this was the closest moment to a true "discussion," and the only time students referenced a similarity related to the goal of the discussion. The second level of the construct map is meant to describe this type of discourse where students are beginning to talk and the teacher is using curricular resources to support a conversation, but the discussion is not fruitful to produce conversations about the target mathematical ideas. The next excerpt is from later during the same day.

[00:35:55.26]
Ms. West: So, for example, I'm going to let you see mine so I can show you all four of them in mine. So, for example, if you're talking about frequency. I have frequency here because you can tell how many got the same score. I have order, because mine goes from 50 to 300 in order. Mine has bins because I put mine into groups, 50 to 75. So how many are in each group if it's 50 to 75?
Student: A bunch?
Andrew: 25
T: Thank you Andrew. And then scale. Mine shows scale because it shows that there's holes such as here. This would be a hole. Nobody measured in between 76 and 100. So this graph shows all four. Expectations, you are going to take the notebooks in front of you and you are going to go on a gallery walk. You are going to start at your graph. You are going to tell me what does your graph have? Does it have frequency, order, bins, scale? Four words, that's all I need for number one if it has all four. Then we're

going to rotate to the next one. You're going to go to the next one and you're going to see whether it has frequency, bins, scale, order. And we are going to do the whole entire thing.
[00:37:59.20]

In between the first excerpt and the second the students discussed the invented displays without much reference to the data display concepts we intended the conversation to focus on. The second excerpt illustrates Ms. West's strategy for supporting students to see these ideas. She began by listing four words on the board: order, frequency, bins, and scale. She then showed her display and how she made use of these ideas. However, the concepts were treated as labels instead of epistemic tools to show (or hide) things about data. In fact, the next task for students was to use the labels on each invented display. This illustrates level three of the construct map. The class talked about relevant concepts, but didn't discuss what each shows or hides about the data. The ideas didn't serve an epistemic role in the conversation.

Mr. North

This whole class discussion occurred the day after students invented their displays. Mr. North selected the inventions to talk about and hung them at the front of the room as the class discussed them. I have selected excerpts from the conversation to illustrate how the idea of grouping was initiated and developed during the conversation.

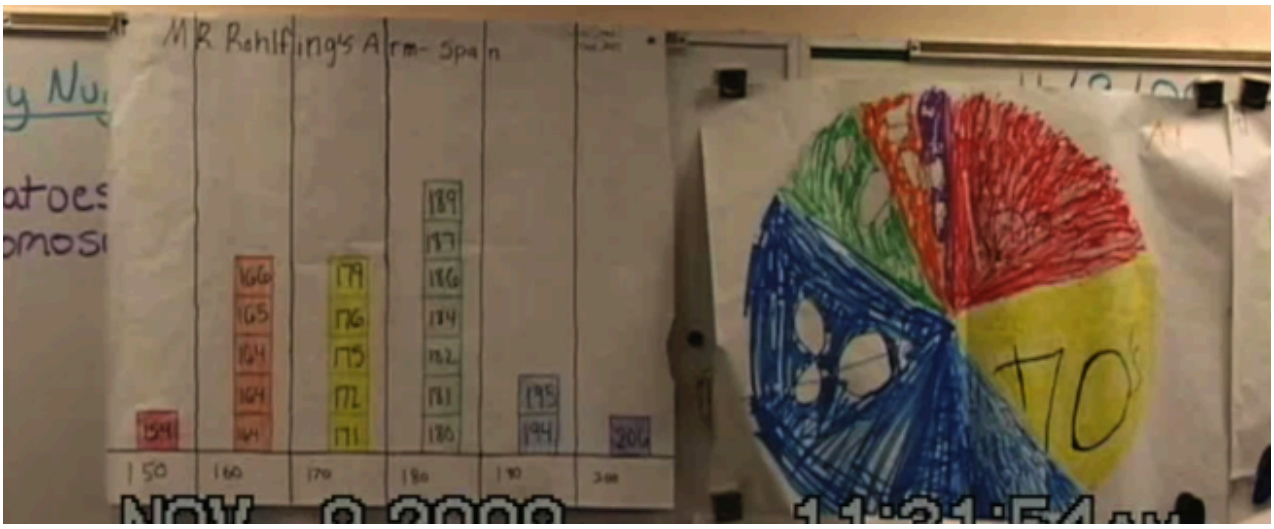


Figure 9: Displays being discussed in Mr. North's class.

[00:23:29.50]
Mr. North: What did you guys want us to notice about that graph *((Graph on left in figure 9))*?
Student: Um, we wanted you to notice that we put the number of guesses in with, like, the tens. Estimates of tens. Like, there's 1, 2, 3, 4, 5, 6, 7 estimates for, around in the 180's, so we wanted people to see that.
Student: And that it's higher than the other ones.
Student: Yeah
Mr. North: Is that easy for people to notice?
Students: Yeah
Student: It's sort of like the pie graph, but it's more of a graph-graph.
Mr. North: Okay, so let's compare the two graphs. What does one graph do better than another? What do you like about one graph? Because you're right, they are very similar. But what does one graph do compared to another graph? Barbara?
Barbara: Their graph *((Graph on left in figure 9))* shows all the numbers that are in the 180's.
Mr. North: Okay, so each specific number, which is what Kerri wanted, and if they had time, they were going to do, right? So, showing each individual number and still being able to see that yup, 180's are the most.
[00:25:00.20]

Mr. North begins this excerpt by asking the authors of the invented display what they wanted the class to notice. The authors indicate that they wanted readers to see that there were more measurements in the 180s than in any other decade. They described that readers could see this because it is the highest, and that they could count to see how many measurements are in the group. Mr. North engaged the rest of the class with this idea by

asking, “Is that easy for people to see?” This provoked one student to notice a similarity between two of the displays on the wall by stating that it was “sort of like the pie graph.” In contrast to the conversation in Ms. West’s class, this conversation brought out a similarity that was related to the concept of grouping. In fact, the two displays the students were comparing in Mr. West’s class did not share many superficial similarities. They were conceptually similar rather than visually similar. Notice how Mr. North continued to use the students’ comments to engage the rest of the class is a discussion of the idea. He did this by asking if others agreed, by asking probing questions such as “what does one graph do compared to another graph?” and by connecting the comments to students’ comments from earlier in the conversation. In this short excerpt these students were discussing conceptual similarities between two displays with an eye towards the epistemic nature of the choices. The pie graph hides the individual measurements while showing relations between groups of measurements. The “graph-graph” shows the same relations between groups while retaining the information about individual measurements. Level 4 of the construct map describes this type of whole class discussions where classes are talking about mathematical concepts as epistemic tools that can tell them something about their data. The next excerpt is from a few moments later during the same class.

[00:26:34.20]
Mr. North: So, they put them in bins, and they chose ten. You guys chose ten. And, I'm curious as to why was ten the lucky number that got chosen for these two graphs and I think I asked this group, do you guys have any reason why you, was there any reason to choose ten? Why not 5? Or 2? Or a bin size of 20?
Student: Cause, then you'd have to make more bars and it would be more time consuming.
Mr. North: Ok.
Student: And it's easier to read. It's also kind of, um, with the numbers, with the amount of numbers that you have.
Mr. North: Mhmm.
Alan: And the smaller range that's in there, I think ten is the best number to choose because it's easier to see the range and stuff.

Mr. North: When you say range, do you mean range between this number and this number? *((Teacher points to the extreme values of the graph on the left))* Or range between this number and this number? *((Teacher points to maximum and minimum values within a bin on the graph on the left))*
Alan: No, like...
Mr. North: Like the range is 150 to 200, or the range inside the bin?
Alan: Well, more like, which one has the most numbers.
Mr. North: Ok. Martha?
Martha: Um, sort of like what Paul said, like if you said, like if they do a bin of 20, that would, it's more, it would have more numbers in it, and 10 is not too little like 5, but it's not too big like 20, or 40 or even 50.
Mr. North: Mhmm.
Martha: Cause then it would just be a lot of numbers and it wouldn't even make sense, because, like if you did a bin of 200, it would be...
Mr. North: Everything's in one bin, right? All you did was just stack everything...
Martha: It wouldn't really show you anything. It would just show you the numbers.
[00:28:24.12]

Mr. North uses the conversation about grouping in the invented displays to support students to imagine alternative grouping choices and the effect on the shape of the data. Notice that the initial rationale for a bin size of 10 was for efficiency. The student said that if the bin sizes were smaller then the creator would have to draw more bars. However, the next student's comment emphasized the epistemic nature of the choice. After the teacher worked to clarify the student's thinking for the class by questioning the meaning of his use of the word "range" it became apparent to other students that the choice of bin size has implications for what readers can see about the data shape. They went on to illustrate the idea by imagining an absurd bin size, 200, and describing the change in the shape of the data. The last student comment exemplifies the epistemic nature of the conversation. She was not concerned by efficiency, but said that the absurd bin size "wouldn't really show you anything." This conversation illustrates the highest level of the construct map. These students were talking about the mathematical concepts in a way that built a "larger system of epistemic meaning."

The construct map can be thought of as descriptions of milestones along a trajectory. This is helpful for thinking about how teachers might "move" from the lower levels to the higher ones. Initially teachers might make use of the tasks and materials, and even work to engage students in talking about the invented methods. However, it often takes time for teachers to begin to understand the intentions of the discussion questions and to be able to recognize worthwhile student ideas. It takes even more time to know how to ask questions that provoke students to consider the epistemic implications of the ideas. The conversations in Mr. West's class didn't happen by accident. Just before many of the student comments are thoughtful and timely questions that provoked a new way of seeing the displays. There is a second way, though, that this construct can be thought of as a trajectory. It is a helpful framework for thinking about the trajectory within a whole class discussion. The class first engages with the invention tasks, then the teacher begins to unearth the concepts in the student products using fairly open ended questions, and moves to the higher levels when the students begin to see and discus the epistemic nature of the inventions. This is the fidelity trajectory for this project, and one way of thinking about the purpose of the measure is to describe "how far" along the trajectory the whole class discussions were.

It's important to remind the reader that this is not meant to be a judgment about teacher quality. It certainly takes a highly skilled teacher to facilitate conversations that are high on the construct map. However, many other factors are in play too. For some classes these ways of talking are very similar to discourse in their homes, while for others it is a very new interaction. The leaders in some schools might encourage these kinds of conversations while other administrators discourage them. These two excerpts serve to illustrate the construct map, but they show very little about the contexts that produced the conversations.

For Ms. West, the excerpts were taken during her initial attempts to drastically change her teaching practice in a school that is committed to developing student discourse. However, these kinds of conversations were very new to Ms. West's students. Mr. North was new to these curricular materials, but had collaborated with the designer of the Data Modeling materials for a number of years. In some ways his students were more familiar with these kinds of conversations, but he still had plenty of work to figure out how to make the discussions productive.

Construct Operationalization

Since the whole class discussions provided the most information about the use of student inventions I could not observe a number of arbitrary lessons during the course of implementation. Instead, I needed to schedule observations during moments where the instructional role of student-invented methods was most evident. This led me to observe the days when students were sharing, comparing, and discussing their invented methods in the display, measure, and model reviews. In figure 10 I represent the rationale for this choice. Unit 1 has four distinct phases, 1) measuring an object, 2) inventing data displays, 3) sharing and comparing invented data displays, and 4) assessment of student learning and continued support for learning. Figure 10 summarizes the main activities students and teachers engage in to carry out these interactions. The excerpts from the Ms. West and Mr. North took place during phase three. Phase three provides the most information about the construct map since we can observe interactions between teachers and students about the displays, and can determine which of the display concepts the class explicitly discussed. Another advantage of observing during these moments is that I can record information about previous phases. For example, I can infer from the data

students are using if they had opportunities to measure in phase 1 and if they had chances to invent in phase 2.

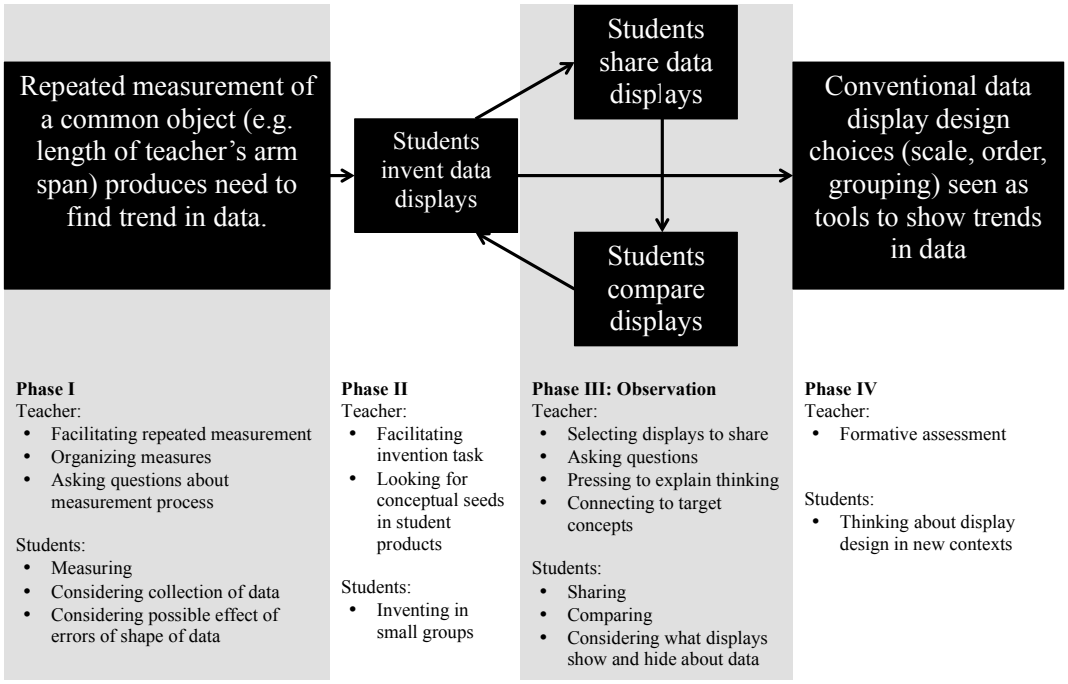


Figure 10: Rationale for observing during display review.

I designed and planned for one classroom observation for each teacher during unit 1, unit 2, unit 3, unit 5, unit 6, and unit 7. I did not observe unit 4 in the first year of the study because in this unit students used the recently developed statistics with new types of data. This was revised in the second year so that Unit 4 now emphasizes new kinds of processes that are also characterized by signal and noise. I observed unit 4 in the second year, but since there were no comparable observations from the first year I did not include these in the fidelity scaling.

Since the unit of analysis was one class meeting I designed variables to index three relevant aspects: digital images of students inventions, summary variables, and 5-minute segment variables. Figure 11 is an image of the three types of observable variables. The images of student invented methods provided data on the kinds of products students created. However,

they do not contribute to the quantitative fidelity scale. For the remainder of this section I describe the theoretical rationale for the variables and the scoring rules used to create fidelity items from these variables.

Summary variables

The classroom level summary variables are useful for indexing the structural aspects of the program, such as use of tasks or data context. The variables primarily identify the existence of a particular task or curricular tool. They do not judge the extent to which the tools were used in the ways that are faithful to our purposes. Hence, these variables primarily provide information about the lowest levels of the construct map. I have listed all of the summary variables in Appendix D, but I only used a subset of them as part of the fidelity scale. I will describe these and the rationale for selecting them later when I discuss the scoring rules.

Segment variables

To index the quality of the whole class discussions around the curricular tools I needed to account for the interactional nature of the conversation. It was not feasible to index every interactional turn during live observations, so instead I created "segment" variables which observers score in 5-minute adjacent segments for the duration of the class. These variables are binary, and observers scored them if they were observed at least once in a five-minute segment. For example, if one student contributed a comment on the conceptual aspects of a student-invented method then the sInvented and sProcedural variables were scored. So, for each five-minute segment we can tell if the variable was observed at least once. This provides some rough temporal and frequency information while also making the live observation manageable. Segment variables fell under three categories, student contributions, teacher practice, and unit

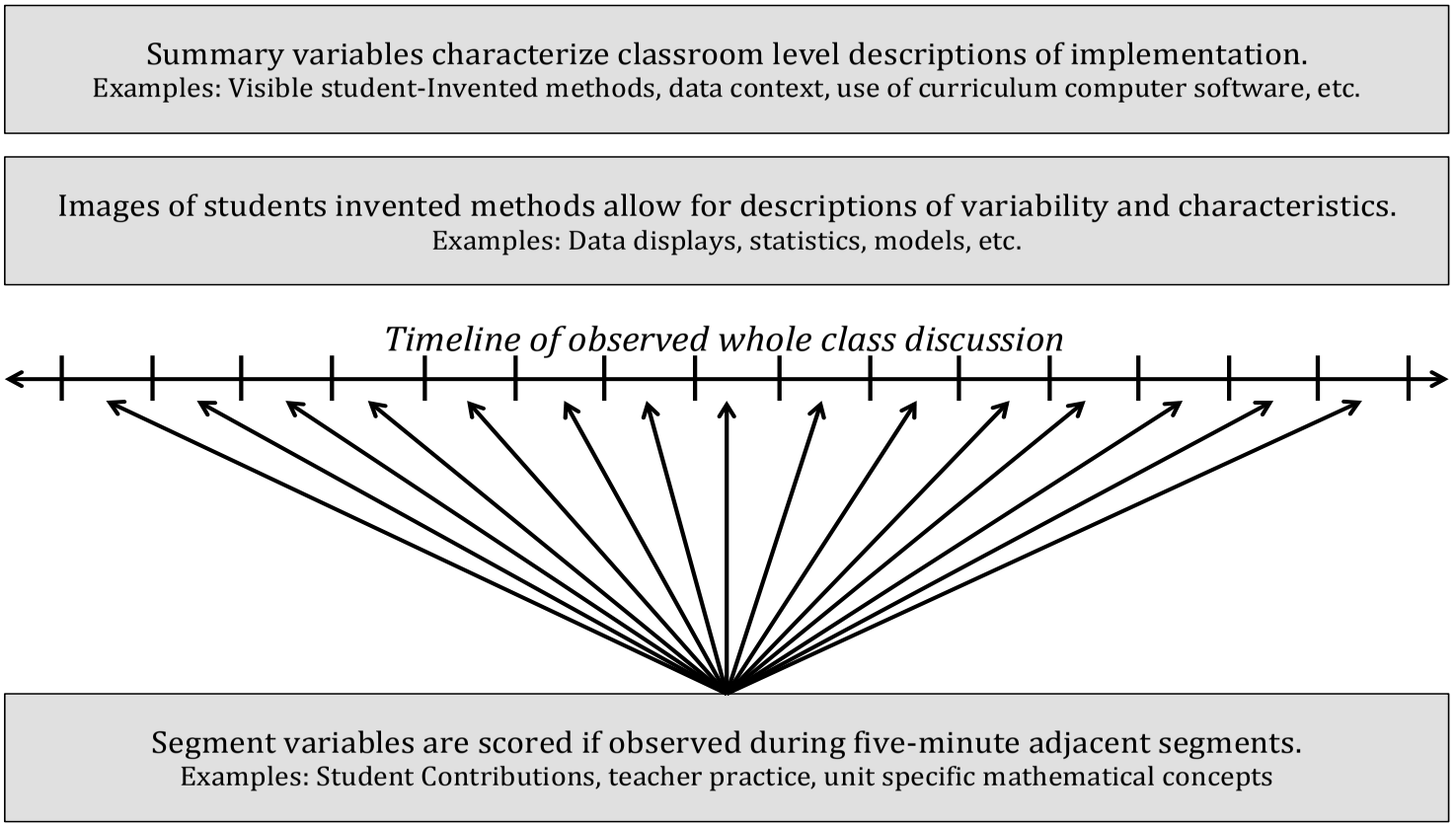


Figure 11: Image of observable variable system

specific mathematical concepts. Table 1 shows the segment variables for unit 1. The teacher and student variables are used for all of the units, but the mathematical concepts are unit specific. See Appendix C for the full set of segment variables and their corresponding units.

The "student contributions" variables account for three ways students can contribute to the whole class discussions. These three variables only provide a very small amount of information about student contributions, but the information they do provide allows for some very rough distinctions about the lower levels of the construct map. The sInvented variable indicates that students were discussing invented methods. This variable was scored any time the

Table 1: Display review segment variables

	Observable Variables	
	Name	**Description**
Student Contributions	sInvented	Did students discuss invented displays?
	sProcedural	Did students make comments or ask questions about the **conceptual** elements of the invented displays?
	sConceptual	Did students make comments or ask questions about the **procedural or calculational** elements of the invented displays?
Teacher Practice	tInitSelect	Did the teacher select student-invented displays to be shared?
	tCompare	Did the teacher compare different ways to display data?
	tDiscussQ	Did the teacher use questions similar to the ones in the curriculum to support students to think about and discuss displaying data?
	tConnectOthers	Did the teacher make connections between different students' thinking?
	tConnectBigIdeas	Did the teacher connect student thinking to the big ideas?
	tPressExplain	Did the teacher press students to explain their thinking?
Unit Specific Mathematical Concepts	iOrder	Was the order of a display talked about?
	iScale	Was the scale of a display talked about?
	iGrouping	Was the grouping in a display talked about?
	iShape	Was the effect of design decisions on the shape of a display talked about?
	iShow	Did the teacher and/or students talk about what one or more displays show about the data?
	iHide	Did the teacher and/or students talk about what one or more displays hide about the data?

whole class discussion was related to student inventions. The sProcedural and sConceptual variables provide information about how the students were talking about the invented methods. The sProcedural variable indicated that students were talking about the procedural aspects of a mathematical idea or method. I don't intend for this variable to be seen in a negative light, but as an important part of any conversation about a mathematical method. Observers scored this variable any time students were talking about *what* was done or *how* it was done. For example, this would be scored when students in unit 1 talked about how they created their displays. On the other hand, observers scored the sConceptual variable when students talked about *why* something was done a particular way or when they talk about a conceptual implication of a procedural choice. For example, this would be scored when students discussed what a design choice in their display showed or hid about the data.

The "teacher practice" variables account for a number of productive strategies for facilitating whole class discussions. One of the first challenges for teachers is to select from the wide variability in student strategies that hold potential for productive conversation, and in what order to discuss them (Stein et al., 2008). Observers scored the tInitSelect variable when teachers deliberately selected which student invented method to focus the conversation on. Of course, during a live observation it is very difficult to determine if the selection and sequence are productive during a 5-minute segment of time. Observers scored this variable very generously in order to maintain a reliable scoring protocol. Teachers also must work to provoke and elicit students' ideas about the particular invented methods under consideration. This effort serves many purposes. First, it provides information to the teacher about student thinking. It also allows students in the class to hear how their peers are making sense of the ideas, and to compare the different ideas (Cobb, Wood, Yackel, & McNeal, 1992). It does more than support conceptual

understanding, though. It develops a collective practice that generates and refines knowledge about the mathematical ideas (Lehrer & Lesh, 2003; Horn, 2008). Observers scored tDiscussQ variable when teachers used discussion questions to provoke and elicit student thinking. Like the previous variable, it is very difficult to determine how productive a particular question is within a 5 minute segment of time, so this was scored generously too. Observers marked it if the questions appeared to be aimed at supporting student discussion. Not all questions were scored, though. Questions that were aimed at evaluating a closed ended response did not count. For example, if a teacher asked, "what color is the 150s data?" observers would not have scored the variable. However, if the teacher followed this by asking, "why do you think this group colored the data in the 150s group this color?" the observers would have scored it. Often students need teacher support to explain their thinking in a way that communicates to others the conceptual ideas they are considering. Observers scored the tPressExplain variable when teachers asked additional questions to press kids to further elaborate their thinking.

After working to elicit student thinking teachers have to work to compare the different ideas that highlight the target mathematical concepts of the lesson. The tCompare variable looks for moments when teachers are juxtaposing two different ideas during the discussion. The tConnectOthers and tConnectBigIdeas variables are designed account for moments when the teacher is connecting different, but related student ideas (tConnectOthers) or making connections between students' ideas and the conventional mathematical concepts (tConnectBigIdeas).

Student discussion is not the end goal in these conversations. Teachers have the responsibility to do more than just engage students in a conversation, they need to engage them in a conversation about worthwhile mathematical ideas (Ball, 1993; Chazen & Ball, 2001; Lehrer & Lesh, 2003; Stein et al. 2008). Each unit in the Data Modeling sequence is designed to support

particular ideas and practices around data, statistics, chance, and modeling. For this reason the segment variables focused on these ideas are specific to each unit. Table 1 shows the unit 1 variables that index discussions that focus on the target mathematical concepts (Appendix C has all variables for all units). Observers score these variables regardless of the person talking about them (student or teacher) and regardless of the duration of the focus on the idea. Some of these variables, such as iOrder, iScale, and iGrouping, focus on particular mathematical concepts. For example, in unit 2 the concepts are measures of center such as mean, median, and mode. Observers scored these variables when the *idea* was being discussed, so many times they were scored even if the word labeling the variable was never uttered in the class. For example, students often talk about "binning" their data in unit 1. This is an example of a time when observers would score the iGrouping variable, even if the class did not use the word "grouping." In unit two a common example is found when students discuss invented measures that focus on which data points were measured the most often. Observers scored iMode during these conversations.

Some of the unit specific variables focus on how the class discussed the concepts. For example, in unit one we want kids to discuss what the different design concepts (order, scale, grouping) show and hide about the data. This is the motivation for the variables iShow and iHide. In units two and three we want students to be treating the statistics as measures of distributional characteristics. I created variables that index moments when members of the class are making correspondences between the statistics and the data that produced them, or with new imagined data. I also designed variables to account for discussions of a statistic's replicability and generalizability.

Instrumentation

Wilson (2004) includes instrumentation within the item block. However, the instrumentation required to collect the data described above is very different than what is needed for a typical student assessment. In fact, our experience supported Pickering's assertion (1995) that the machines we build have agency to influence our theories. For me, technology had both affordances and constraints that significantly influenced the ways I was able to index the program theory.

I used the *Filemaker* software and the iPad tablet to instrument our observation measure. This step was anything but trivial, and as I mentioned before, changed many of our measurement plans. For example, even the size of the screen limited the number of variables observers could score while the ability to automate five minute scoring segments provided an opportunity to collect more fine grained information. With the help of Chris Hancock, a well-known software developer in the education community, we developed an interface that organized the data structure, recorded variables, automated five-minute intervals, and allowed for the collection of images. In addition, it provided an infrastructure to support data collection in the field and to electronically transmit it to the master database for the project. In our circumstance, project staff conducted observations over 1,500 miles away from campus, but with these tools we were able to electronically transmit data so that we could view it within minutes of the observation. Figure 12 provides a visual example of the interface and the relationship between the components.

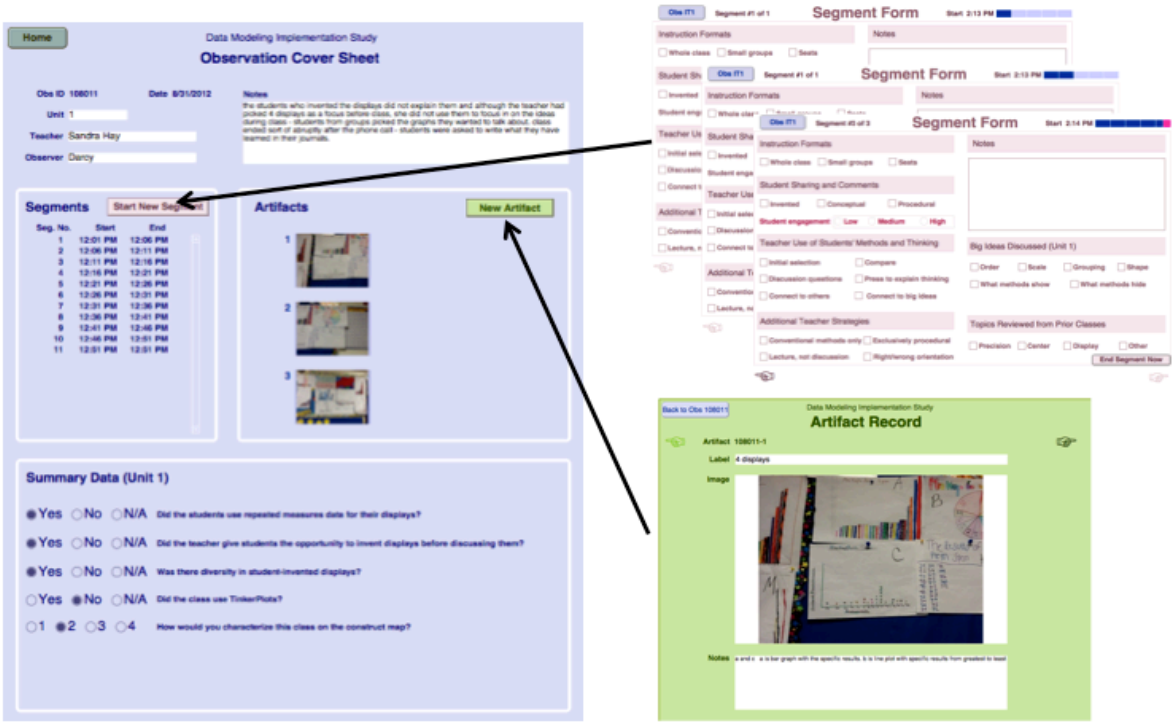


Figure 12: Fidelity observation instrumentation

Data Collection

This measure was used during a two year randomized field experiment testing the efficacy of the Data Modeling curriculum materials. During year 1 21 teachers participated in the Data Modeling professional development and used the materials in their classes while 20 teachers served as comparison classes as they continued with "business as usual" instruction. During the second year 39 teachers used the Data Modeling materials and PD, with 40 teachers serving as a comparison group. After both years of the study all comparison teachers received the Data Modeling PD and materials to use in their classes.

During the course of this project I trained and oversaw an observation team that collected the data I am using for this analysis during live observations. During the summer of 2012 I conducted a weeklong training for 8 classroom observers. During this training we

discussed the program theory, the curriculum units and the fidelity measurement system. I conducted the training concurrent with the teacher training so we could go observe while the teachers rehearsed their strategies for conducting the whole class discussions. We also independently scored videos from classrooms using the Data Modeling materials and compared our scores to judge reliability. Observers were required to agree with anchor scores at least 80% of the time when scoring these videos in order to conduct live classroom observations. I also led ongoing training for four different Saturdays across the school year. During the weekend training meetings we would discuss issues arising from the ongoing observations and practice the observation variables for upcoming units (especially the unit specific variables). During the first year 5 of the 8 observers resigned.

During the summer of 2013 I conducted another weeklong training with a new team of observers. Three members of the previous observation team participated in this team, and one of them served as an anchor coder and team leader. The training for the second year was very similar to the first year. However, we had a more thorough video record from the previous year, and we used this record to conduct more reliability coding than in the first year. All observers were required to meet the 80% benchmark again, which resulted in two of the new members resigning because of their inability to meet the required agreement (they did, however, continue to work on other aspects of the project). During the second year we also conducted random double observations to maintain an ongoing analysis of our agreement. The double observations in live classrooms typically produced more agreement than the videos, so there were no observers that met the original benchmark but later failed to agree at or above 80% of the time during live observations.

Table 2: Observation totals

	Year	Unit 1	Unit 2	Unit 3	Unit 5	Unit 6	Unit 7
Control	1	10	16	3	N/A	N/A	N/A
	2	5	5	0	N/A	N/A	N/A
Treatment	1	21	17	17	15	12	7
	2	39	33	33	24	11	5
	Total	75	71	53	39	23	12

Table 3: Sets of observations

	Year	Units 1-2	Units 1-3	Units 1-5	Units 1-6	Units 1-7
Control	1	8	2	0	0	0
	2	5	0	0	0	0
Treatment	1	17	16	14	10	6
	2	32	32	24	9	5
	Total	49	48	38	19	11

The first year observation team conducted a total of 118 classroom observations from both treatment and control classes during the 2012-2013 school year and the second year team conducted 155 observations during the 2013-2014 school year. These observations include data from units 1-3 and 5-7. Although observers conducted 22 unit 4 observations during year 2 I am not including this data here because it was not collected in year one. The observations from units 1-3 include both control and treatment classes in order to inform differences in classroom practice when teaching these concepts. Since we saw a significant difference in classrooms during year 1 we only collected control observations from a very small number of classes in year two to confirm the difference we saw in year 1. Our research team identified lessons in control classes that were focused on Data Modeling concepts (data display for unit 1, measures of center for unit 2, and measures of variability for unit 3). However, there were no comparable lessons for

units 5-7 in the control classes, so we were not able to construct a comparison. Table 2 summarizes the frequency of observations for each unit for both years. For the treatment group, we observed units 1-5 in significantly more classes than we did in units 6 and 7. This was primarily due to the fact that teachers scheduled units 6 and 7 to be taught after statewide testing, but many teachers did not complete these two units due to end of year schedules. Table 3 gives frequencies for different sets of observations.

Through the process of observing large numbers of classrooms during the first year there were inconsistencies and challenges that emerged with some of the segment variables. In fact, there are a number of variables that I eliminated from the measurement system for these reasons that are not represented in this dissertation. For example, I attempted to capture student "engagement" in the discussion using a rough three point scale (low, medium, high), but found that after much effort I was not able to define these in a meaningful way to provide reliable information. In addition, the first year variables for teacher practice included four variables that observers were to score if an entire segment was dominated by the practice (such as lecturing). However, we found that having two different criteria to score the variables ("if it is seen once" for some and "if the segment is dominated by it" for others) was too difficult to manage during observations so I eliminated anything that did not fall under the "if it is seen once" rule. In addition to these, I had to eliminate two segment variables shown in table 1, sInvented and tConnectBigIdeas, from this analysis. These two variables are still a part of the measurement scheme, but are not included in this analysis because I refined the definition of the variables between year 1 and year 2 in order to better characterize the construct map.

I originally defined the sInvented variable to index students talking about their own invention. I imagined that this would allow me to see how often students were sharing their

inventions during the discussion. Many of the teachers, though, used a strategy that I did not anticipate. They left the authors of particular inventions anonymous. So, in a class it was impossible for our observers to determine when a student was talking about their particular invention. This meant that we were not indexing many moments when students were talking about student invented methods because the variable was defined to exclude any talk that did not include a students clearly describing their own invention. Between the first and second year I changed the definition of this variable to index any moment when students were talking about student inventions. With this change, the variable was scored significantly more in the second year than in the first year. While this is an important variable to include in the measurement system in the future, it is not informative in this data because of the difference in coding.

I originally defined the tConnectBigIdeas to be generously applied when teachers and students made connections between the students' ideas and the mathematical concepts. However, under this definition we were not able to score the variable reliably. So, in the second year I narrowed the focus of the variable to index when teachers explicitly made a connection between a particular student idea and the conventional mathematical concepts. This change allowed us to score the variable reliably, but it also reduced the frequency of the code. We collected very little data during the first year of the study to inform rater reliability. This is a weakness of the study, and was a product of the challenges of scheduling observations during particular moments in the Data Modeling sequence. During the first year of the project there were a large number of teachers that taught the method reviews on the same day because the project coaches were in town. In the second year these lessons were more spread out and we were able to conduct 23 double observations across the 7 Data Modeling units. We paired an observer with one of the two anchor observers on the team for these observations. The two would score the class

independently during the lesson and would compare their scores immediately after the class. These conversations focused on disagreements in scoring and the rationale behind each observer's scores. Each observer then sent their data, without altering it, to our database and I would calculate the percent of segments where the observers agreed for each variable.

Item Scoring and Outcome Space

Table 4: Relationship between unit 1 segment variables and construct map

	Observable Variable	**Construct MAP Level**				
		1	**2**	**3**	**4**	**5**
Student Contributions	Did students share-invented displays?		X			
	Did students make comments or ask questions about the **conceptual** elements of the invented displays?				X	
	Did students make comments or ask questions about the **procedural or calculational** elements of the invented displays?			X		
Teacher Practice	Did the teacher select student-invented displays to be shared?				X	
	Did the teacher compare different ways to display data?				X	
	Did the teacher use questions similar to the ones in the curriculum to support students to think about and discuss displaying data?		X			
	Did the teacher make connections between different students' thinking?					X
	Did the teacher connect student thinking to the big ideas?				X	
Unit Specific Mathematical Concepts	Was the order of a display talked about?			X		
	Was the scale of a display talked about?			X		
	Was the grouping in a display talked about?			X		
	Was the effect of design decisions on the shape of a display talked about?					X
	Did the teacher and/or students talk about what one or more displays show about the data?				X	
	Did the teacher and/or students talk about what one or more displays hide about the data?				X	

I developed scoring rules to mark out the outcome space of the observation data into qualitatively different categories that map onto the construct map. This outcome space is intended to index the construct map, so the scoring rationale and rules were designed to create a quantitative scale that can be interpreted in terms of the qualitatively different levels of the map. In this section I will explicitly describe the mapping between the program theory, as represented in the construct map, and the quantitative scores generated from the observation variables.

The segment variables were designed to provide information about particular levels on the construct map. For example, variables focused on unit specific concepts indicate higher levels of the construct if they are observed. Table 4 shows each segment variable for unit 1 along with the level of the construct map it is designed to inform. Remember, the student contribution and teacher practice variables are unit general, but the concepts are specific to this unit.

Table 5: Data profile from 45 minutes of instruction

	Segment								
	1	2	3	4	5	6	7	8	9
sConceptual	1	1	1	1	1	1	1	1	1
sProcedural	1	1	1	1	1	1	1	1	1
tInitSelect	1	1	1	1	1	0	1	0	0
tDiscussQ	1	1	1	1	1	1	1	1	1
tCompare	0	1	1	1	1	1	1	1	1
tPressExplain	1	0	1	1	1	1	1	1	1
tConnectOthers	0	1	1	1	1	1	1	1	1
iOrder	0	0	0	0	1	0	0	0	0
iGrouping	0	1	1	1	1	1	1	0	1
iScale	0	0	0	0	0	0	0	1	0
iShape	1	0	0	0	0	0	0	0	0
iShow	1	1	1	1	1	1	1	0	1
iHide	0	1	1	1	1	1	0	1	0

However, the descriptive statistics alone are not sufficient because it can lead to misleading conclusions. For example, although the concept variables provide information for

higher levels of the construct map they must be observed with other high level variables for classroom interactions to resemble the types of instruction we are supporting. For this reason I developed scoring rules to account for co-occurrences of particular variables in ways that discriminates between the qualitatively different levels described in the construct map.

To generate these scoring rules I first qualitatively analyzed the segment variables looking for profiles that indicated higher and lower levels of the construct map. For example, the profile in table 5 indicates that the class observed had teacher and students interactions described by higher levels of the construct map. In this matrix of binary codes the columns represent each 5-minute segment and rows represent each variable. A "1" indicates that the variable was observed at least once in the segment, and the columns are temporally ordered. So, the first column describes the first five-minute segment of instruction. Notice that the variable tInitSelect suggests that the teacher is selecting which student invented methods to keep at the center of the discussion, which also suggests that the conversation is focused on student inventions. The unit specific mathematical concepts (iOrder, iGrouping, iScale, iShape, iShow, iHide) allow us to see that the concept of grouping data dominated the conversation and that there was very little discussion of order, scale, or shape.

There was also discussion about what the displays show and hide about the data, although it appears this was primarily focused on what ordering shows and hides. Last, the teacher practice variables describe a classroom in which the teacher is asking discussion questions, pressing kids to explain their thinking, and comparing different ways of thinking across the class. While no individual variable can inform much about the quality of discussion, taken together these variables support the inference that this discussion about student invented method included

worthwhile conversation about the concept of grouping, but much less discussion of scale and order.

I used principles from these qualitative distinctions to design scoring algorithms to map data profiles to appropriate levels of the construct map. In Appendix E I have provided the full scoring rules for the items that constitute the quantitative scale. I designed the items to determine the support for student invention, the nature of the teacher and student contributions to the discussion, and the mathematical ideas that the class discussed. The items that look for support for student invention primarily look for the existence of particular tasks and materials that support invention. Since observers did not see classes in which the invention activities took place I can only make rough distinctions about teacher support. For example, I look to see if inventions are visible in the class, if there is diversity in student invention, and the context of the data that the students used when inventing. The items focused on student and teacher contributions to discussion index the ways students are talking and the ways teachers are using their strategies. Since we observe the actual discussion I can make more fine-grained distinctions. For example, in the item that judges the use of discussion questions classes would get a "1" if a discussion question was used and a "2" if the teacher also pressed kids to further elaborate their thinking after the original question. These are still rough distinctions, but are finer than the distinctions around invention. The unit specific concept items make the most fine-grained distinctions. In these items I index which concepts classes talked about, and how they talked about them. For example, for the items in unit 1 I gave a "1" if the concept was discussed (scale, grouping, order). I scored it a "2" if the concept was discussed while talking about what it shows about the data, and a "3" if they talked about what it shows and hides about the data.

I applied the scoring rules to create the items in two steps. First, I coded each class observation to produce polytomously scored items at the 5-minute segment level. So, I generated each of the items described in Appendix E for every 5-minute segment using the segment variables. Second, I used these segment level scores to create fidelity scores at the classroom level. I used the highest segment level score observed more than once during an observation as the overall score for the class. I looked for those that happen more than once for two reasons. First, a simple sum is not useful because it would create a scale in which "more" is not necessarily better. For example, there is nothing in our theory that suggests that discussing scale for 45 minutes is any better than discussing it for 30 minutes. In fact, it could be bad if a class was dominated by only one concept, but a sum of the scores would indicate "more fidelity." The

Table 6: Scoring rules applied to unit 1 profile

	Segment									
	1	2	3	4	5	6	7	8	9	
sConceptual	1	1	1	1	1	1	1	1	1	
sProcedural	1	1	1	1	1	1	1	1	1	
tInitSelect	1	1	1	1	1	0	1	0	0	
tDiscussQ	1	1	1	1	1	1	1	1	1	
tJuxtapose	0	1	1	1	1	1	1	1	1	
tPressExplain	1	0	1	1	1	1	1	1	1	
tConnectOthers	0	1	1	1	1	1	1	1	1	
iOrder	0	0	0	0	1	0	0	0	0	
iGrouping	0	1	1	1	1	1	1	0	1	
iScale	0	0	0	0	0	0	0	1	0	
iShape	1	0	0	0	0	0	0	0	0	
iShow	1	1	1	1	1	1	1	0	1	
iHide	0	1	1	1	1	1	0	1	0	Score
Item 1	0	0	0	0	0	0	0	4	0	0
Item 2	0	5	5	5	5	1	3	0	1	5
Item 3	0	0	0	0	5	0	0	0	0	0
Item 4	4	4	4	4	4	0	4	4	0	4
Item 5	2	1	2	2	2	0	2	2	0	2
Item 6	1	1	1	1	1	0	1	0	0	1
Item 7	0	3	3	3	3	3	3	3	3	3

second rationale for using a minimum of two segments is that we value sustained attention to particular concepts. I did not use the highest observed at least once because I didn't want to allow for the possibility of scoring an entire class session by an interaction that was only briefly observed. Table 6 illustrates these rules with the data profile from table 5. Notice that these scoring rules use the segment data to obtain scores at the same unit of analysis, the class observation level, as the summary variables and the artifacts.

These scores and the classroom level summary variables jointly constitute the quantitative fidelity scale. These classroom level scores, while not informative about within lesson variability, are useful to look for relationships between the overall quality of whole class discussions during the method reviews and student learning as measured by pre and post assessments. Since the measures of student learning are on the scale of an entire school year the classroom level indices are the most useful in characterizing the learning opportunities available to students during each unit in a way that can be related to the student measures.

Item Modeling and Validity Analyses

Partial Credit Model

I modeled the classroom level item scores with a Partial Credit Model (Masters, 1982) using Conquest software (Wu, Adams, & Wilson 1998) to examine the relationship between these item scores and the construct map. This model is a member of a family of generalized linear mixed models that estimates the probability of observing each scoring category as a function of the item's difficulty and the class being observed. It can be thought of as a version of the multinomial conditional logit model. Equation 1 is one mathematical formulation of this model describing the probability of classroom j obtaining a category score of y_{ji} on item i when the item has (m+1) scoring categories.

$$P(y_{ji}|\theta) = \frac{e^{[\sum_{k=0}^{y_{ji}}(\theta-\delta_{ik})}}{\sum_{h=0}^{m_i} e^{[\sum_{k=0}^{h}(\theta-\delta_{ik})}} \quad (1)$$

This model estimates the probability of observing a particular score on an item as the difference between a classroom (often conceptualized as a person in achievement testing) trait level and the difficulty of an item score. These parameters can be used to calculate Thurston Thresholds for each scoring category. These thresholds are the point at which a classroom has an equal probability of scoring at or above the level as they do below the level. In other words, there is a 50% chance of being scored at or above the category. So, these are cumulative probabilities. For example, if an item has four scoring categories (0, 1, 2, 3) the Thurstone Threshold for "2" would describe classroom trait score at which a classroom would have the same 50% chance of being being scored either "0" or "1" and a 50% chance of being scored either "1" or "2." These can be plotted on a Wright Map that places the item thresholds and classroom traits on a logit scale that can be used to relate the classroom scores and item level difficulties. A logit is the logarithm of the odds, termed "log-odds", of observing an item category given a particular classroom estimate. This is equivalent to the logarithm of the ratio of the probability of observing the category to the probability of not observing the category (or the ratio of the probability to its complement). I used the Wright map to look for evidence that categories indexing the same construct levels have similar thresholds across different items. I also looked for evidence about the relative distance between scoring categories for the items.

In addition to the Wright Map I used the following Conquest generated parameters to investigate each item:

1. Correlation between each item and the total score

- This provides evidence about the items moving in the same direction as the total score. Since all of the items are intended to index the same construct, each item should be positively correlated with total scores. Negative correlations would suggest that people scoring low on the item are getting higher total scores and would call for closer inspection.

2. Point biserial correlation between each category and total score
    - The point biserial coefficient is the same as the Pearson coefficient, but for the case when one variable is binary. Since each category is binary (you were in it or you were not) this describes the relationship between being in each item category and the total score. While the item itself might have a positive correlation, there might be particular scores that are not related with total scores in the ways we expect them to be. Ideally, the point biserial correlations should increase from negative values to positive values as the position on the construct map increases.
3. Item fit parameter (weighted mean square)
    - This describes the ratio of the mean of the squared observed residuals to the mean of the expected residuals. If these two are identical then the index is 1 (ideal). The conventional acceptable range is between .75 and 1.3. Items outside of this range indicate poor fit to the observed data, which would require closer examination to determine the cause.
4. Item reliability parameter
    - This parameter describes the extent to which items estimated high on the logit scale can be validly seen as qualitatively different from items lower

on the scale. This is a particularly important parameter since the scale is conceptualized as a quantification of our fidelity definition. So, it is important for high numbers on the scale to be qualitatively different from lower numbers since descriptions of higher fidelity levels are qualitatively different from lower levels.

The overall sample size for the data collection is significantly lower than what is typically used with this type of analysis. For this reason I pooled the responses from the items across units and across both years. This resulted in non-uniform sample sizes for the items. For example, item 5 in Appendix D (How did the teacher use discussion questions to facilitate a conversation about invented methods?) was generated for each unit. So, this item has scores for each of the units that the team observed. On the other hand, the unit specific items were only observed during their corresponding units. So, there are a smaller number of observations for each of these items than for the unit general items. We conducted a sufficient number of observations in units 1, 2, 3, and 5 to model the unit specific items, but the very small numbers of observations in units 6 and 7 made it impossible to model these items with the partial credit model. For this reason I did not include the observations for units 6 and 7 in the partial credit model.

Since I treated each observation as an independent event to increase the number of "independent" observations of the unit general items I violated an assumption of the partial credit model, that all units are independent of each other. In reality these observations have a very complicated cross-classified nesting structure. Although this violates an assumption of the model, it will provide more reliable estimates of the item parameters. In the next chapter I discuss the implications of this violation.

Validity Information from External Measures.

I compared the inferences I made from the observation measure to two external sources of evidence to further interrogate the validity of the data. The first source of evidence came from the professional judgments of the project's instructional coaches. This research project was intentionally designed to maintain a barrier between the coaching activities and the measurement activities. This is primarily due to the desire for teachers to trust the coaches as non-evaluative supporters of instruction. The coaches were not allowed access to the fidelity measures, and they did not use them in their practice. One productive consequence is that the judgments made by the coaches about each teacher's practice can be compared to the observation data since they were generated independent of each other.

During the first year of the project the coaches completed "coaching logs" after each time they traveled to the research site to meet with a teacher. In the logs the coaches described any communication with the teacher before the lesson they attended, the lesson they participated in, and the de-brief conversations after the lesson. The descriptions of the lessons typically focused heavily on the concepts the class discussed, although there was some attention to teaching practice as well. I investigated the extent to which descriptions of teacher practice in these coaching logs corresponded to the inferences made from the classroom observation data in a random subsample of the observations. I first identified the moments in the coaching logs where classroom instruction was described during days in which coaches and fidelity observers were in the class at the same time. Then, I examined the nature of these descriptions and the ways in which the descriptions were similar and different than the inferences made from the observation data. There were many descriptions that could not be compared because they fell outside of the scope of the measure. So, I first identified descriptions in the coaching logs that fell within the

scope of the measure and then compared them to the observation data. For example, one of our coaches, Lydia, described in her log about a unit 2 whole class discussion that we observed. She said the students in the class only invented methods that resembled the mode, and that the teacher "had her students work on the invented methods for most of the class period." She went on to share that the class only talked in a whole class setting for 10-15 minutes. Within this 10 to 15 minutes she stated that the students discussed two invented methods similar to mode, and that they realized "one of the methods could not be replicated and that the other was problematic because the data had two values with a frequency of 6 so it was impossible to determine which they should select." In this description there are four claims that can be tested against the observation data, 1) that the class only talked in a whole class discussion for the final 10-15 minutes of class, 2) the class only discussed methods similar to mode, and 3) the class discussed the replicability of the methods by 4) mapping between the methods and a visible data set. In this example I looked at the observation data and found that 1) the class engaged in a whole class discussion for the final two segments (10 minutes), 2) the iMode variable was the only statistic scored in these segments, 3) the iReplicability variable was scored in the final segment, and 4) the iLinkVisDist (link to visible distribution) was scored in the final two segments. This is an example of the observation data agreeing with the coaching log description on the testable claims made in the log.

The second source came from measures of student learning. We have evidence from previous research that the interactions described in the higher levels of the construct map support the development of student thinking around data, statistics, chance, and modeling better than the lower levels. I estimated the relationship between the classroom level scores and posttest measures using multilevel linear models with students nested within teachers. I created 7 models

that estimated the effect of each unit's fidelity scores on two different outcomes, a composite student posttest score and the subscale posttest score that corresponds to the unit. For example, unit 1 focuses on displaying data, so the subscale model uses the Data Display (DAD) subscale as the dependent variable. I estimated this relationship while controlling for student level pretest score, gender, ELL status, and ethnicity. The fidelity scores were the only classroom level variable. These models only included data from the first year of the study because the data from the second year student measures is still being scored.

CHAPTER IV

RESULTS FROM MEASUREMENT MODEL AND VALIDITY ANALYSES

Descriptive Statistics for Variables

Descriptive Statistics for Unit General Segment Variables

In table 7 I have provided the number of 5-minute segments in each unit across all observations and the percent of those segments in which an observer scored the variable. Between year one and year two there was significant turnover in the observation team. In spite of the turnover, there was great consistency in the percent of segments that each variable was observed, and also of the relationships between the variables. For example, the proportion of segments in which a discussion question was being asked were very similar between year one and year two. Also, the student talk variables (sConceptual and SProcedural) are very consistent between year one and year two. As I noted earlier, this was not the case for the variable indexing discussion of invented methods and for the variable indexing teachers making connections between student contributions and mathematical ideas. Because of this I have dropped them from this analysis.

It is already apparent from these proportions that teachers asked many discussion questions and students were talking, which is encouraging since we were observing whole class discussions. However, this doesn't provide much information about he quality of the questions or the student comments. These proportions have also pushed our theories about what we would expect to see in aggregate numbers from a group of classes engaging in the kinds of whole class discussion we are trying to support. It is true that proportions are not all that matters, or even the

Table 7: Percent of 5-minute segments observed for unit general items

Unit	Cohort	sConcept.	sProced.	tCompare	tDiscQ	tPressExp.	tConnectO.
				2012-2013			
1	1	61%	70%	50%	86%	53%	17%
2	1	46%	60%	16%	68%	54%	13%
3	1	53%	63%	17%	71%	50%	7%
5	1	58%	61%	10%	78%	32%	2%
				2013-2014			
1	1	61%	51%	45%	85%	57%	20%
	2	53%	53%	58%	84%	74%	13%
2	1	49%	62%	13%	75%	39%	3%
	2	69%	66%	37%	85%	71%	22%
3	1	53%	62%	17%	89%	35%	1%
	2	70%	69%	15%	83%	56%	9%
5	1	76%	25%	17%	91%	44%	1%
	2	56%	37%	8%	79%	40%	2%

most important things in determining quality. However, it is worth considering if there are any thresholds that are informative about quality. For some, such as the student talk and the discussion question variable, it is clear that more is better. However, it's not so clear for the compare variable. How much time should a class compare different ways of thinking? Surely there needs to be some time for making sense of the individual ideas, but how much time? In unit 1 the classes spent roughly half the conversation time comparing. But in unit two this dropped in year one, and it remained low for the first cohort of teachers in year two. While it is still unclear where the threshold lies for sufficient amount of time devoted to comparing ideas, I suggest that the percentages in units 2, 3, and 5 for cohort 1 are less than desirable.

The cohort design also allows for comparison between teachers in their first year and teachers in their second year. This comparison is somewhat problematic in this study because of the nature of the attrition from year 1 to year 2 in the first cohort. Three teachers from the same school dropped out after the first unit in year one, and they were three of the highest scoring

classes in the first year. Also, another teacher in the first year that had great success with the Data Modeling resources was promoted to a position for district wide coaching. While she still participated in the study as a coach, she did not teach classes in the second year. This attrition might explain some of the differences in cohort one and cohort two teachers during year two. Cohort two classrooms had a higher percentage of many important variables, even though they were in their first year of using the materials and the other cohort was in their second year. This is not what we expected. We expected to see growth from year one to year two, which we thought would show up as a positive difference between cohort one and two. While many of the teachers in year one did score higher in year two, the overall percentages remained lower than the second cohort.

Descriptive Statistics for Unit Specific Variables

The segment variable proportions tell very little about the extent to which the questions and student talk produced conversations about the mathematical ideas that the Data Modeling materials are designed to support. The unit specific variables show which items classes talked about more often, and how often they engaged with the epistemic nature of them. I have provided the proportion of segments each unit specific variable was scored in table 8.

In unit 1, classes talked about order for almost half of the segments, and grouping for a large proportion of the time. They talked about scale much less often, though. This is consistent with the fact that order and grouping are more accessible than scale, so these often dominate the discussions. This highlights one of the shortcomings of this measure, though. I know from anecdotes that many classes don't get to scale, but talk about it the next day. In year two, the first cohort talked about scale over twice as much as the second cohort. Since we only observed one day of the whole class discussion it is likely that we missed many of these conversations. Also

notice that classes talked much more about what displays show than about what they hide or influences on the shape of the data.

Table 8: Percent of 5-minute segments observed for unit specific items

Year	Cohort							
		Unit 1						
		iOrder	**iScale**	**iGrouping**	**iShape**	**iShow**	**iHide**	
1	1	41%	25%	40%	26%	69%	37%	
2	1	52%	27%	33%	14%	67%	30%	
2	2	54%	11%	39%	19%	66%	20%	
		Unit 2						
		iMode	**iMed.**	**iMean**	**iReplic.**	**iGeneral**	**iLinkVis Dist**	**iLink Imag**
1	1	35%	37%	22%	11%	21%	29%	14%
2	1	32%	38%	21%	16%	40%	27%	18%
2	2	53%	44%	40%	23%	32%	38%	11%
		Unit 3						
		iRange	**iCenter Clump**	**iDev.**	**iReplic.**	**iGeneral**	**iLinkVis Dist**	**iLink Imag**
1	1	21%	34%	21%	5%	17%	12%	5%
2	1	34%	44%	28%	5%	15%	38%	4%
2	2	41%	34%	17%	13%	18%	34%	9%
		Unit 5						
		iTheor. Prob.	**IEmp. Prob.**	**iSample Size**	**iSamp. Distrib.**	**iCenter Stats**	**iVariab. Stats**	
1	1	43%	52%	43%	45%	19%	13%	
2	1	45%	59%	32%	52%	23%	18%	
2	2	29%	51%	31%	40%	12%	7%	

In unit two, classes talked about mode and median for similar proportions of time, between 35% and 53% of the time. The first cohort only talked about the mean 21% and 22 percent of the time in year one and year two respectively. Cohort two classes, though, talked about mean much more. Their talk about mean was very similar in percent to their talk about the other two statistics. We would like to see more classes talking about the other four variables at unit 2 more often than they did. These are ideas that we hoped they would explore with each of the three statistics, but this seems unlikely given the percent of time for each. For example, cohort two classes only mapped between a statistic and an imagined distribution 11% in year

two, and cohort one only 18%, though this is an improvement from year one. Imagining new distributions and considering a new statistic in the imagined scenario is one of the most powerful ways to think about issues of generalizability. This is likely why cohort one classes doubled the percent of time they discussed the generalizability of particular statistics in year two.

In unit 3 classes talked about range and center clump methods for measuring variability. The center clump variable refers to any statistic that focuses on the center of a distribution. The conventional measure is the IQR. However, students often invent unconventional approaches that are grounded in the same ideas as the IQR. For example, sometimes students measure the width of the middle 1/3 of the distribution. Initially students often invent methods that don't make use of proportion, such as counting 10 data points on each side and then finding the distance between them. All of these are coded as center clump statistics in this measure. Similar to unit two, these two statistics provide the most widely accessible entry for students so they are often talked about before deviation methods. Deviation statistics refer to any measure that focuses on deviations from a center reference point. The most common conventional measures are the standard deviation and variance measures. However, we do not intend for students to think about these in middle school. Students often invent statistics that share conceptual similarities with these conventional measures, such as the sum of the distances between each data point and the median. The goal of the Data Modeling approach is to use these ways of thinking to build an understanding of measures such as median absolute deviation from the median. Just like the mean in unit two, this measure is the most challenging for classes to talk about, and this bears out in the lower percent of time this idea is discussed. Of course, the measure has the same limitations as in unit two because we don't know how many classes talked about this idea during the day following our observation.

Unit five is unique in that the growth effect that we expected for the cohort one teachers appears to be somewhat present. The proportions increased for the first cohort teachers in all but one variable. The growth is small in some cases, but there is a consistent increase. The classes talked about sampling distributions that they created, but rarely used variability or center statistics to make sense of the sampling distributions.

Overall the consistency in the percentages across the cohorts and the two years provide some evidence that the items were scored in a consistent manner. There were only two common members between the first year observation team and the second year team, so the consistency between the two years in the percent of segments scored for each variable suggests that these two observation groups were following the same criteria for the items.

Descriptive Statistics for Item Scores

Although the proportions above are informative about the relative amount of time spent on particular ideas they do not inform the ways the classes talked about the ideas. It is often times the co-occurrences of particular items that are most revealing. I designed the levels of the classroom level fidelity items to capture this. Table 9 shows the items and the percent of scores in each category. The dark cells are levels of the construct map that particular items do not inform. Remember that the unit of analysis for the proportions is different than in tables 7 and 8 so comparisons between the percentages are not very meaningful. Tables 7 and 8 give the percent of 5-minute segments that a variable was scored. Table 9 uses the whole class as the unit. These are marked if the level is scored in at least two 5-minute segments within a class. So, while table 7 tells us that teachers pressed students to further elaborate their thinking during around 50% of the segments table 9 tells us that teachers pressed kids to elaborate their thinking after asking discussion questions at some point during the lesson 91% of the time (level 2 of the

tDiscQ item). Table 9 also pools both years and cohorts since the sample size is too small to make the distinctions with the classroom level scores.

Table 9: Percent of classes scored at each level

	Level 1	Level 2	Level 3	Level 4	Level 5
sTalk	8%	91%			
tDiscQ	9%	91%			
RMData	95%				
OPPInv	97%				
InvDiv	92%				
tCompare				18%	51%
InvSelect		84%			
SharedUnd		85%			
iScale			7%	14%	34%
iGrouping			5%	34%	43%
iOrder			5%	39%	45%
iMode			9%	18%	49%
iMedian			16%	13%	53%
iMean			11%	7%	38%
iRange			16%	20%	38%
iCenterClump			27%	9%	42%
iDeviation			29%	4%	29%
VisDiffVar		80%			
iProbability			68%		
iSampDist			41%	12%	38%
iSampleSize			53%	18%	15%

The items that index the lower two levels of the construct map are populated by a large majority of the classrooms that were observed (all > 80%). On the highes level, though, classes only talked about the ideas in ways that addressed their epistemic nature between 29% and 53% of the time. There are significant differences between units, too. In unit 1, classes rarely talked about the display concepts without addressing their epistemic nature in some regard (level 3 of iScale, iGrouping, and iOrder). On the other hand, classes were more likely in units two and three to talk about the statistics without addressing their replicability, general use, or correspondence to real or imagined data. These are significant shortcomings since our intent was

to support students to develop an understanding of statistics as measures of distributional characteristics.

Partial Credit Model Results

Item and Classroom Thresholds

The item level estimates generated from the Partial Credit Model are represented as Thurstonian thresholds in a Wright Map in figure 13. If an item has k scoring categories, then there are k-1 thresholds. The threshold represents the transition from one category to the next. The threshold estimates represented in the Wright Map show classroom estimate needed to have a 50% chance of scoring at or above the particular category. The Wright map places the item level estimates and the classroom estimates on the same logit scale. This is one way to describe the probabilistic relations between classrooms and item levels since the logit is the logarithm of the odds (p/1-p). On the right side of the scale the item levels are displayed. I have separated the item levels into columns that correspond to the level on the construct map the level is intended to index. For example, the item that indexes the use of the student invention task is in the first column since it differentiates between classes at the lowest level and classes that are not on the construct. It doesn't provide any information about the higher levels. On the other hand, the highest level of the item that indexes how classes talked about scale is placed in the column at the far right. This item differentiates between the highest level and the levels below. Readers can interpret the vertical location of item level as the likelihood of that level being observed. This is sometimes thought of as the "difficulty" of the item level. So, the purpose of separating the item levels into their corresponding construct levels is to see if items mapping to higher levels on the construct are more difficult than those at lower levels.

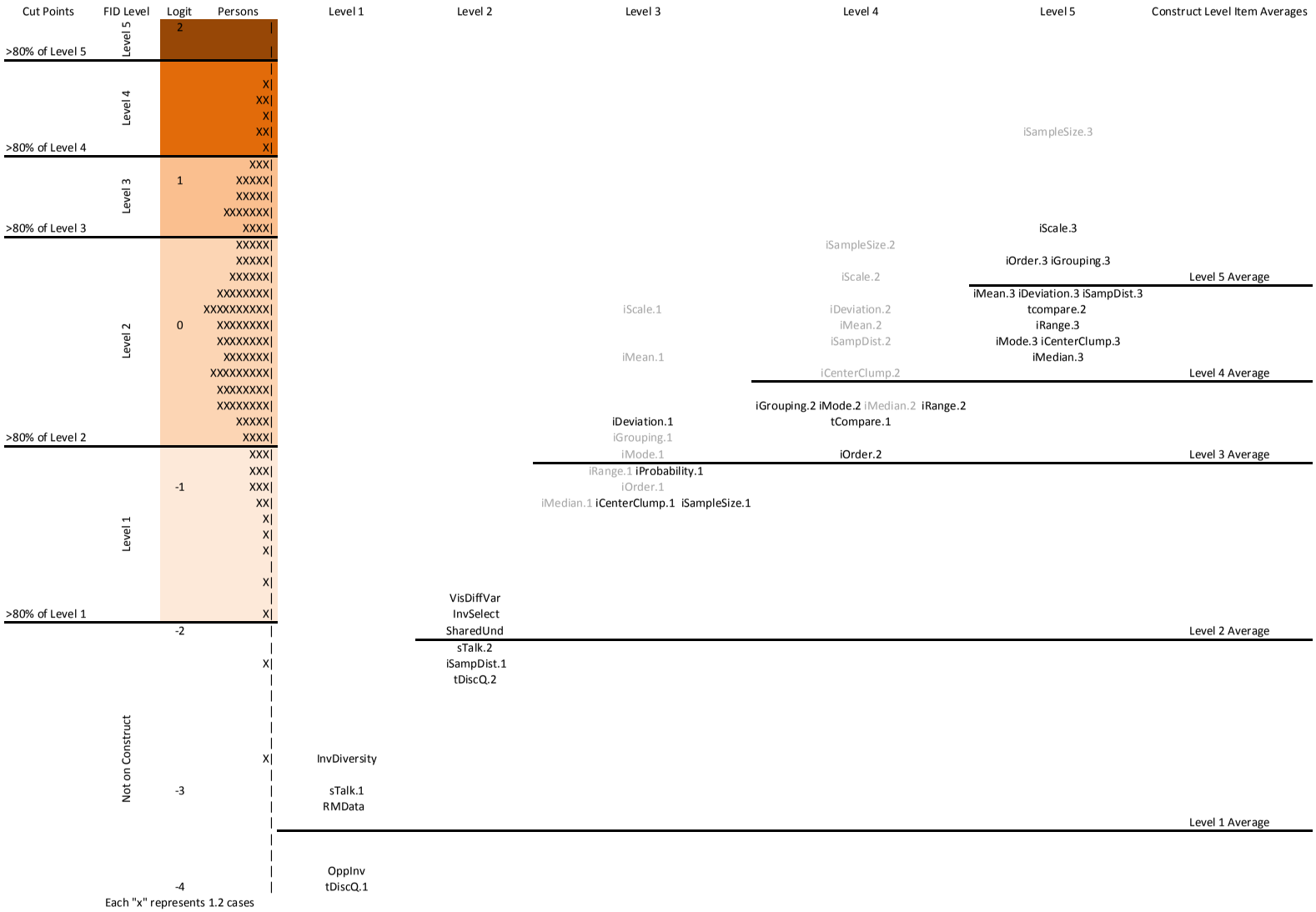


Figure 13: Wright map from partial credit model

Since the probability of observing a particular level is a function of the difference between the classroom estimate and the item level estimate the two can be related on the scale. For example, if an item level and classroom have identical logit estimates (same horizontal location on the Wright Map) then the model estimates that an observer would have a 50% chance of seeing evidence of the level in that particular class. The higher a classroom estimate is from an item level estimate, the higher the probability of observing the level in the class. Inversely, the lower a classroom estimate is from an item level estimate, the lower the probability of observing the level.

On the item side of the scale, the item scores that correspond to the first level of the construct map were the easiest to observe. The five thresholds at level 1 ranged from -4.03 to -2.77 with an average of -3.37. In almost every class the students, at a minimum, had opportunities to invent methods before the observers arrived (OppInv) and they used repeated measures data (RMData). In addition, the widespread diversity in student-invented methods (InvDiversity) suggests that teachers almost universally structured the invention task in a way that supported students to genuinely engage with the ideas. Teachers also used discussion questions (tDiscQ.1) in almost every class we observed and students talked at least procedurally about the student inventions (sTalk.1).

The six thresholds for items scored at level 2 ranged from -2.36 to -1.87 with an average threshold of -2.07. These items were more difficult to observe than level 1 items but much easier than the higher levels. Remember that this level represents a move from only using tasks and materials to using them along with strategies, such as pressing students to elaborate their thinking, to engage students in talking about them. The levels of the discussion question items illustrate the rationale behind this difference. The discussion question item is scored at level 1

when teachers only ask students questions that have the potential to support discussion. As I said before, this was an item that was scored generously since it is often difficult to determine a question's quality in a 5-minute segment. You can think of Ms. West's question in chapter three as an example of the minimum an observer would need to see in order to score it at level 1. Although the question in Ms. West's class didn't really support a discussion, it had the potential to. It was open ended, and the students provided opportunities to develop important ideas around order. However, the teacher did not notice the fruitful moments to press kids to elaborate their thinking, which is what an observer would need to see to score it at level 2. This distinction is why I consider the first level of scoring to be consistent with the first level of the construct map. Using discussion questions that are suggested in the curriculum is a similar task as using activities provided by the curriculum. On the other hand, the second level of scoring indicates a classroom where student responses to the questions are being further elaborated. The distance between these scoring levels, and the difference between the two levels in general, suggest that this is not a trivial difference in classroom practice.

At level 3 the eleven thresholds ranged from -1.15 to .17 with an average threshold of -.75. Notice that at level three and above there are some item thresholds that are grey. These are levels at which there was a low frequency of scores (< 8 scores). Sometimes this indicates a score that should be collapsed with another score since it is rarely observed. However, given the small sample size of the calibration group, and the meaningful differences in codes grounded in the construct map, I decided to keep these codes in the measurement scale. The overall sample size for this study is smaller than what is typical for this kind of modeling work, and this fact is even more true for these items than the others. This is because the items in levels 3, 4, and 5 refer to the mathematical ideas, which are unit specific. Although I decided to keep these codes in the

measurement system their threshold estimates for should be considered with the low sample size in mind. The implication is that these estimates are much noisier than the others, that is to say that the sample-to-sample variability is so large that the location in this calibration is less trustworthy than the other items. However, for this study I used these estimates to calculate the average of the level three thresholds because of the conceptual meaning. The items scored at this level refer to classes where the mathematical concepts were at least talked about, even if the conversation did not address the epistemic nature of the ideas. For example, in unit 1 the items are scored at this level if the class at least talks about scale (iScale.1), order (iOrder.1), and grouping (iGrouping.1). Again, Ms. West's class is an example of a class that just meets the minimum requirements to be scored at this level. The students never got to discuss the ideas, but observers scored this even if just the teacher talked about them. These estimates suggest that even with this generous coding this level is very different than levels one or two. There is a big difference between getting students to talk and talking about worthwhile mathematical ideas. This is not a new idea, but these threshold values provide an empirical estimate of how much more different.

The 12 thresholds at level 4 ranged from -.85 to .56 with an average of -.2. The unit specific items at this level indicate classrooms that were beginning to discuss worthwhile mathematical ideas in ways that addressed their epistemic nature. This is operationalized somewhat differently across units. For example, in unit 1 the grouping item would be scored at this level (iGrouping.2) if the class discussed what grouping decisions show about the data, but didn't go on to discuss what it hides about the data. For units two and three the items are scored at this level if the class discussed the replicability or generalizability of a statistic (center in unit 2 and variability in unit 3) without explicitly establishing correspondences between the statistics

and different distributional states. This happened when teachers would ask "do you think anyone could follow the rules for this statistic and get to the same number you did?" or "do you think this statistic would always work well to measure variability?" Notice that the difference between level 3 and level 4 is much smaller than the previous differences. In fact, these levels have some overlap between them. This suggests that while the levels are different from each other, they are much "closer" to each other than to the other levels. One way to conceptualize this difference is to think of it as the difficulty of moving a class from one level to another. So it is hard to move from level 2 to level 3, but once you are at level three it is much easier to move to level 4.

There is one unit general item scored at level 4, the item that indexes when teachers compare different invented methods (tCompare.1). The fact that this item has a similar estimate to the level four unit specific items supports one of the Data Modeling approach's most important instructional theories, that the epistemic nature of the mathematical ideas is most transparent when different invented methods are compared and contrasted. For example, this is the strategy Mr. North used to support a discussion of grouping. The "pie graph" and the "graph-graph" both grouped by decades, but used different representational strategies that differentially showed or hid aspects of the data. Later in the class the students compared a display that did not group the data to these and talked about the implications for these decisions for readers.

The twelve thresholds for level 5 ranged from -.20 to 1.31 with an average of .29. These scores represent the highest level of the construct map because they refer to discussions that treated the concepts as epistemic tools. The difference between level 4 and level 5 is more a matter of degree than difference. For example, in unit 1 the level 4 scores represent conversations about what display concepts show about the data, but at level 5 students talk about what the concepts show and hide about the data. For units 2 and 3 the students not only discuss

the replicability or generalizability of a statistic, but also build correspondences between the statistics and different distributional states. In addition, the tCompare item is scored at level 5 if the teacher compares different invented methods, but also makes connections between different student ideas in the class. This was a difference illustrated by Ms. West and Mr. North. Ms. West compared invented displays, but Mr. North compared while making reference to the differences in students' ideas in the conversation and in the displays. The differences between these scores on the wright map suggest that it takes significant work to move from just comparing, to more meaningful comparisons where students ideas are discussed. Also, for both levels of the items the threshold estimates are consistent with the thresholds for the items focused on indexing the ways mathematical ideas were discussed in classes.

The top three levels also provide interesting information about the variability within each level. This suggests that the concepts are not equally likely to be discussed. For example, look at the difference between the iScale item and the iGrouping item. Both of these are unit 1 items, but it appears that grouping was more likely to be discussed than scale. This difference highlights the relationship between instructional practice and context in which it plays out. During the summer professional development meetings the leaders used grouping as a focus since students often invent with interesting differences in how they group data, and these differences have clear and accessible implications for the shape of the data. With this experience, and with the fact that student inventions reliably produce opportunities to make this idea visible, it makes sense that it would be discussed more often, and that the relationship would hold true for all three levels.

The left side of the scale provides the distribution of classroom logit estimates. The item estimates are concentrated on the lower end of the classroom distribution, but it is important to remember that if classrooms and item levels have the same logit values there is a 50% chance of

observing at or above the level in the classroom. From a fidelity perspective this is not terribly encouraging. I expect to see the higher construct levels represented in classrooms with high fidelity more often than 50% of the time. I have provided cut points on the classroom side of the map for the points at which the classrooms would have 80% chance or higher of observing each level to show how the distribution is partitioned with a much more conservative expectation. Under the 80% requirement virtually every class is above the lowest cut point. However, the level 4 has very few classrooms in its region, and there are no classes that have an 80% chance or above of exhibiting level 5 scores.

I am not arguing for 80% as a standard, but rather suggesting that the measurement model provides opportunities to think about a classroom's fidelity in new ways. Fidelity is typically thought of as a static location for a class. However, there is great variability from class to class and from unit to unit. I argue that it is more productive to think about fidelity in probabilistic terms. Instead of asking about what levels particular classrooms are at, we can ask about the likelihood of observing particular levels in a given class.

The Wright map allows for this kind of question. For example, consider again Ms. West and Mr. North. During my description of their classes I focused on the levels that the excerpts most clearly illustrated. However, this characterization alone could lead one to expect that Ms. West's would not be scored on levels 3 or 4, and that Mr. North's class would only be scored on the highest level possible for each item. This is what is called Guttman scaling (Guttman, 1950). However, the Guttman requirement, that any subject that scores on a particular level must always me able to score at any level below, it rarely observed in practice (Kofsky, 1966; Wilson, 2004). Mr. North and Ms. West are a perfect example. Although Ms. West scored on levels three and below for most of the items, she scored on level 4 for the tCompare item. Mr. North, while

scoring on levels 5 for iScale and iGrouping, only scored on level 3 for iOrder. If fidelity is thought of as an exact location then this situation is problematic. However, if fidelity is thought of as a probabilistic relationship between the classrooms and the construct levels then these situations are not only acceptable, but are expected.

In table 10 I have provided the probabilities of observing items at each level in Ms. West's and Mr. North's classrooms. I used the average item threshold at each level to calculate these probabilities. In comparing we can consider the differences in probabilities for the 5 levels. There is very little difference in the likelihoods of observing items for level one. At level two, though, the two probabilities begin to become farther apart. By level 5 Mr. North's class has a 70% chance of being scored at this level, while Ms. West only has a 39% chance.

Table 10: Probability of observing item levels

	Level 1	Level 2	Level 3	Level 4	Level 5
Ms. West	97%	86%	67%	52%	39%
Mr. North	99%	96%	87%	80%	70%

This is also helpful for characterizing the extent to which the treatment condition classrooms used the curriculum in ways consistent with the intentions of the design. For example, every classroom observed but two have a greater than 80% chance of being scored at or above level 1. In fact, 90% of the cases observed had a 60% chance of being scored at level 2. 72% of the classes had a 60% chance of being observed at level three. This suggests that as a group, virtually all the whole class conversation made use of the materials and tasks and used some strategies to engage kids with talking about their invented methods. However, in only 21% of the observed classes did we have at least a 60% chance of seeing the target concepts talked about at or above level 3. This paints a picture of a group of teachers that universally made use of

the materials and engage kids in conversations, but that still have much growing to do in order to make the conversations as fruitful as they could be from the perspective of the program theory.

It is important to remember that the classroom estimates in this model do not account for the nesting within each teacher or within unit. This means the percentages in the previous paragraph could be misleading. A closer look reveals that many of the highest scores on the scale were from unit 1 observations. Only one of the scores in the highest 20% of scores was from a unit 3 observation. This suggests that the mathematical domain greatly influences the difficulty of facilitating a high quality whole class conversation. This also provided evidence of the validity of the person scores since we have known for a while that teachers have a much more difficult time with measures of variability than they do with data displays. This Wright map provides an empirical estimate of how much more difficult it is. To shift the probabilities from Ms. West to Mr. North requires moving over 1 logit on the scale. Considering that the range of the entire distribution is around 4 logits this is a very significant distance.

Item Statistics from Partial Credit Model

I have provided the item correlations with total score and the item fit statistics in table 11. All items in the measure are positively correlated with the total score (sum of all item scores), indicating that all items move in the same direction as the total score. All of the correlations are statistically significant, but this is not particularly informative since this significance only tells us if it is truly different than zero. Notice that the items on the lower levels of the construct map are not as highly correlated with the total score as the items that index higher levels. This is primarily because there is much less variability in the lower level scores since observers saw them in most of the classes. For the higher levels, though, the scores are more highly correlated with the total score (between .59 and .80). In addition to the overall item correlation I examined

Table 11: Item correlations and fit statistics

	Correlation	Weighted MNSQ
sTalk	0.25	1.13
tDiscQ	0.41	0.98
RMData	0.41	0.97
OPPInv	0.51	0.95
InvDiv	0.48	0.93
tCompare	0.41	1.23
InvSelect	0.32	1.01
SharedUnd	0.55	0.90
iScale	0.73	1.05
iGrouping	0.78	0.89
iOrder	0.73	0.97
iMode	0.64	1.14
iMedian	0.70	1.04
iMean	0.59	1.19
iRange	0.61	1.19
iCenterClump	0.63	1.11
iDeviation	0.57	1.06
VisDiffVar	0.47	0.94
iProbability	0.47	1.02
iSampDist	0.80	0.96
iSampleSize	0.78	0.88

the point biserial correlation between each item level and the total score. The correlations generally increased within an item from lower scores to higher scores. There were five instances in which this was not true, but all five cases occurred with categories that have very low sample sizes.

All of the item fit statistics are within the conventionally acceptable range of .75 to 1.33, suggesting that the residuals were very similar to the expected residuals. The reliability estimates suggest reliability in the item estimates, with the item separation reliability being .96. This indicates that the differences in item parameters can be reliably thought of as differences in the properties of the items and item levels. The person separation reliability, which in this case refers to the reliability of the classroom estimates, was .69. This is much lower than the item reliability. This is due to the fact that the standard errors of the person estimates were somewhat large, and

is somewhat troubling given that this is a measurement system designed to provide information about individual classrooms.

Classroom Fidelity Estimates

Since I pooled the data across units and both years this model produced a classroom estimate for each of the observations. This means that each classroom receives an estimated fidelity score for every observation. For example, if a classroom was observed for units 1, 3, and 5 in both years they would receive one fidelity score for each of these observations, 6 in total. One advantage to this is that I can examine the relationships across the units for individual teachers and in aggregate. One challenge, though, is deciding how the multiple estimates might be combined into a summary fidelity score for each teacher.

Table 12: Average classroom logit estimates

	Year	Unit 1	Unit 2	Unit 3	Unit 5
	1	0.486	-0.272	-0.763	-0.244
	2	0.167	0.076	0.041	0.102
Mean		0.326	-0.097	-0.360	-0.070
Difference		-0.318	0.349	0.804	0.346

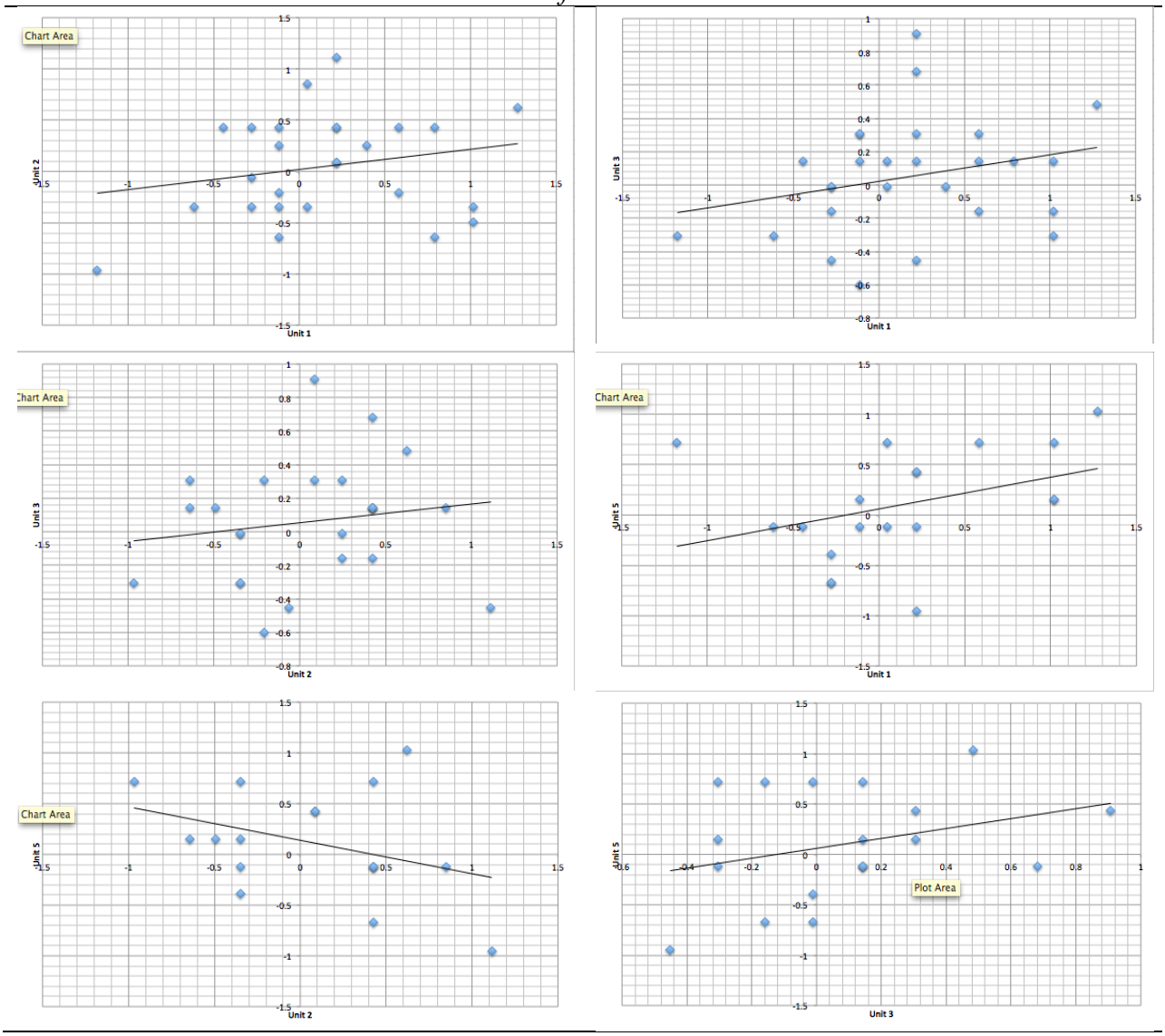
In table 12 I have provided the average fidelity scores for each unit in each year. Both years show similar trends with unit 1 having the highest average score, units two and five having similar scores, and unit three having the lowest. In many ways this matches our expectations. Unit 1 is on data display, and the concepts are typically familiar to teachers. On the other hand, unit 3 focuses on measures of variability. It is often the case that teachers in our projects first learn about measures of variability, other than range, in our professional development. In fact, during the first year of the project the PI observed one class during unit three and was surprised to hear the teacher explain to the class that the median was a good measure of variability. This paired with the observation data from more classrooms led to a redesign of the professional

development to better support teachers with this new idea. With these kinds of experiences it is not surprising that the lowest average estimate was in the first year's unit 3 observations.

It is encouraging, though, that the unit 3 average increased by .8 logits between year 1 and year 2. In fact, all of the averages increased in year 2 other than the unit 1 observations. Closer examination of the item scores revealed that the differences were due to lower scores on the iScale and iGrouping items in unit 1. Both were talked about less frequently in the classrooms, and the classes talked about what these concepts hide about the data less often. iOrder was talked about more often, but not enough to offset the other differences. Students talked about what the concepts show about the data the about the same percent of time in the second year as they did in the first year.

The fidelity estimates between pairs of units were positively correlated in all but one instance. Table 13 summarizes this with scatterplots of every possible pair of units. As you can see, 5 of the 6 show a clear positive relationship between estimates. This supports the inference that teachers with higher scores in the earlier units are more likely to have higher scores in the later one, while also allowing for variability in the degree to which this is true.

Table 13: Pairwise correlations between units for classroom estimates



Validity Information from Relationships with External Data Sources

Overall, the coaches' judgments about individual teachers in the first year of the project and the inferences supported by the measurement system were very consistent. Remember, the coaches did not have access to the observation system. In fact, although they knew observations were taking place, they did not know what the observers were recording. There are two pieces of information from the coaches that I used to add to the validity evidence for the measurement system. The first piece was the result of our coaching and PD team living in a different region of

the country from the study. During the first year of the project the team decided to identify a select group of teachers that were especially successful with the Data Modeling approach to serve as "on the ground" coaching support for the second year teachers. They selected six teachers based on their observations and coaching time. I compared the fidelity scores of these six teachers to the entire sample and found that their scores were largely in the higher end of the distribution. The average classroom estimate for the population is approximately zero, and the average theta score for the observations from the teachers selected to be coaches was .4, almost half a logit higher. In fact, there were 7 observations that had greater than 80% chance of being scored on level 4 items, and all seven of these were from the teacher/coach classrooms.

The coaching logs provide information about mainly the unit specific items in units 1 and 3. This is because these units we observed during the same day the coaches participated in class for these two units in the first year. Out of these common dates there were 27 "testable claims" made in the logs across 6 different coaches and thirteen different teachers. The claims were almost exclusively about the ideas the class talked about, and sometimes about how they were talked about. For example, if a log claimed that a unit 1 class talked about scale, order, and grouping I went to the observation data to see if it supported the same inference. Of the 27 claims, 6 came from Unit 1 and all were supported by the observation data. This left 21 claims in unit three, and all but three were supported by the observation data.

There were many claims in the logs that could not be tested against the observation measure, and a number of these highlighted some of the shortcomings of the system. There were meaningful differences in instruction that this measure did not capture. For example, in one log the coach said the teacher used a lot of the questions, and the students talked a lot, but that the teacher asked the questions in a ways that hinted at what he wanted students to say. The coach

said this produced a conversation where students were saying a lot of things about the ideas, but were not sure what it was they were saying or why it mattered. She said that the class seemed more like the kids trying to figure out what the teacher wanted to hear rather than genuinely engaging in inquiry around the data. However, in a class like this the measurement system was not able to pick up on these issues. The variables suggest that this class was talking about the ideas in meaningful ways, and it is likely that nuances such as this are lost in a number of observations.

The multilevel models relating the fidelity scale and the student assessment scale suggested a number of positive relationships between classroom fidelity and student thinking. Table 14 shows the coefficients for the 8 models. Four of the models pair the unit specific fidelity estimates with the composite posttest student scores. The other four models pair the unit specific fidelity scores with the post test subscale scores from corresponding subscales (Unit 1 with Data Display (DaD), Units 2 and 3 with Conceptions of Statistics (CoS), and Unit 5 with Chance (Cha)).

The unconditional model estimated that 10% of the overall variability was from between classroom differences. Unsurprisingly, the pretest scores were the strongest predictor of the posttest composite scores and the posttest subscale scores. The gender, ELL, and ethnicity variables are not included in the table. The gender variable was negatively correlated with all outcomes suggesting that boys, on average, scored lower than girls in the Data Modeling classes. Neither the ELL nor the Ethnicity variables were statistically significant ($p < .10$).

The Unit 1 fidelity variable had a statistically significant positive relationship with the composite posttest ($p = .096$) and the DaD posttest subscale ($p = .068$). The coefficient for the unit 2 fidelity variable was positive, but not statistically significant ($p = .218$). However, the

Table 14: Regression coefficients from multilevel models

DV	Pre Composite	Pre Subscale	U1 Fidelity	U2 Fidelity	U3 Fidelity	U5 Fidelity	% Classroom Variance
Posttest							10%
Posttest	.252**		.141*				4.56%
Post DaD		.156**	.160*				4.26%
Posttest	.253**			0.078			4.61%
Post CoS		.099**		.174*			3.58%
Posttest	.244**				0		8.22%
Post Cos		.110**			0.009		8.33%
Posttest	.240**					0.18	5.27%
Post Cha		.142**				.292**	2.20%

* Significant at .10 level

** Significant at .05 level

coefficient was positive and significant when predicting the CoS subscale (p = .066). The same is true for the unit 5 fidelity variable (p = .123 for composite post test model and p = .039 in the Cha subscale model). It is important to note that these models are underpowered to detect significant relationships at the classroom level due to the small numbers of classrooms. The study is designed to have sufficient power for the main effect between treatment and control groups (~ 80 classes total). However, this study only uses the treatment classes in which we conducted observations, which drastically reduces the sample. This means that although the positive coefficients are not all statistically significant the consistent positive direction is worth noting for units 1, 2, and 5. In addition, the percent of unexplained between classroom variance was significantly lower in all models with units 1, 2, or 3 predictors.

Unit 3 is a different story. The unit 3 fidelity variable was not correlated with either outcome. This is a striking difference when compared to the other units. This is likely due to the fact that the fidelity scores in unit three during the first year were some of the lowest in the study. The average score was -.721 in the first year, and rose to .042 in the second year. Since these models only use the first year data I am not able to test this conjecture. However, as soon as the second year student data is scored this will be an important finding to interrogate further.

Rater Reliability

In table 15 I have provided the percent of segments in which observers agreed for each variable. I calculated the agreement at the segment variable level because this provides the most actionable information about the extent to which the observation system can be reliably used by multiple raters. To calculate these percentages I used the 18 double observations from units 1, 2, 3, and 5 in the second year of the study. This data represents a random sample of 10% of the observations conducted in these units in year two. However, they are not equally distributed across the units. This is particularly problematic for unit 1 since we only conducted two double observations in this unit. This sample consists of two double observations from unit 1, four from unit 2, six from unit 3, and six from unit 5. Remember that the student and teacher variables (those beginning with "s" or "t") are common to all the units, so these percentages represent the agreement across the full sample of double observations. The variables beginning with "i" are unit specific, so the sample is much smaller for each of these.

Overall the observers agreed in their scoring a very large proportion of the time. Observers agreed greater than 80% of the time for all but four of the variables. This is not surprising for some of the variables, such as tDiscussQ, since the observers were trained to code them generously. For other variables, though, this was quite an accomplishment considering the nuanced differences the observers were trained to record. For example, observers agreed almost 90% of the time when scoring iLinkImagDist. This variable records moments where students imagined new data distributions and made correspondences to the statistics they were discussing. This is an important aspect of the conversations, but one that takes significant training for observers to see.

Table 15: Percent agreement for segment variables

	% Agreement
sInvented	94.65%
sConceptual	80.35%
sProcedural	76.97%
tInitSelect	93.86%
tCompare	90.97%
tDiscussQ	92.09%
tPressExplain	79.08%
tConnectOthers	93.71%
iOrder	100.00%
iScale	75.00%
iGrouping	91.67%
iShape	100.00%
iShow	83.33%
iHide	91.67%
iMode	92.12%
iMedian	93.88%
iMean	95.94%
iReplicability	86.38%
iGeneralizability	88.71%
iLinkVisDist	84.21%
iLinkImagDist	89.46%
iRange	88.22%
iCenterClump	78.76%
iDeviation	83.95%
iTheoreticalProb	86.66%
iEmpiricalProb	85.09%
iOdds	93.81%
iSampleSize	93.89%
iSamplingDistrib	93.11%
iCenterStats	98.61%
iVariabilityStats	96.10%

Chapter V

DISCUSSION

Representing Program Theory

In this study I have developed, implemented, and studied a measurement system for characterizing the extent to which whole class discussions in Data Modeling classrooms are faithful to the kinds of discussions the materials are designed to support. The notion of fidelity, although widely viewed as critical to studies of large-scale use of designed learning programs, has had very little consistency in meaning or in measurement. There has been recent work to bring more consistency to this field of work, but these conversations have almost exclusively focused on issues of data collection and data analysis and have given much less attention to foundational questions regarding underlying theory and measurement. I have attempted to bring these issues to light.

I am far from the first to suggest that fidelity measures should be rooted in underlying program theories. This is a widely held conviction. However, there has been little agreement on what a program theory is, and what would constitute a sufficient representation of one. Without an explicit discussion of this issue researchers have implicitly and differentially focused on three different aspects of a program theory: program structure, program processes, and underlying latent constructs. In most cases the emphasis has been on program structure and processes, with much less attention to underlying constructs. Here I have argued that a sufficient program theory should represent all three, as well as the relations among them, and that the construct map is a particularly useful tool to represent underlying constructs in the context of fidelity.

The representation of the Data Modeling program theory that I sketched in chapter three does just this. The construct map provides a conceptual description of the target phenomenon with two features. First, it described the phenomenon along a trajectory. This is particularly useful in a fidelity context because it provides a theoretical rationale for what counts as "more faithful." The construct map, though, is also related to the particular design strategies in the Data Modeling approach to realize the construct in classroom interactions. The tasks, materials, and instructional strategies are not simply things to do, but are seen as tools to bring about the instruction described in the construct map, and hopefully the kinds described in the higher levels. One advantage to this kind of representation is that it makes the theoretical intent of the design explicit to the reader. This provides the opportunity to make more meaningful comparisons between alternative designed programs. It is possible to imagine a program that is rooted in a very similar underlying construct, but makes use of very different design strategies to realize it. In this case the comparison could focus on the operational designs and the relative success at realizing the underlying construct. On the other hand, one can imagine a very different kind of comparison in which a designed program is fundamentally motivated by a different underlying theory. In this case the question "what works better?" makes little sense because the two designs are built to do different work. The question under this circumstance would be "which underlying theory do we value more?" Without explicit attention to these three aspects of a program theory attention to these details is not possible.

Representations of program theory also have implications for how one might conceptualize program implementation. If structure is the only aspect of program theory represented then researchers might conceptualize implementation as a straightforward task of using the program components. On the other hand, if processes are represented then

implementation might be thought of as component use with a particular level of quality. In contrast, this dissertation represents program theory in a way that unifies structure, process, and underlying latent construct. Here, the *what* of the program, the tasks and materials, and the *how* of the design, the teacher instructional strategies, are merely components of an overall design intended to bring about the *why* of the program, opportunities for students to represent, measure, and model variability. Under this framework implementation takes on a different meaning than the other two I just described. In our teacher professional development meetings we focus on bringing about opportunities for students to engage with the ideas, and we describe the Data Modeling materials as tools to do this work. It's not that we present them as optional, but that the intent of the tools is kept at the front. Implementation, then, is the *use* of the tools to realize the instructional theories.

It's worth noting, though, that representing program theory is not solely a representational activity. It is also a meaning making activity. During quality design studies many of the relations between underlying theory and the designed learning environments realized in action are worked out in fine detail, but fidelity measurement pushes on theory in new ways because of new constraints. Measurement is expensive, so it requires thought about which aspects of the theory are most relevant and it requires designers to consider which design elements are essential. At the same time, it requires thought about which of these elements are feasible to measure. These kinds of questions can bring a new understanding of a design. While underrepresented in this study, this was the case here. Many iterations of piloting produced multiple changes in how I conceptualized and represented both the construct and the mapping between the underlying theory and the visible instantiations in the Data Modeling design.

But this is not a dissertation only about representing program theory. It is a dissertation that claims that a measurement scale is a productive way to think about and empirically study a program theory. There are at least two questions that must be answered to support this claim.

Research Question One:
Can we validly and reliably measure the extent to which students are supported to represent, measure, and model variability in Data Modeling classrooms?

The temptation with questions about measurement validity, which researchers sometimes succumb to, is to provide a once-and-for-all statement. This temptation makes sense. How nice it would be to build meaningful measurement systems that can be validated in a final and certain way so future researchers can pick them off the shelf and use them without spending energy and resources on questions of validity. Hopefully it is clear at this point that I'm going to resist this temptation. In fact, my wording of the research question was my first strategy for this resistance. I didn't ask if this *was* a valid measurement system. I asked if this construct *can* be validly measured.

Instead of a final *yes* or *no*, this question suggests that the answer is but a first step in a much longer process. It provokes a "*yes, but*" kind of answer. In this section I will argue that the data I have constructed with this system validly informs many relevant questions about the kinds of whole class discussions supported by the Data Modeling program. So, part of the answer to this question point backward to what I did, and what I think it means. However, I will also argue that a part of this system's validity is found in its potential use in the future. Another part of the answer to this question is found in the things that might be changed about the system to make it useful for even more powerful inferences in the future. I will address both of these within the categories suggested by the Standards for Educational and Psychological Testing (American

Educational Research Association, American Psychological Association, National Council for Measurement in Education, 1999; Wilson, 2004).

Evidence Based On Measurement System Content

The Data Modeling program theory is rooted in a commitment to providing students accessible experiences that provoke the need for new mathematical tools. However, these experiences are not sufficient to grow foster understandings of worthwhile mathematical ideas. The Data Modeling program is designed to leverage students' sense making in these experiences to support the development of more powerful ways of thinking. The ideas are leveraged in whole class discussions as students share their thinking, compare different ways of thinking, and engage with mathematical ideas as epistemic tools. Although there are a number of student ideas that teachers can reliably expect to see during instruction, the realized form of the instruction is hopefully shaped by the particular ways students talk about and engage with the ideas. So, faithfulness to the intent of the tools can look different from classroom to classroom. In fact, if a large number of classrooms looked identical during instruction it might suggest that the instruction is overly routinized and the materials are not being used as intended.

I believe that the observable variables, observation protocols, and scoring rules described in this dissertation formed a measurement system that was able to identify many meaningful differences in classroom instruction. The summary variables focused on lower levels of the construct map by indexing the presence of critical structural pieces of the design such as invention tasks and data context. The segment variables provided a more detailed picture of the whole class discussions. Although the student talk variables made very rough distinctions, they did represent an important difference in the ways students often talk about invented methods, the difference between describing *what* someone did and describing *why* they made that choice or

the conceptual implications of the method. The teacher practice variables provided information about teaching strategies that we have found to be powerful in leveraging student thinking, such as asking questions and comparing different student ideas. The teacher variables also spanned a wide range of the construct map. I interpreted the mere presence of a question as evidence that teachers at least made use of the suggestions in the curriculum materials (level 1). I saw questions with teacher presses for more elaborated answers as an indication of a higher value for student discourse (level 2). Teacher strategies that included the comparison of different strategies and discussion of the different ideas in the class represented instruction at the highest levels (Level 4 and Level 5).

In addition to these domain general variables, this measurement system also coordinated the general with the specific mathematical ideas the units were designed to support. This provided information about the higher levels of the construct map since levels 3, 4, and 5 describe instruction in which the class discussed the relevant mathematical ideas. These variables are also useful to identify different ways classes discussed the ideas. For example, in unit two the variables look for classes to discuss the median. In addition, the variables account for comments that suggest students engaged with the median as a measure. The variables look for students to talk about the replicability and generalizability of this technique, and to consider correspondences between different distributional states and the numbers the statistic would produce in each state.

The content of these variables have the potential to support credible inferences because they are rooted in the theories described in the construct map. I designed each variable and scoring rule to provide information about particular regions of the construct map. As a set, I designed them to cover the range of the construct, with information about each level. Because of

these commitments, the variables in this measurement system can all be interpreted in relation to the underlying instructional theory.

However, there were many aspects of the Data Modeling instructional design that this measurement system is not designed to capture. I have committed my efforts to use the quality of the method reviews as a proxy for classroom instruction, which means this measurement system can only provide valid inferences in this specific context. It cannot be used to support meaningful inferences about other sections of the instructional sequence. This is very important to remember because the resources needed to schedule all teachers at the same moments in the sequence make it tempting to observe a number of randomly selected instructional days. This would seriously undermine the validity of the data because the system is not designed to be sensitive to differences in many of the classroom interactions one would observe under this scheduling strategy.

There are also a number of important differences in whole class discussions that this measurement system can't detect. Although this system accounts for kinds of student talk, teacher strategies, and target concepts, it is not able to string them together into interactional turns. Because of this, it's impossible for me to determine if a question was a timely question, or if a comparison of two student ideas was a useful comparison. Instead, I have to rely on the assumption that the presence of the target concepts suggests more productive teaching strategies, but I can't test this conjecture with this measurement system. The Data Modeling design is not aimed at just supporting conversations, but at supporting a set of epistemic practices that students are then disposed to deploy in a variety of contexts to represent, measure, and model variability in data. With these limitations on the practicality of observational measurement the data that we have to make inferences about these practices vastly underrepresents the complexity of them.

Evidence Based on Observation Process

Observers used the measurement system during live observations in the vast majority of classes. This is an important part of the observation process because raters had a full range of vision in the class. We trained observers to sit in the back of the class, but in an area where they could hear and see as many students as possible.

However, classroom observations have a troubling history. An observer's presence in a classroom changes the social organization in ways that can significantly influence instruction. This is particularly true in the day of high stakes observations. We used a number of strategies to ward against this. First, I discussed the observations with all teachers during the summer professional development. While I didn't share the details of the observation system, I explained that the measurement system is intended to help inform our efforts to support their practice. I was clear that it was not an evaluation, and that the data would be held in confidence. One of the biggest concerns for the teachers was that we might identify poor instruction and inform school leaders. This was never our intention, and we worked to communicate this to teachers.

Even with these efforts it was clear that the presence of an observer changed the nature of the classrooms in consequential ways. First, it is likely that the teachers most committed to scheduling observations times were also the teachers that would at least use the Data Modeling materials. This is likely one reason why these variables were scored at such a high rate. Also, teachers sometimes felt as if they should perform in a unique way during observations. This often led teachers to ask observers questions after an observation directed at finding out "how they did." It was common for teachers to ask, "What did you think about the lesson?" as observers left the classroom. We trained observers to engage with these kinds of questions with

genuine, but value neutral responses such as “I really enjoyed seeing your class today” or “I’m so thankful you allowed me to come watch your class today.”

We also made great efforts to delineate between instructional coaches and observers. This was challenging because teachers would sometimes ask observers to provide suggestions for instructional strategies or next steps. We trained observers to never provide instructional guidance for two reasons. First, if teachers began to see observers as instructional coaches it would be very difficult to complete observations because the variables required sustained attention to reliably record. Second, the observers were knowledgeable about the measurement system, but received no training on the overall Data Modeling sequence. This paired with the lack of knowledge about each teacher’s students meant that even under the best intentions it is unlikely that the observers would be equipped to provide helpful advice.

The agreement during live double observations I reported in chapter four suggests that these observation processes provided contexts in which observers could reliably score the fidelity variables. However, we never viewed agreement as something that was established, but something to continually maintain. A significant portion of the weekend trainings consisted of group discussion about challenges during observations. These sometimes focused on procedural aspects, such as where to park at particular schools. Primarily, though, observers brought examples from their recent observations that they found challenging to score. The entire team would engage in a conversation about the conceptual elements of the variables as we worked out how a particularly challenging example would be scored.

In spite of all of these efforts there are always threats to validity in live observations. In addition, some of our efforts, while supporting more valid data collection, have provoked new questions for future research. Under this project’s design the teachers never saw the data

collected in their class, and we tried to engage observers as little as possible in the classroom interactions. We tried to make the observers and the data as "invisible" as possible to the teachers. However, this meant that the teachers could not use the data or the measurement framework as a tool to think about their instructional practice. In the future I'd like to consider ways in which the teachers can become a meaningful part of the measurement process, and ways to represent the data that would be useful to teachers as they reflect on the lessons.

Evidence Based on Internal Structure

The item threshold estimates represented in the Wright map empirically support many of the conjectures in the construct map. It is important to note that I have explained the progress of this work in a way that seems very linear. However, the construct map in this dissertation is the result of multiple iterations with a number of revisions. Since this is the first measure of its kind, my initial theorization of the fidelity construct was challenged as I used it to measure, model, and interpret differences in classroom instruction. This happened during the early stages of operationalization as I noticed that there were important differences not represented in the map. For example, the original map did not have what is now level three. However, as I piloted the early measurement instrument I recognized that there were a number of classes that talked about key ideas, such as scale or median, but did not address the epistemic nature of the ideas. The measurement model also challenged the theory since I initially expected teachers to outright reject the curriculum design at lower levels, and that teachers' use of discussion questions would be evidence that students were likely talking about important ideas. Modeling the observation data made it clear that this was not the case with this sample, although previous studies did include teachers that rejected the design so we expect to see this from time to time. This is why I revised the construct map to describe lower levels as classrooms where curricular tools were

being deployed, and the higher levels as classrooms where the tools were being used to produce conversations about the intended ideas. On the other hand, I revised the variables and scoring along the way too. Challenges to the construct map did not always result in changes to the construct map. More often I was led to change the ways I operationalized the map to provide more valid evidence. The changes to the sInvented (are students talking about invented methods?) variable that I reported in chapter 3 are just one example. After the first year of data collection I noticed that the sInvented variable, which indexes the lowest level of the construct map, was more difficult to observe than many variables indexing higher levels. This did not lead to a change in theory, but a closer look at the observable evidence. Only then did I realize that the definition of the variable required students to be talking about their own inventions, which observers were unable to determine when the inventions were anonymous. Unfortunately the change in the variable led to its exclusion from this study, but it's refinement will provide for more meaningful interpretations in future work.

A first glance at this might lead the reader to think that this is evidence against measurement validity since I modified the construct map and the variables along the way as it was challenged by the data. Surely I just gamed the system! However, this perspective, although common, is "diametrically opposite to good practice" (Wilson, 2004, p. 161). As I described in chapter two, measurement is fundamentally a modeling activity in which theory influences observation, but observation often leads one to change their original theorization. The previous paragraph is but an example of figure 4 (from chapter two) in action. So, this iterative process is not a threat to validity, but evidence of ongoing modeling work to maintain and improve validity.

The Wright map shows that the item scores increase in difficulty as they correspond to higher levels of the construct map. The item fit statistics were all within the conventionally

acceptable range, and the items were all positively correlated with the total score. The item threshold estimates at levels 3, 4, and 5 leave a number of issues uncertain, though. This sample provided very few observations in some of the levels, so the estimates have large errors. Because of this it is difficult to say if the model supports the construct map and the item design. For example, there are some items in level 3 that are more difficult than level 4 items. The same is true between levels 4 and 5. However, since the estimates have such large errors more work needs to be done to determine if this is a problem with the underlying theory, the variables, or the scoring rules.

The Wright map also revealed that items at the bottom two levels of the construct map were very easy to observe in this population. As I reported in table 9, over 90% of the teachers in this study were scored at level 1 or 2 for these items. Under some circumstances this kind of information can be evidence that there is no need to retain the items in the instrument because they don't differentiate between qualitatively different groups. However, the sample that I used to calibrate this measure consisted of teachers that participated in significant professional development and ongoing coaching support. We would likely observe a larger proportion of classrooms not scored on these levels in a population that was not provided with similar resources and opportunities to learn. For example, if teachers were simply given the Data Modeling materials without adequate professional development I expect many would not make use of the most important tasks. It is important to retain these variables in the system to account for these classrooms in the future.

There are two items, though, that might be reconsidered in light of the Wright map. For all of the items I used the highest score observed for more than one five-minute segment. It is possible that this was too generous for the sTalk and tDiscussQ items. The sTalk item indexes

student talk using two levels, talking procedurally and talking both procedurally and conceptually. The tDiscussQ item indexes teachers' questions using two levels, teachers using discussion questions and teacher using them while pressing kids to elaborate their thinking. Ten minutes of attention for other items, such as those focused on unit specific concepts, often makes sense. For example, if a class talked for ten minutes in a high quality way about the median it is very possible that this is sufficient. However, this is less certain for the sTalk and tDiscussQ variables because it might be problematic if students only talked for ten minutes of the class or if a teacher only asked questions for ten minutes. In large part, the classes scored on these variables had many more than the minimum of 2 5-minute segments at the level in which they were scored. However, this scoring rule is one aspect of the measurement system that should receive additional attention in future iterations.

Evidence Based on Relationships with External Data

The coaches' professional judgments about classroom instruction agreed with the measurement system in many respects. This comparison suggested that the observers and coaches had a similar vision for the kinds of conversations we were trying to support. It is important to remember that these two groups, coaches and observers, did not have any contact during the project. The comparison also suggests that there were a number of important distinctions the coaches made in their professional judgments that the observation system was not able to make. It isn't that the observation measure disagreed with the coaches, but that the measure was blind to important instructional characteristics, such as the quality of the sequencing of questions over the course of a whole class discussion.

While there is still much uncertainty in the relationship between the observation measures and the student measures, the coefficients in table 14 suggest that there is likely a positive

relationship between units 1, 2, and 5 observation measures with the student measure. This was especially true for the corresponding student subscale measures as the coefficients from all three of these were significant at the .10 level. However, the unit 3 measures were not correlated with the outcome measures. It is possible that this was because of the lack of classrooms at the higher ends of the construct map in unit 3, but this is an area that warrants a closer look once all year two student data is ready for analysis.

Evidence Based on Measurement Consequences

The evidence from this measurement system was already consequential in informing our ongoing professional development for teachers. For example, in the first year of the study it was clear that teacher were talking about grouping much more than the other ideas in unit 1, and that many needed much more support to understand concepts related to measuring variability in unit 3. These examples led to changes in the professional development to support the specific challenges teachers were having. For the example from unit 1, the consequences show the importance of indexing unit specific concepts. Without attention to these it would have been impossible to tell which ideas classes talked about more than others. In the unit 3 example the variables focusing on a shared understanding of variability as a measurable attribute were critical. In addition, the unit specific concepts allowed us to see classrooms where kids were talking, and teachers were using many of the strategies we support, but these resulted in conversations suggesting that measures of center and measures of variability are conceptually the same. This measurement system provided a key source of data, along with coaching logs and observations during professional development, to generate knowledge about the use of the Data Modeling materials during classroom instruction.

At the end of two years of implementation this measurement system is the one of the primary tools that will allow us to determine the nature of the instruction in intervention classrooms, and the extent to which it resembles the kinds of instruction the Data Modeling program is intended to support. Other researchers in our project built case studies of particular types of teachers to develop fine grained understandings of how teachers recontextualize teaching strategies in particular classroom contexts (Pfaff, 2013) and how teachers and students framing of classroom activity changes over time (Shinohara & Lehrer, 2013). While these studies have made significant theoretical contributions to our understanding of the Data Modeling design, and will likely influence future iterations of this measure, they are unable to inform questions about classroom instruction across the entire sample of teachers. I will discuss this issue further when I address research question two, but here I will just note that without this measure it would be unclear what theory we are testing with the random assignment. This measure will contribute to our understanding of any differences, or lack of differences, between children in the intervention group and children in the comparison group.

Table 16: Summary of research question 1 discussion

	Support	Challenges
Measurement System Content	• Item scores inform construct map • Indexes students, teachers, and content • Unit specific concepts	• 1 Day Observations • Unable to link particular interactional turns
Observation Process	• Framed observations for teachers • Observers trained to not be coaches • Rater agreement	• Live observations change classroom contexts • Teachers still viewed observations as performances • Teachers not involved in process
Internal Structure	• Wright Map • Acceptable Fit Statistics • Item scores positively correlated with total score	• Noisy classroom estimates • Sparsely populated categories at levels 3, 4, and 5 • *Student Talk* and *Discussion Question* items
Relationships with External Data	• Agreement with coach judgments • Positive relationships between most unit observation scores and student post test measures	• Some coach distinctions are invisible to the measure • Unit 3 not related to student post test measure
Measurement Consequences	• Informed PD • Informs the nature of the realized intervention	

Research Question Two:
What do we learn about differences in classroom instruction with such a measurement system?

As I said before, this measurement system is the primary source of evidence of the nature of instruction in Data Modeling classes. Anyone that has worked with teachers knows that some teachers will make use of the instructional designs and teaching strategies in powerful ways while others will use the tools, but in more superficial ways. With this measurement system I have attempted to make explicit the characteristics that constitute "powerful" and "superficial,"

as well as point in between. In doing so, I have provided empirical evidence of classroom instruction that is rooted in the underlying instructional theories guiding the design.

The logit scale estimated by the measurement model suggested that the differences between the qualitative states of the construct map are not all equal. First, the difference between level 1 and level 5 items on the map was over four logits. To put this into perspective, Ms. West would have a 97% chance of being scored on an item at the first level of the construct, but only a 39% chance at the highest level. Since this is a logarithmic scale the probability shifts from the bottom to top levels are not the same at all points on the scale. Mr. North, by comparison, had a 99% change of being scored on a level 1 item, which is similar to Ms. West. However, he still had a 70% change of being observed on an item at the highest level. Ms. West's score was just below the average classroom score, so this is a shift that is representative of a number of observations. Mr. West is representative of classroom scores at the highest end of the distribution in this sample. The lowest classroom score in this sample was -2.18. This classroom would have a 77% chance of being scored on at least level one for items that include this level, but only an 8% chance of being observed scored at the highest level for items stretching to level 5. Although the changes in probabilities are much more extreme at the lower levels, it is clear that the distance between solely implementing the tasks and materials and using them to talk about the ideas in intended ways is a significant one.

The difference between the average level 1 threshold and the average level 2 threshold was 1.3 logits, and the difference between levels 2 and 3 was 1.32. This suggests that the differences can be thought of as very similar. So the difference between a class only using materials and a class beginning to support student discussion is similar to the difference between starting a class discussion, and facilitating a conversation that talks about key ideas without

addressing their epistemic underpinnings. If you thought of this as the work it takes at a teacher to move the class to each level this shows that even at the bottom levels of the construct map it takes significant teacher work to make these shifts.

The difference between level 3 and 4 was .55 logits, and between 4 and 5 was .49. This might suggest that the hardest work for a teacher is in getting the ideas on the table, and once out it is not as difficult to get kids to talk about them in productive ways. However, we should be very careful about this interpretation because of the number of thresholds with large standard errors. In fact, the distance between level 3 and level 5, 1.04 logits, is very similar to the distances at the lower levels. It is possible that it is easier to get kids to begin to talk about the epistemic implications, but it is still very challenging to get them to talk about the epistemic implications in a robust way. In unit 1 this would represent the difference between talking about an idea like scale (level 3), talking about what scale shows about the data (level 4), and talking about what scale can show and hide about the data (level 5). In units two and three it is the difference between talking about a statistic (level 3), beginning to talk about talking about a statistic's replicability or generalizability (Level 4), and building correspondences between the statistics and different distributional states to talk about replicability and generalizability (level 5). And in unit 5 this represents the difference between talking about different sample sizes of trials (level 3), the effect of changing sample size on the center of a sampling distribution (level 4), and the effect of the changes on both the center and variability of a sampling distribution (level 5).

I have been talking about the item level thresholds in terms of the average threshold for each level, but these estimates also suggest that the distances between levels are not the same for all items. Consider the iScale and the iSampleSize items. The level three threshold for iScale is

.168 logits and for iSampleSize is -1.14. This suggests that it is much easier to support students to talk about differences in sample size in unit 5 than scale in unit 1. However, once students are talking about scale it is much easier to get them to discuss what scale shows and hides about a distribution, which has a threshold of .684, than to move a discussion of sample size into a conversation about the effects of sample size on the center and variability of a sampling distribution, which has a threshold of 1.305. This is very important to consider when supporting teachers. Sometimes it is easy to get an idea on the table, but difficult to move beyond initial conceptions. On the other hand, sometimes the idea itself is difficult to begin a conversation about, but once students see it they can much easily develop a more robust understanding of it. Often times professional development talks about the development of a mathematical idea in domain general terms, which ignores these important distinctions for teachers.

In addition to the differences between individual items, there were noticeable differences from unit to unit in the classroom estimates. As I reported in chapter 4, unit 1 whole class discussions resembled our intentions much more than units 2, 3, and 5 on average. This pattern held true for most teachers in the study. Ms. West, for example, had a classroom estimate in unit 1 of -.12, but had estimates of -.64, -.76, and -.39 for units 2, 3, and 5 respectively. This again points to the effect of content domains on classroom instruction. The ideas in unit 1 were much more accessible to teachers, which made it easier to support a class discussion about them (although still hard work!). On the other hand, statistics and probability proved to be more difficult concepts to grasp, which made instruction more challenging.

Due to these kinds of variability it is difficult to characterize the realized Data Modeling instruction in the intervention classes in general terms. It is clear that this experiment can not be called a test of the Data Modeling instructional theory since so few students were given

opportunities in units 6 and 7 to model sample to sample variability and make inferences in light of this variability. While there was great variability in the fidelity of implementation in the Data Modeling classes, on average lessons in unit 1 were more faithful to the intentions of the design than in the lessons in the later units. Even in unit 1, though, classes talked about what design choices show about the data much more often than what the choices show and hide. Instruction in unit 3 resembled the intentions of the design the least, especially in the first year. Many classes did not even build a shared understanding of the concept of variability as a measurable attribute.

The coefficients that I reported in table 14 suggest a consistent, if not always statistically significant, positive relationship between the variability in classroom instruction and the variability in student outcomes during the first year of the study. This is particularly noticeable for the models using corresponding subscales as outcome measures and pretest covariates. In units 1, 2, and 5 a one-logit change in the fidelity scale is related to significant change in the student outcome subscales, .16 logits in unit 1, .17 in unit 2, and .29 in unit 5.

All together, these findings shed new light on the cost of improving instructional quality at a large scale. In this study we provided materials, software, and ongoing professional development and coaching. All of these are extremely expensive and require teams of highly trained professionals. Even with these resources there were many classes where observers had a low probability of observing items that are higher on the construct map. These resources successfully supported most of the teachers to make use of the materials and to begin to engage students in conversations around the materials. However, if we think that we would want more than a 50% chance of observing an average item at the higher levels then it is clear that there is much work left to do.

Future Work

This study leaves much work to be done. The sample in this study is less than desired, so it will be important in the future to collect measures on more samples to add to the model. This will provoke additional iterations and refinement to the scoring rules in this study. This is especially true in the case of unit specific items at levels 3-5 of the construct map. It's likely that additional data will give a clearer picture of how to improve measurement at these levels.

Additionally, this kind of measurement should be conducted in other mathematical domains. Is it true that the differences in levels observed here hold for, say, irrational number? What about geometry and space? Will similar instrumentation be productive, or are there new demands in these contexts that would require an entirely new system? Can classroom work products be more successfully incorporated into the measurement system? Questions like these are important to address in order to build a better understanding of the "distance" between qualitatively different states of instructional quality across content domains.

More work should also be done to try and quantify the resources needed to move along particular levels of the fidelity construct map. Clearly you at least need to produce the materials to move into the first level, which is a non-trivial cost in itself. The resources we deployed in this study were successful to move the center of the distribution of class estimates to the same area of the scale as the level three items, but it's not clear what additional resources are needed for additional movement. Time is likely one additional resource, and one that should be studied more in the future.

Last, future work is needed to identify ways to make the measurement system and the data it can produce to support teachers to think about their own practice. I am now creating representations of the construct map at each level with video examples of instruction at the

different levels. I am also coordinating these representations with the tools designed to support teacher practice, such as Erin Pfaff's discourse moves (Pfaff, 2013). In the future I plan on studying the ways this kind of representation can support teachers as they collaborate with one another, coaches, and researchers. Many current representations of teaching practice are at the extreme ends of a scale, either exemplary practice or poor practice. I see this as a productive framework for representing practice along a trajectory in order to support teachers to think about moving incrementally towards an ideal goal, but also being able to track progress along the way.

This study shows that it is possible to use a measurement framework to study classroom practice from a fidelity perspective. By explicitly representing program theory, mapping this theory onto an observable measurement system, and modeling the observed data to estimate a measurement scale, I believe that this study is an example of how a construct modeling approach to fidelity measurement would move this field of work forward.

BIBLIOGRAPHY

Abry, T. D., Rimm-Kaufman, S. E., Larsen, R. A., & Brewer, A. J. (2011). Applying New Methods to the Measurement of Fidelity of Implementation: Examining the. *Educational Research,* 78(1), 33-84. https://www.sree.org/conferences/2011f/program/downloads/abstracts/375.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation,* 33, 5-18. http://dx.doi.org/10.1002/ev.1443

Brown, S. A., Pitvorec, K., Ditto, C., & Kelso, C. R. (2009). Reconceiving fidelity of implementation: An investigation of elementary whole-number lessons. *Journal for Research in Mathematics Education*, 363-395. http://www.jstor.org/stable/40539344

Chard, D. J., Baker, S. K., Clarke, B., Jungjohann, K., Davis, K., & Smolkowski, K. (2008). Preventing early mathematics difficulties: The feasibility of a rigorous kindergarten mathematics curriculum. *Learning Disability Quarterly*, 11-20. http://www.jstor.org/stable/30035522

Chen, H. T., & Rossi, P. H. (1980). The multi-goal, theory-driven approach to evaluation: A model linking basic and applied social science. *Social forces*,*59*(1), 106-122. http://www.jstor.org/stable/2577835

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale

cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127-166. http://www.jstor.org/stable/10.5951/jresematheduc.42.2.0127

Cobb, P., Confrey, J., diSessa, A., Lehrer, R. & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, 32(6), 3–12. http://dx.doi.org/10.3102/0013189X032006003

Cohen, D.K. (2011). *Teaching and Its Predicaments*. Cambridge, MA: Harvard University Press.

Cordray, D. S. & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E., McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation (pp 103-124).* Washington, DC: American Psychological Association. http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=2005-16601-006

Crawford, L., Carpenter, D. M., Wilson, M. T., Schmeister, M., & McDonald, M. (2012). Testing the Relation Between Fidelity of Implementation and Student Outcomes in Math. *Assessment for Effective Intervention*, *37*(4), 224-235. http://dx.doi.org/10.1177/1534508411436111

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23-45. http://dx.doi.org/10.1016/S0272-7358(97)00043-3

DiSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *The journal of the learning sciences*, *13*(1), 77-103.

Doabler, C., Baker, S. K., Smolkowski, K., Fien, H., Clarke, B., Cary, M. S., & Chard, D. (2011). Impact and Implementation Analyses of the ELM Kindergarten Mathematics

Intervention. *Society for Research on Educational Effectiveness*. http://eric.ed.gov/?id=ED530368

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. *The handbook of evaluation: Policies, programs, and practices*, 56-75. http://cgu.edu/PDFFiles/sbos/Roles for theory in contemporary.pdf

Ertle, B., Ginsburg, H. P., & Lewis, A. E. (2006). Measuring the efficacy of Big Math for Little Kids: A look at fidelity of implementation. In *annual conference of the American Educational Research Association, San Francisco*.

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W. S., & Erickson, S. (1987). Innovation in education and criminal justice: Measuring fidelity of implementation and program effectiveness. *Educational Evaluation and Policy Analysis*, *9*(4), 300-311. http://www.jstor.org/stable/1163769

Ford, M. J. (2010). Critique in academic disciplines and active learning of academic content. *Cambridge Journal of Education*, *40*(3), 265-280.

Ford, M. J., & Forman, E. A. (2006). Redefining disciplinary learning in classroom contexts. *Review of research in education*, *30*, 1-32.

Forman E.A. & Ford M.J. (2013). Authority and accountability in light of disciplinary practices in science. *International Journal of Educational Research*

Fuchs, L. S., Fuchs, D., Yazdian, L., & Powell, S. R. (2002). Enhancing first-grade children's mathematical development with peer-assisted learning strategies. *School Psychology Review*. http://psycnet.apa.org/psycinfo/2003-04356-010

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*(3), 606-633.

Hagermoser Sanetti, L. M., & Kratochwill, T. R. (2009). Toward Developing a Science of Treatment Integrity: Introduction to the Special Series. *School Psychology Review*, *38*(4), 445-459. http://eric.ed.gov/?id=EJ867973

Hulleman, C.S. & Cordray, D.S. (2009). Moving from the lab to the field: the role of fidelity and achieved relative intervention strength. *Journal of research on educational effectiveness*, (2) p. 88-110 http://dx.doi.org/10.1080/19345740802539325

Hall, G., & Loucks, S. (1975). Levels of use of the innovation: A framework for analyzing innovation adoption. *Journal of Teacher Education*, 26(1). http://dx.doi.org/10.1177/002248717502600114

Horn, I. S. (2008). Chapter 3: Accountable Argumentation as a Participation Structure to Support Learning through Disagreement. *Journal for Research in Mathematics Education. Monograph*, 97-126.

Jones, S. R. & Kim. M. J. (2011). Enacting a New Curriculum: A Teacher's First Attempt with Data Modeling. Poster presented at the 2011 meeting of the National Council of Teachers of Mathematics Research Pre-session, Indianapolis, IN.

Kuhn, T.S. (1970). *The Structure of Scientific Revolutions, 2nd. ed.* Chicago: University of Chicago Press.

Latour, B. (1999). *Pandora's hope. Essays on the reality of science studies.* Cambridge, MA: Harvard University Press.

Lehrer, R., & Romberg, T. (1996). Exploring children's Data Modeling. *Cognition & instruction*, 14(1), 69–108.

Lehrer, R., & Kim, M.J. (2009). Structuring variability by negotiating its measure. *Mathematics education research journal*, 21(2), 116–133.

Lehrer, R., Kim, M. J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM*, *43*(5), 723-736. http://www.fisme.science.uu.nl/staff/marjah/download/ZDM/Lehrer-etal_2011_measures-of-distribution.pdf

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data.* (pp. 149-176). New York: Lawrence Erlbaum.

Lehrer, R., Kim, M.J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning*, 12(3), 195–216.

Lieber, J., Butera, G., Hanson, M., Palmer, S., Horn, E., Czaja, C., ... & Odom, S. (2009). Factors that influence the implementation of a new preschool curriculum: Implications for professional development. *Early Education and Development*, *20*(3), 456-481. http://www.tandfonline.com/doi/abs/10.1080/10409280802506166 - .Uo_RCmRgaDQ

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. *New Directions for Program Evaluation*, 27, 7-28. http://dx.doi.org/10.1002/ev.1398

Lipsey, M.W. (1993). Theory as method: small theories of treatments. *New directions for program evaluation*, 57, p. 5-38 http://dx.doi.org/10.1002/ev.1637

Masters, G. N. (1982). A Rasch model for partial credit scoring.*Psychometrika*, *47*(2), 149-174.

McNaught, M. D., Tarr, J. E., & Sears, R. (2010, April). Conceptualizing and Measuring Fidelity of Implementation of Secondary Mathematics Textbooks. Results of a Three-Year Study. In *Annual Meeting of the American Educational Research Association (Denver, CO*.

http://cosmic.missouri.edu/aera10/AERA2010_Measuring_Implementation_Fidelity_100505.pdf

Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, *35*(11), 1012.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2) p. 13-23 http://www.jstor.org/stable/1176219

Mills, S. C., & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning system. *Educational Technology Research and Development*, 48(4), 21–41. http://dx.doi.org/10.1007/BF02300498

Mislevy, R. J. (1993). Foundations of a new test theory. *Test theory for a new generation of tests*, 19-39. http://www.dtic.mil/dtic/tr/fulltext/u2/a215437.pdf

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463-469.

Mowbray, C.T., Holter, M.C., Teague, G.B., Bybee, D. (2003). Fidelity criteria: Development, Measurement, and Validation. *American Journal of Evaluation* 24(3) p. 315-340. http://www.stes-apes.med.ulg.ac.be/Documents_electroniques/EVA/EVA-GEN/ELE EVA-GEN 7386.pdf

Munter, C., Garrison, A., Cobb, P., & Cordray, D. (2010). Evaluating Math Recovery: Measuring Fidelity of Implementation. *Society for Research on Educational Effectiveness*. http://eric.ed.gov/?id=ED514496

Munter, C., Wilhelm, A. G., Cobb, P., & Cordray, D. S. (2014). Assessing Fidelity of Implementation of an Unprescribed, Diagnostic Mathematics Intervention. *Journal of Research on Educational Effectiveness*, *7*(1), 83-113.

Nelson, M.C., Cordray, D.S., Hulleman, C.S., Darrow, C.L., & Sommer, E.C. (2012). A Procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The journal of behavioral health services and research*. 1(22) http://dx.doi.org/10.1007/s11414-012-9295-x

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, 78(1), 33-84. http://dx.doi.org/10.3102/0034654307313793

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning*, 5(2&3), 131–156.

Pfaff, E. (2013). The role of coaching in rehearsals of ambitious math discussions. *Proceedings of the 2013 Annual Meeting of the American Educational Research Association.* San Francisco, CA: AERA.

Pickering, A. (1995). *The mangle of practice. Time, agency and science.* Chicago: The University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226668253.001.0001

Sandoval, W. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences*, *23*(1), 18-36.

Shinohara, M., & Lehrer, R. (2013). *Epistemology in Flux: How Teachers Interpret the Practice of Inventing Statistics.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Spillane, J. P. (2004). *Standards deviation: how schools misunderstand education policy*. Cambridge, MA: Harvard University Press.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning*, *10*(4), 313-340.

Swafford J.O., Jones G.A., Thornton C.A. (1997). Increased knowledge in geometry and instructional practice. *Journal of Research in Mathematics Education*, 28(4): p. 467–483. http://dx.doi.org/10.2307/749683

Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199278220.001.0001

W.K. Kellogg Foundation. (2004). Using logic models to bring about planning, evaluation, and action: logic model development guide. Battle Creek, MI.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, 61, 620-630. http://dx.doi.org/10.1037/0022-006X.61.4.620

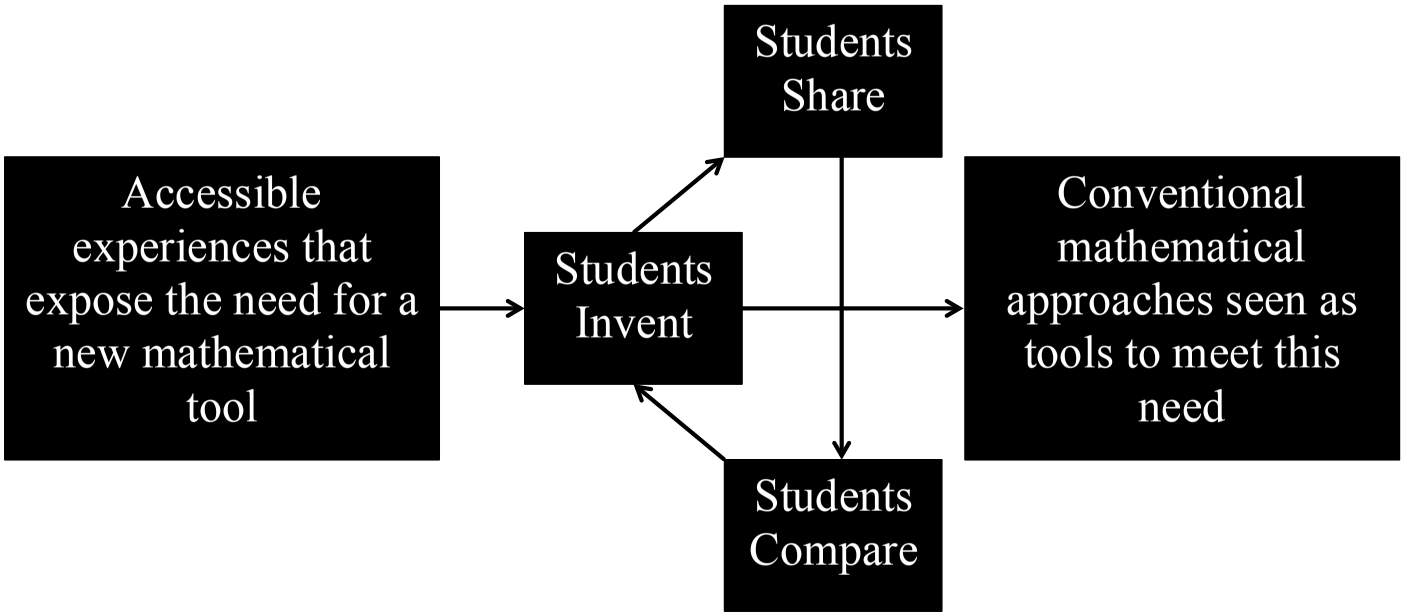
Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Program Evaluation*, 76, 41-56. http://dx.doi.org/10.1002/ev.1086

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P., & Peck, K. (2009). A Multisite Cluster Randomized Trial of the Effects of CompassLearning Odyssey [R] Math on the Math Achievement of Selected Grade 4 Students in the Mid-Atlantic Region. Final Report. NCEE 2009-4068. *National Center for Education Evaluation and Regional Assistance*. http://eric.ed.gov/?id=ED507314

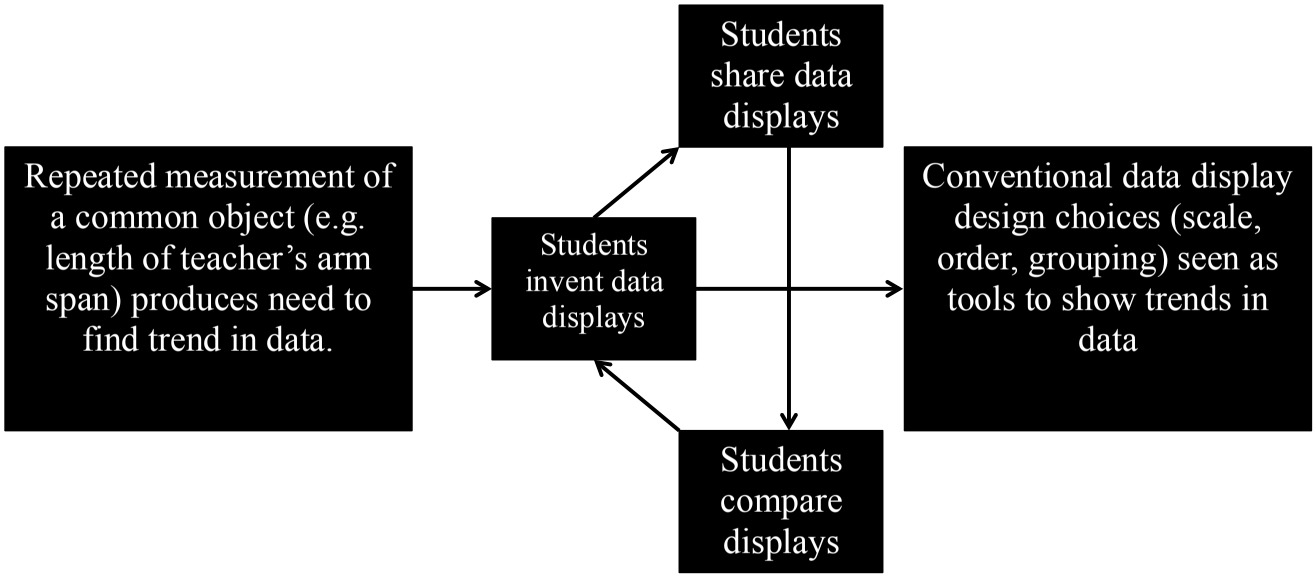
Wilson, M. (2004). *Constructing measures.* Mahwah (NJ): Lawrence Erlbaum.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalised item response modelling software*. ACER press.

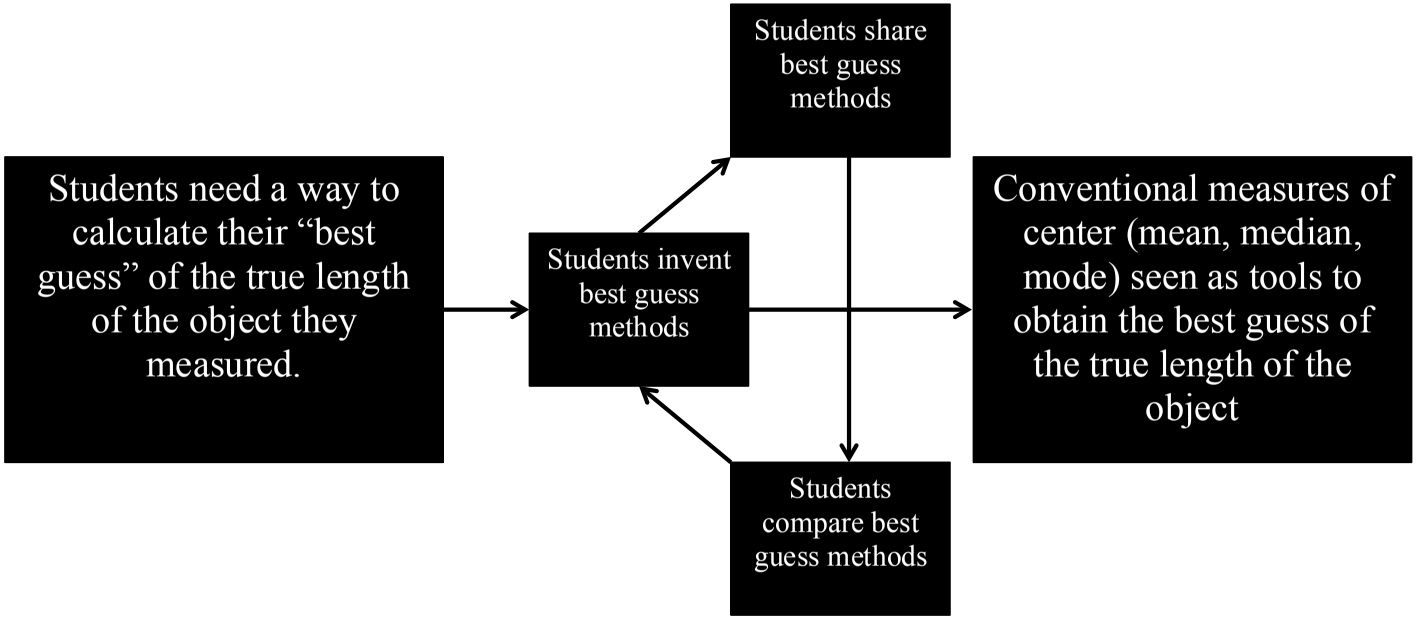
Appendix A
Conceptual Logic Model



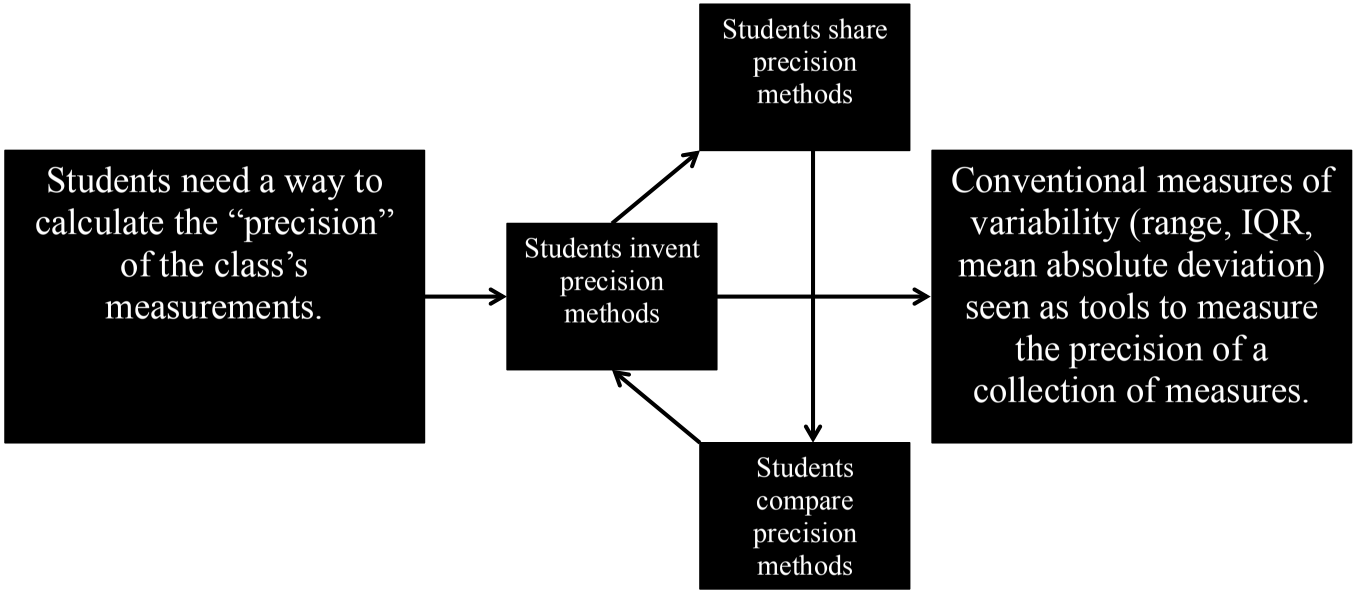
Unit 1 Operationalized Logic Model



Unit 2 Operationalized Logic Model

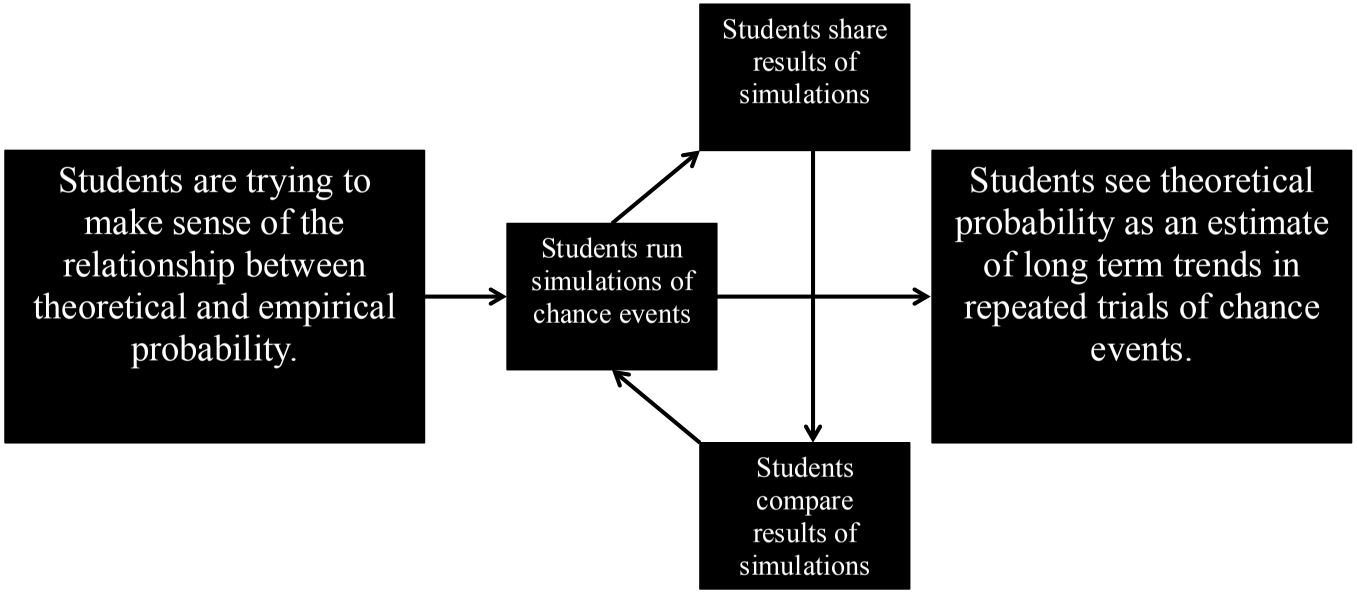


Unit 3 Operationalized Logic Model

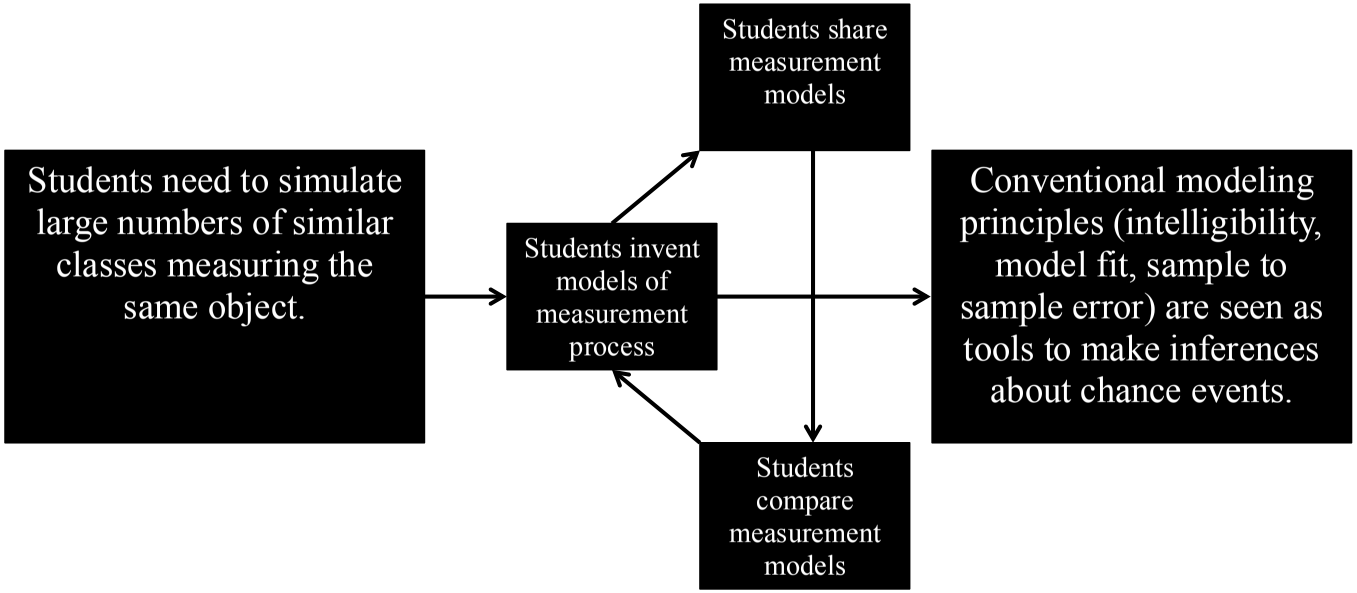


Unit 5 Operationalized Logic Model

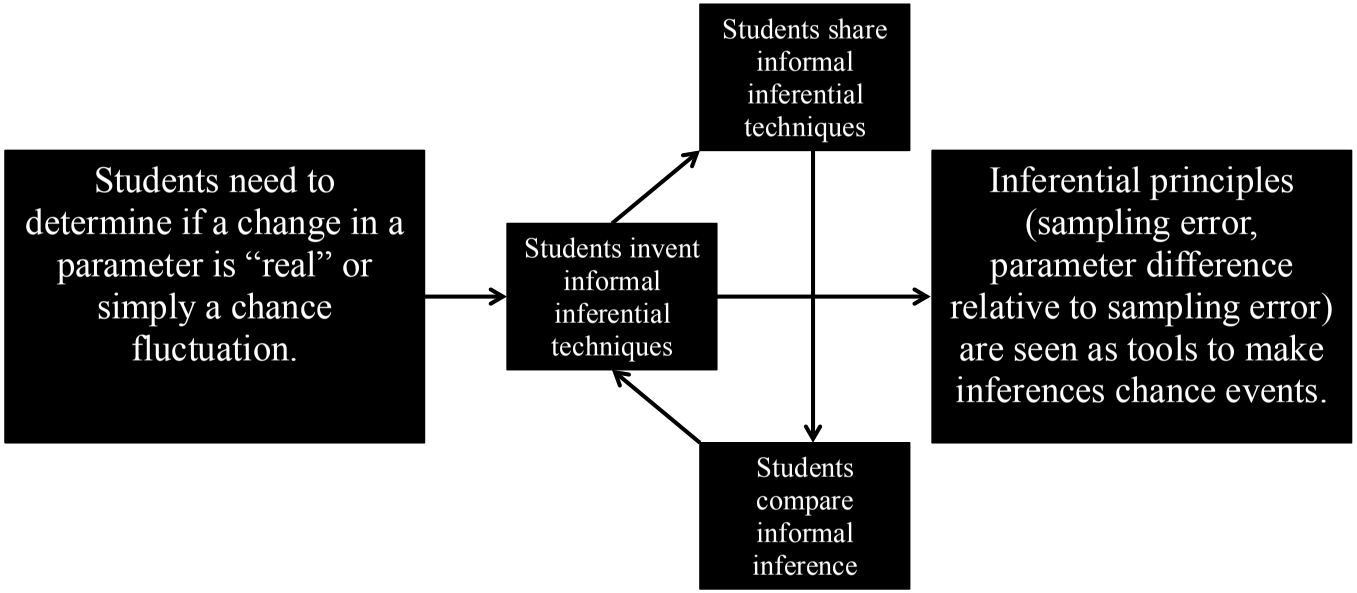
During unit 4 students use statistics in new contexts to consider meaning and use with different kinds of data. Unit 4 was not observed.



Unit 6 Operationalized Logic Model



Unit 7 Operationalized Logic Model



Appendix B
Unit Specific Construct Maps of Instructional Practice

Unit 1

Level	Practice	Example
5	Teacher selects and juxtaposes contrasting examples based on the DaD and MRC constructs that span the levels 2-4 of DaD and levels 1-3 of MRC. Teacher questions and strategies, such as tracing data cases between displays, help students notice the relations between the big ideas of the unit. 1) Discuss how the data is ordered in displays 2) Discuss the scale of the displays 3) Discuss how the data is grouped in displays 4) Discuss what displays show 5) Discuss what displays hide 6) Discuss the effect of design choices on the shape of data	The teacher could ask many of the questions from unit 1, but in addition leads the class to consider how these principles are related to changes in the shape of the data, and to what the displays show and hide about the data. They might use questions similar to the ones below to help students make these connections. 1) "What is misleading about this display? Why?" 2) "What does this display show that the other ones hide? What choices help to show this?" 3) "Where is this data point on the other graph?"
4	Teacher selects displays to be shared based on the DaD construct that span levels 2-4. The class talks about these design principles and begins to address the epistemic implications. 1) Discuss how the data is ordered in displays 2) Discuss the scale of the displays 3) Discuss how the data is grouped in displays 4) Discuss how design choices determine the shape of the data.	Teacher might ask questions such as the following without any coordination between the implications of different display features: 1) "Jay's group put their data in order, but some did not. Do you think this is an important difference?" 2) "If no one measured 145, do we need to have it on our graph?" 3) "Some grouped their data by tens, while others did by fives. Which do you like better?" 4) "Does anyone see a trend in this graph?"
3	Student invented methods are seen as an instructional resource to promote key concepts such as: 1) Discuss how the data is ordered in displays 2) Discuss the scale of the displays 3) Discuss how the data is grouped in displays	Teacher might ask questions such as the following: 1) Which of these displays ordered their data? 2) How did this display group the data? 3) Does this one have scale?
2	Student invented displays are seen as an instructional resource to support student discourse.	Teachers ask questions and press students to elaborate their thinking while students talk about the invented displays. However, the conversation does not address ideas about scale, order, grouping, or what the displays show and hide.
1	Student invented displays are seen as an instructional resource.	Teachers make use of the following tasks: 1) Collecting repeated measures data 2) Inventing data displays 3) Display review

Unit 2

Level	Practice	Example
5	Teacher selects and juxtaposes contrasting student-invented methods and leads a whole class discussion that gives students the opportunity to think about the concepts from level 3 on the Conceptions of Statistics construct map, but also works to establish the relationships among the different measures. (1) What is important to consider in measuring the "best guess" of a group of repeated measurements. (2) The type of thinking behind each measure. There should be examples of agreement (mode), center (median), and fair share (mean). (3) The clarity of each measure (4) The generalizability of each measure (5) How each measure corresponds to changes in distribution	Teacher uses any of the questions given in the level 4 examples, but in **addition** might ask questions similar to the following: 1) "We have seen that the mode and the median use different parts of the data. How do these different approaches affect our 'best guess'?" 2) "The mean, median, and mode all use different parts of the data to measure precision. What are the advantages and disadvantages of each approach?" 3) "What happens to the mean is the largest measurement is much larger than the rest of the data? The median? The mode?" 4) "When we look at the different ways to measure "best guess', which seems to give the best indication of the actual length? Why do you think so?"
4	Teacher selects student-invented methods to be shared and leads a whole class discussion to guide students to consider one or more of the principles of measuring center. The class begins to discuss the epistemic implications of the ideas. (1) What is important to consider in measuring the "best guess" of a group of repeated measurements. (2) The type of thinking behind each measure. There should be examples of agreement (mode), center (median), and fair share (mean). (3) The replicability of each measure (4) The generalizability of each measure	The teacher might ask any of the questions the following, but in a manner that does not establish the relationships between the measures. 1) "Would this statistic be a good indication of our 'best guess' of the actual length?" 2) "What does the mode focus on about the data? Is this good or bad? Why?" 3) "If everyone in the class followed these directions would we all get the same answer? Why is this so important?" 4) "Is there a scenario when this method might give us a misleading "best guess" measure?" 5) "Where in the data would this statistic be?"
3	Student invented methods are seen as an instructional resource to promote key concepts such as: 4) Mean 5) Median 6) Mode	Teacher might ask questions such as the following: 1) "How did john calculate his method?" 2) "Can everyone understand this method?" 3) "Did Jane do the median right?" 4) "What are the steps to calculating the mean?"
2	Student invented statistics are seen as an instructional resource to support student discourse.	Teachers ask questions and press students to elaborate their thinking while students talk about the invented statistics. However, the conversation does not address ideas about median, mean, mode, generalizability, replicability, or correspondence to distribution.
1	Student invented statistics are seen as an instructional resource.	

Unit 3

Level	Practice	Example
5	Teacher selects and juxtaposes contrasting student-invented methods and leads a whole class discussion that gives students the opportunity to think about the concepts from level 4, but also works to establish the relationships among the different measures. (1) What is important to consider in measuring precision of a group of repeated measurements. (2) The type of thinking behind each measure. There should be examples of range, center clump (IQR), and deviation from a measure of center. (3) The scale of each measure. (4) The clarity and generalizability of each measure (5) How each measure corresponds to changes in distribution	Teacher uses any of the questions given in the level 4 examples, but in addition might ask questions similar to the following: 1) "The range, IQR, and average deviation all use different parts of the data to measure precision. What are the advantages and disadvantages of each approach?" 2) "What does a measure of zero tell you about the data if you are using the range? The IQR? The average deviation?" 3) "When we look at the different ways to measure precision, which seems to give the best indication of how much our data tends to agree? Why do you think so?"
4	Teacher selects student-invented methods to be shared and leads a whole class discussion to guide students to consider one or more of the principles of measuring variability. The class begins to discuss the epistemic implications of the ideas. (1) What is important to consider in measuring precision of a group of repeated measurements. (2) The type of thinking behind each measure. There should be examples of range, center clump (IQR), and deviation from a measure of center. (3) The scale of each measure. (4) The clarity and generalizability of each measure	Or the teacher might ask any of the questions the following, but in a manner that does not connect the student inventions to the mathematical concepts behind each type of measure and does not establish the relationships between the measures. 1) "Would this statistic be a good indication of how much our measurements tend to agree?" 2) "What part of the data does this statistic use? 3) What does it mean if this statistic is zero? 4) "What happens to this measure as the data gets more spread out?"
3	Student invented methods are seen as an instructional resource to promote key concepts such as: 7) Range 8) Center clump measures (i.e. IQR) 9) Deviation Measures	Teacher might ask questions such as the following: 1) "How did john calculate his method?" 2) "Can everyone understand this method?" 3) "Did Jane do the IQR right?" 4) "What are the steps to calculating the range?"
2	Student invented statistics are seen as an instructional resource to support student discourse.	Teachers ask questions and press students to elaborate their thinking while students talk about the invented statistics. However, the conversation does not address ideas about range, center clump, deviation, generalizability, replicability, or correspondence to distribution.
1	Student invented statistics are seen as an instructional resource.	

Appendix C
Data Modeling Observable Segment Variables

	Unit 1	Unit 2	Unit 3	Unit 5	Unit 6	Unit 7
Student Contributions						
Did students share invented methods?	X	X	X	X	X	X
Did students make comments or ask questions about the **conceptual** elements of the invented methods?	X	X	X	X	X	X
Did students make comments or ask questions about the procedural or calculational elements of the invented methods?	X	X	X	X	X	X
Teacher Practice						
Did the teacher select student-invented methods to be shared?	X	X	X	X	X	X
Did the teacher compare different ways of thinking?	X	X	X	X	X	X
Did the teacher use questions similar to the ones in the curriculum to support students to think about and discuss different ways of thinking?	X	X	X	X	X	X
Did the teacher make connections between different students' thinking?	X	X	X	X	X	X
Did the teacher connect student thinking to the big ideas?	X	X	X	X	X	X
Did the teacher press students to explain their thinking?	X	X	X	X	X	X
Unit Specific Mathematical Concepts						
Was the order of a display talked about?	X					
Was the scale of a display talked about?	X					
Was the grouping in a display talked about?	X					
Was the effect of design decisions on the shape of a display talked about?	X					
Did the teacher and/or students talk about what one or more displays	X					
Did the teacher and/or students talk about what one or more displays hide about the data?	X					
Was mode (or similar ideas) talked about as a way of measuring center?		X				
Was median (or similar ideas) talked about as a way of measuring center?		X				
Was mean (or similar ideas) talked about as a way of measuring center?		X				
Was a statistic's replicability talked about?		X	X			
Was a statistic's generalizability talked about?		X	X			
Was a statistic's correspondence to a visible distribution talked about?		X	X			
Was a statistic's correspondence to an imagined distribution talked about?		X	X			
Was Range (or similar ideas) talked about as a way of measuring center?			X			
Was Center Clump (or similar ideas) talked about as a way of measuring center?			X			
Was Deviation from Center (or similar ideas) talked about as a way of measuring center?			X			
Was theoretical probability talked about as a measure of chance?				X		
Was empirical probability talked about as a measure of chance?				X		
Was odds as a measure of chance talked about?				X		

Was the relationship between sample size and variability in estimates talked about?	X		
Did students create or talk about a sampling distributions?	X		
Is the center statistic interpreted as an estimate of the probability of the generating device?	X		
Is the variability statistic interpreted exolicitly as an estimate of the sample to sample variability?	X		
Did the teacher help students make sense of the models meaning, and of the relationship between model and output?		X	
Did the class explicitly discuss model fit?		X	
Did the class talk about a distribution of generated data?		X	
Did the class explicitly talk about the sources of variability reflected in the random components?		X	
Did the class explicity discuss the meaning of the non-random component?		X	
Is a center statistic used to create a sampling distribution?		X	
Is a variability statistic used to create a sampling distribution?		X	
Did the teacher help students make sense of the models meaning, and of the relationship between model and output?			X
Did the class talk about the model generated sampling distribution of the median?			X
Did the class talk about the model generated sampling distribution of the IQR?			X
Did the class compare the median of the new data to the sampling distribution of the medians?			X
Did the class compare the IQR of the new data to the sampling distribution of the IQR?			X
Did the class use regions of the sampling distrubution to guide inference?			X
Did the class quantify regions of the sampling distribution to guide inference?			X

Appendix D
Classroom level variables

Variable Question	Units	Options		
Did the class use tinkerplots?	1, 2, 3, 4, 5, 6, 7	Y	N	N/A
Did this class have a right/wrong orientation to mathematics?	1, 2, 3, 4, 5, 6, 7	Y	N	N/A
Did the teacher primarily lecture?	1, 2, 3, 4, 5, 6, 7	Y	N	N/A
How would you characterize this class on the construct map?	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5		
Did the students use repeated measures data?	1, 2, 3	Y	N	N/A
Did the teacher give students the opportunity to invent before the whole class discussion?	1, 2, 3	Y	N	N/A
Was there diversity in student-invented methods?	1, 2, 3	Y	N	N/A
Did the teacher give students an opportunity to try out the statistics before discussing them?	2	Y	N	N/A
Did the class have a shared understanding of a measure of center as the "best guess" of the true length of the attribute measured?	2	Y	N	N/A
Was a "fair share" interpretation of the mean discussed?	2, 4	Y	N	N/A
Was a "balance" interpretation of the mean discussed?	2, 4	Y	N	N/A
Did the teacher provide distributions with visibly different variability for the students to use the statistics to compare?	3	Y	N	N/A
Did the classhave a shared understanding of precision as "the extent to which the measurements agree?"	3	Y	N	N/A
Was a count of a data point in the center clump discussed?	3	Y	N	N/A
Was the percent of data in a center clump discussed?	3	Y	N	N/A
Was a quantification of center clump other than middle 50% discussed?	3	Y	N	N/A
Did the students use statistics of center on repeated measures data?	4	Y	N	N/A
Did the students use statistics of center on production process data?	4	Y	N	N/A
Did the students use statistics of center on natural variation data?	4	Y	N	N/A
How did the class interpret statistics of center when used on repeated measures data?	4	Field		
How did the class interpret statistics of center when used on production process data?	4	Field		
How did the class interpret statistics of center when used on	4	Field		

natural variation data?				
Did the students use statistics of variability on repeated measures data?	4	Y	N	N/A
Did the students use statistics of variability on production process data?	4	Y	N	N/A
Did the students use statistics of center on natural variation data?	4	Y	N	N/A
How did the class interpret statistics of variability when used on repeated measures data?	4	Field		
How did the class interpret statistics of variability when used on production process data?	4	Field		
How did the class interpret statistics of variability when used on natural variation data?	4	Field		
Did students work independently or in small groups to run the simulations?	5	Y	N	N/A
Did the teacher run the simulations on a computer in front of the whole class?	5	Y	N	N/A
Did the class have a shared understanding of probability as the ratio of target outcomes to all outcomes?	5	Y	N	N/A
Did the teacher support students to coordinate theoretical with empirical probability?	5	Y	N	N/A
Did the class create sampling distributions of the proportion of a desired outcome?	5	Y	N	N/A
Did the class discuss the inteligibility of a model?	6	Y	N	N/A
Did the teacher run the models on a computer in front of the whole class?	6	Y	N	N/A
Did the class discuss the difference between the constant and randodm components of a model?	6	Y	N	N/A
Did the class create sampling distributions of a measure of center?	6	Y	N	N/A
Did the class create sampling distributions of a measure of variability?	6	Y	N	N/A
Did the class create a "bad model"?	6	Y	N	N/A
Did students compare a measure of center from a new sample to a sampling distribution of a measure of center?	7	Y	N	N/A
Did students compare a measure of variability from a new sample to a sampling distribution of a measure of variability?	7	Y	N	N/A
Did students quantify a region of the sampling distribution that defines the difference between chance fluxuations and real difference?	7	Y	N	N/A

Appendix E

Each item gets one score for an observation. Items scored at the segment level will use the highest score observed more than once.

Unit General Polytomous Items:

Items About Student Participation:

Item 1: How did students talk about their inventions?

sProcedural + sConceptual*2

- 0 = Students did not talk conceptually or procedurally
- 1 = Students talked only procedurally about their inventions (FID 2)
- 2 = Students talked conceptually about their inventions (FID 3)
- 3 = Students talked both procedurally and conceptually about their inventions (FID 3)

(Codes 2 and 3 are combined)

Items about Teacher Strategies

Item 2: How did the teacher use discussion questions to promote a conversation about invented methods?

(Discussion Questions) + (Discussion Questions*Press)

- 0 = Teacher did not use DQ
- 1 = Teacher used DQ to talk about inventions (FID 1)
- 2 = Teacher used DQ and pressed kids to explain their thinking (FID 2)

Item 3: How did the teacher make use of the invented methods?

Initial select

- 0 = Teacher did not select displays to be discussed
- 1 = Teacher selected displays to be discussed (FID 2)

Item 4: Did the teacher compare different methods for displaying data?

Compare + (Compare*Connect to others)

- 0 = Did not compare
- 1 = Teacher compared different ways to display data (FID 4)
- 2 = Used comparison to connect different student ideas (FID 5)

Unit 1 Polytomous Items:

Item 5: How were student inventions used to support scale?

iScale + (iScale*iShow) + (iScale*iHide*2)

- 0 = Scale not talked about
- 1 = Inventions were used to talk about scale (Fid 3)
- 2 = Inventions were used to support a discussion about what scale shows (Fid 4a)
- 3 = Inventions were used to support a discussion about what scale hides (Fid 5)
- 4 = Inventions were used to support a discussion about what scale shows and hides (Fid 5)

(Codes 3 and 4 combined into same level)

Item 6: How were student inventions used to support Grouping (Class)?

Grouping + (iGrouping*iShow) + (iGrouping*iHide*2)

- 0 = Grouping not talked about
- 1 = Inventions were used to talk about Grouping (FID 3)
- 2 = Inventions were used to support a discussion about what Grouping shows (FID 4a)
- 3 = Inventions were used to support a discussion about what Grouping hides (FID 5)
- 4 = Inventions were used to support a discussion about what Grouping shows and hides (FID 5)

(Codes 3 and 4 combined into same level)

Item 7: How were student inventions used to support Order?

iOrder + (iOrder*iShow) + (iOrder*iHide*2)

- 0 = Order not talked about
- 1 = Inventions were used to talk about Order (FID 3)
- 2 = Inventions were used to support a discussion about what Order shows (FID 4a)
- 3 = Inventions were used to support a discussion about what Order hides (FID 4b)
- 4 = Inventions were used to support a discussion about what Order shows and hides (FID 5)

(Codes 3 and 4 combined into same level)

Items From Summary Variables:

Item 8: What was the context of the data?

Summary variable asking about data context

- 0 = Not Repeated Measure
- 1 = Repeated Measure (FID 1)

Unit 2 Polytomous Items:

Item 9: How were student inventions used to support mode?
Mode + (*Mode*ReplicabilityORGeneralizability) + (*Mode*Link to imagined or visible Dist*2)

- 0 = Mode not talked about (NL)
- 1 = Inventions were used to talk about calculating mode (Fid 3)
- 2 = Inventions were used to support a discussion about the replicability or generalizability of Mode (Fid 4a)
- 3 = Inventions were used to support a discussion about how Mode corresponds to distribution (Fid 5)
- 4 = Inventions were used to support a discussion about how Mode corresponds to distribution in order to interrogate the measure's replicability or generalizability (Fid 5)

(Codes 3 and 4 combined into same level)
Item 10: How were student inventions used to support Median?
Median + (* Median *ReplicabilityORGeneralizability) + (* Median *Link to imagined or visible Dist*2)

- 0 = Median not talked about (NL)
- 1 = Inventions were used to talk about calculating mode (Fid 3)
- 2 = Inventions were used to support a discussion about the replicability or generalizability of Median (Fid 4a)
- 3 = Inventions were used to support a discussion about how Median corresponds to distribution (Fid 5)
- 4 = Inventions were used to support a discussion about how Median corresponds to distribution in order to interrogate the measure's replicability or generalizability (Fid 5)

(Codes 3 and 4 combined into same level)
Item 11: How were student inventions used to support Mean?
Mean + (* Mean *ReplicabilityORGeneralizability) + (* Mean *Link to imagined or visible Dist*2)

- 0 = Mean not talked about (NL)
- 1 = Inventions were used to talk about calculating Mean (Fid 3)
- 2 = Inventions were used to support a discussion about the replicability or generalizability of Mean (Fid 4a)
- 3 = Inventions were used to support a discussion about how Mean corresponds to distribution (Fid 5)
- 4 = Inventions were used to support a discussion about how Mean corresponds to distribution in order to interrogate the measure's replicability or generalizability (Fid 5)

(Codes 3 and 4 combined into same level)
Items From Summary Variables:

Item 12: What was the context of the data?

Summary variable asking about data context

- 0 = Not Repeated Measure
- 1 = Repeated Measure (FID 2)

Unit 3 Polytomous Items:

Item 13: How were student inventions used to support range?
Mode + (*irange*ReplicabilityORGeneralizability) + (*irange*Link to imagined or visible Dist*2)

- 0 = Range not talked about (NL)
- 1 = Inventions were used to talk about calculating Range (Fid 3)
- 2 = Inventions were used to support a discussion about the replicability or generalizability of Range (Fid 4a)
- 3 = Inventions were used to support a discussion about how Range corresponds to distribution (Fid 5)
- 4 = Inventions were used to support a discussion about how Range corresponds to distribution in order to interrogate the measure's replicability or generalizability (Fid 5)

(Codes 3 and 4 combined into same level)
Item 14: How were student inventions used to support center clump measures (i.e. IQR)?
Median + (* icenterclump *ReplicabilityORGeneralizability) + (* icenterclump *Link to imagined or visible Dist*2)

- 0 = Center Clump not talked about (NL)
- 1 = Inventions were used to talk about calculating Center Clump (Fid 3)
- 2 = Inventions were used to support a discussion about the replicability or generalizability of Center Clump (Fid 4a)
- 3 = Inventions were used to support a discussion about how Center Clump corresponds to distribution (Fid 5)
- 4 = Inventions were used to support a discussion about how Center Clump corresponds to distribution in order to interrogate the measure's replicability or generalizability (Fid 5)

(Codes 3 and 4 combined into same level)
Item 15: How were student inventions used to support deviation-based measures?
Mean + (* ideviation *ReplicabilityORGeneralizability) + (* ideviation *Link to imagined or visible Dist*2)

- 0 = Deviation not talked about (NL)
- 1 = Inventions were used to talk about calculating Deviation (Fid 3)
- 2 = Inventions were used to support a discussion about the replicability or generalizability of Deviation (Fid 4a)
- 3 = Inventions were used to support a discussion about how Deviation corresponds to distribution (Fid 5)
- 4 = Inventions were used to support a discussion about how Deviation corresponds to distribution in order to interrogate the measure's replicability or generalizability (Fid 5)

(Codes 3 and 4 combined into same level)
Items From Summary Variables:

Item 16: What was the context of the data?

Summary variable asking about data context

- 0 = Not Repeated Measure
- 1 = Repeated Measure (FID 1)

Unit 5 Polytomous Items:

Item 17: Were Theoretical and Empirical probability discussed?

Theoretical probability*empirical probability

- 0 = Theoretical and Empirical probability not related
- 2 = Theoretical and empirical probability related (FID 3)

Item 18: How did the students use sampling distributions of the outcomes of experiments?

Sampling distributions + (sampling disctributions*center stats) + (sampling distributions * variability stats*2

- 0 = Did not create sampling distributions
- 1 = Sample distributions created, but not statistics used to interpret them (FID 2)
- 2 = Center stats used to interpret sampling distributions (FID 4)
- 3 = Variability stats used to interpret sampling distribution (FID 5)
- 4 = Both center and variability stats used to interpret sampling distribution (FID 5)

(Codes 3 and 4 combined)

Item 19: Did the class discuss the relationship between sample size and the sampling distributions?

Sample Size + (Sample Size*center stats) + (sample size*variability stats*2)

- 0 = Did not talk about sample size
- 1 = talked about sample size
- 2 = talked about relationship between sample size and center stats
- 3 = talked about relationship between sample size and variability stats
- 4 = talked about relationship between sample size and both variability and center stats

(levels 3 and 4 combined)

Items From Summary Variables:

Item 20: Did the class have a shared understanding of probability as the ratio of target outcomes to all outcomes?

Item 21: Did the teacher support students to coordinate theoretical with empirical probability?