Chapter 1

Measurement in Early and Elementary Education

Douglas H. Clements
University of Denver
Jeffrey E. Barrett
Illinois State University
Julie Sarama
University of Denver

THE IMPORTANCE OF MEASUREMENT

Quantitative reasoning and measurement competencies support the development of mathematical and scientific thinking from prekindergarten (pre-K) through Grade 8 (Clements, 2003; Davydov, 1991; So, 2013; Steffe, 1991; van den Heuvel-Panhuizen & Buys, 2008) and are fundamental to science, technology, engineering, and mathematics (STEM) education. Children's knowledge of quantity and strategies for quantifying, as well as their reasoning about quantity, are necessary elements of a working theoretical foundation for measurement (Case & Okamoto, 1996; Piaget, Inhelder, & Szeminska, 1960; J. P. Smith & Thompson, 2007; Steffe & Cobb, 1988; Steffe & Olive, 2010; Tal, 2013). In particular, geometric measurement (e.g., length, area, and volume) bridges the mathematical domains of number and geometry (Sarama & Clements, 2009b). Children forge this critical connection while they establish and reason about quantity throughout the mathematics curriculum.

Measuring is a nontrivial aspect of children's developing mathematical and scientific thinking. Its basis is in the exercise of comparative judgment; children compare magnitudes immediately or observe patterns of change over time by quantifying their experiences and observations. Measurement knowledge and strategies play broadly and deeply into children's understanding of both science and mathematics, making measurement a vital component of pre-K through Grade 8 curricula (So, 2013). For example, measurement experience provides images of variables as quantities in algebra, an important focus of the upper elementary and middle school curriculum, which does not often emphasize a conceptual approach to thinking quantitatively (A. G. Thompson & Thompson, 1996; P. W. Thompson & Thompson, 1994).

The *National Science Education Standards* (National Research Council, 1996) indicated that mathematical measurement plays an essential role in all aspects of scientific inquiry. C. L. Smith, Wiser, Anderson, and Krajcik (2006) pointed out

the integrative role of measurement for learning science conceptually: "Given the centrality of measurement in science and the ways measurement can contribute to conceptual understandings, it is important to start early in developing a rich understanding of the measurement of important physical quantities" (p. 33). Measurement is a central aspect of spatial thinking; a quantitative understanding of space is often essential for posing questions, developing explanations, and communicating results (National Research Council, 2006; Presmeg & Barrett, 2003). Within technological design for spatial problems, children need to choose suitable tools and techniques and work with appropriate measurement methods to ensure precision and accuracy. The authors of the *Atlas of Science Literacy* (American Association for the Advancement of Science, 2001) described measurement concepts as a critical foundation for establishing evidence as well as a basis for modeling processes and systems (cf. Lehrer & Schauble, 2002, 2004; Petrosino, Lehrer, & Schauble, 2003).

In 2006, the National Council of Teachers of Mathematics (NCTM) released *Curriculum Focal Points for Prekindergarten Through Grade 8 Mathematics: A Quest for Coherence* (CFP). With this document, NCTM recommended that teachers focus primarily on target curriculum topics for each grade level from pre-K through Grade 8. The CFP offers guidelines for age-appropriate learning activities, clustered about central themes in mathematics. Furthermore, it portrays measurement as an essential topic, underscoring the important role that measurement instruction plays, especially before middle school. Although the National Governors Association Center for Best Practices (NGA) and the Council of Chief State School Officers (CCSSO) completed the Common Core State Standards for Mathematics (CCSSM; NGA & CCSSO, 2010) after we conducted the research reported in this monograph, the CCSSM also includes geometric measurement as an essential topic. In response to the fundamental need for developmental accounts based in research, we set out to furnish longitudinal accounts that would establish sequenced assessment goals across grades, guide curricular development programs, and strengthen teacher education.

THEORETICAL PERSPECTIVE

We approached this project through a theoretical lens called *hierarchic interactionalism* (Sarama & Clements, 2009b), a synthesis of previous theoretical frameworks that owes much to theorists and researchers too numerous to list here (see Sarama & Clements, 2009b) but that relies heavily on several (e.g., Carpenter, Franke, Jacobs, Fennema, & Empson, 1998; Confrey & Kazak, 2006; Karmiloff-Smith, 1992; Minsky, 1986; Piaget, 1941/1952; Siegler & Booth, 2004; Steffe & Cobb, 1988; Van de Rijt & Van Luit, 1999; Vygotsky, 1934/1986). This research corpus is the basis for the 12 tenets of hierarchic interactionalism (adapted from Sarama & Clements, 2009b):

1. *Developmental progression.* Children acquire content knowledge along developmental progressions of levels of thinking that are topic-specific (see Tenet 2). These progressions are consistent with children's intuitive background knowledge and patterns of thinking and learning, which may be culturally specific, albeit still influenced by "initial bootstraps" (see Tenet 6). Specific concepts and processes characterize each level of thinking.
2. *Domain-specific progression.* Developmental progressions are characterized within a specific mathematical domain or topic, such as measurement, and even measurement of a particular attribute.
3. *Hierarchic development.* Levels of thinking are internally coherent, but the learning process is often incremental and gradually integrative, rather than "stagelike."
4. *Cyclic concretization.* Developmental progressions proceed from sensory-concrete levels, in which perceptual concrete supports are necessary and reasoning is usually restricted to a few cases, to more explicit verbally based generalizations, resulting in integrated-concrete understandings relying on mental representations that serve as models for operations and abstractions (see Sarama & Clements, 2009a).
5. *Mutual development of concepts and skills.* Concepts and skills develop together, each supporting the further development of the other.
6. *Initial bootstraps.* Children are endowed with premathematical competencies and predispositions either at birth or soon thereafter; these competencies and predispositions support and constrain future learning.
7. *Different developmental courses.* Different developmental courses are possible; individual, environmental, and social confluences influence them.
8. *Progressive hierarchization.* Children integrate concepts and skills, building understandings that are hierarchical and that include both generalizations and differentiations.
9. *Environment and culture.* Environment and culture affect the pace and direction of development.
10. *Consistency of progressions and instruction.* Basing instruction on learning research adds value.
11. *Learning trajectories.* Curricula and instruction developed on the basis of full hypothetical learning trajectories are more effective than those that lack such a basis.
12. *Hypothetical learning trajectories.* Teachers interpret hypothetical learning trajectories and realize them through social interaction with children around instructional tasks.

Learning Trajectories

We based our longitudinal study of children's thinking and learning on the construct of a *learning trajectory* (LT). In his seminal work, Simon (1995) stated that a *hypothetical learning trajectory* (HLT) included "the learning goal, the learning activities, and the thinking and learning in which students might engage" (p. 133). We use the term learning trajectory rather than hypothetical learning trajectory for the sake of simplicity in our reference to LTs. Nevertheless, we note that the LTs are theoretical objects of research; as such, they are consistent with HLTs. We view LTs as

> descriptions of children's thinking and learning in a specific mathematical domain and a related, conjectured route through a set of instructional tasks designed to engender those mental processes or actions hypothesized to move children through a developmental progression of levels of thinking, created with the intent of supporting children's achievement of specific goals in that mathematical domain. (Clements & Sarama, 2004, p. 83)

Developmental progressions are the core component of LTs (and the main focus of this monograph), but they have two additional parts: goals and instructional activities. The synthesis of these three parts can be expressed as follows: to reach a certain mathematical competence in a given topic (the goal), children progress through sequential levels of thinking (the developmental progression), supported and assisted by tasks (the instructional activities; see Clements & Sarama, 2004).

Mathematical goal. We determined our goals through a synthesis of three sources. First were the big ideas and core competencies of the domain—clusters of concepts and skills that are mathematically central and coherent, consistent with children's thinking, and generative of future learning. These big ideas and competencies come from several large projects, including those from the NCTM's (2006) CFP, the National Mathematics Advisory Panel's (2008) final report, the CCSSM (NGA & CCSSO, 2010), and related projects (Clements, Sarama, & DiBiase, 2004). For example, at least eight concepts form the foundation of children's understanding of length measurement: understanding of the attribute, conservation (a good example of a concept often assumed by content descriptions, but important psychologically), transitivity, equal partitioning, iteration of a standard unit, accumulation of distance, origin, and relation to number (Clements & Stephan, 2004; Sarama & Clements, 2009b). Our goal for children certainly includes that they learn to "measure and estimate lengths in standard units" (NGA & CCSSO, 2010, p. 20), including selecting appropriate tools, relating measurements by using different units, estimating, and comparing measurements. In addition, we include children's explicit understanding of how such competencies relate to the concepts of equal partitioning, iteration, and origin, as well as others. As an example in measurement of two or three dimensions, the cognitive competence of spatial structuring is fundamental but ignored in most content-only descriptions of mathematical goals.

Developmental progression. A main assumption is that children acquire most content knowledge, such as knowledge of geometric measurement, along developmental progressions of levels of thinking that are particularly consistent with children's intuitive knowledge and patterns of thinking and learning (at least in a particular culture, but guided in all cultures by innate competencies). That is, children progress through domain-specific levels of understanding in ways we can characterize by the concepts and processes (mental objects and actions on them) that build hierarchically on previous levels. These actions on objects are children's main way of operating on, knowing, and learning about the world, including the world of mathematics. As a brief example, our previous work on length LTs revealed that children reliably developed through the levels of thinking, passing through a level in which they measured length by placing multiple units, sometimes leaving gaps between units, before they could iterate a unit and then measure length competently with multiple strategies and tools (Barrett et al., 2011; Sarama, Clements, Barrett, Van Dine, & McDonel, 2011).

Developmental progressions follow a pattern. Initially, at the sensory-concrete level, children need perceptual concrete supports. Next, children create verbally encoded generalizations that have distance from those supports. Finally, children construct integrated-concrete understandings on the basis of mental representations that serve as models for mathematical concepts and operations (Clements & Sarama, 2009; Sarama & Clements, 2009a).

Levels of thinking are coherent, with increased sophistication, complexity, abstraction, and generality often characterizing higher levels (Clements & Sarama, 2014b; Confrey, Rupp, Maloney, & Nguyen, 2012; Maloney & Confrey, 2010; Sztajn, Confrey, Wilson, & Edgington, 2012). Conceptualizing levels at different grain sizes is possible. As an example, the "growth points" of the Early Numeracy Research Project in Australia are developmentally widely spaced (e.g., children might achieve successive points in different grades; Clarke et al., 2002). Our approach (Sarama & Clements, 2009b) is to find levels of finer grain size so that, for example, teachers can see and observe development through several levels within their grade but still retain the defining characteristics of levels, such as constancy (some properties, states, or activities remain consistent at the level), order invariance, and hierarchic incorporation and integration (Sarama & Clements, 2009b; Steffe & Cobb, 1988).

Instructional tasks. Although the emphasis of this research project was the developmental progressions, instructional tasks, the third component of LTs, also played a role. In hierarchic interactionalism, educators often design tasks to present a problem that is just beyond the children's present level of operating. In some cases, sequences of tasks may furnish sufficient goal-directed activity, changing the focus of attention (e.g., from a physically concrete model to a mental image or symbolic representation), reflection (identification and connection of commonalities), and anticipation (prediction of the result of an activity before carrying it out) to engender children's generalization to form abstractions, such as creating a new

object (e.g., a unit of units) that is more general or abstract mathematically (Simon, 2013). In other cases, children may need to actively engage in reformulating the problem or their solution strategies, often with peer interaction and teacher guidance. In reflecting on their activity, children learn whether they have solved the original problem or whether they need to engage in more thinking. This cycle may continue until they build a new level of thinking. A critical mass of such construction in individuals and small groups often allows productive class discussions that can justify, formalize, and symbolize the mathematics (Simon, 2013) and eventually integrate it into a mathematical system (van Hiele, 1986; van Hiele-Geldof, 1984) in a manner appropriate for the children.

Our theory of hierarchic interactionalism does not attempt to include or categorize a full range of teaching and learning processes (e.g., see Clements, Agodini, & Harris, 2013; Clements & Sarama, 2009; Tharp & Gallimore, 1988; van Hiele-Geldof, 1984) but posits that instructional tasks are a main component of effective instruction and that any pedagogical practice is effective to the extent that it activates children's mental actions on objects that support the subsequent level of thinking in the development progression.

Misconceptions Regarding Learning Trajectories

Because of the complexity of LTs, it is no surprise that educators often misunderstand them. One claim is that they are nothing new (e.g., are just "scope and sequences"). The LTs of hierarchic interactionalism share characteristics with previous research programs but differ in significant ways (for a complete analysis, see Clements & Sarama, 2014b). LTs build on previous work, which has developed from simple accumulations of connections (Thorndike, 1922) to more complex views of thinking and learning (Gagné, 1965/1970; Piaget et al., 1960; Resnick & Ford, 1981). All theories attempt to explain psychological sequences, with previous accounts relying on the acquisition of facts and skills over time. LTs include such sequences but are not limited to sequences of "pieces" of knowledge only. They contain descriptions of children's levels of thinking; and it is therefore impossible to summarize them by stating a mathematical definition, concept, or rule (cf. Gagné, 1965/1970). Levels of thinking convey *how* children think about a topic and *why.* Finally, earlier theories often based instruction on a transmission approach. LTs have an interactionalist view of teaching—children interacting with a teacher and other children around instructional tasks, with the teacher using all three components of LTs to guide those interactions.

A second misconception is that LT levels rigidly categorize children (i.e., "put children in boxes"). In contrast, the theory posits that children can operate at multiple levels, with most children working mostly at one level or in transition between two levels. We identify children as being at a certain level when most of their behaviors reflect the patterns of thinking at that level. Often, they show a few behaviors from the next (and previous) levels while they learn. The continued

existence of earlier levels, as well as the role of intentionality and social influences in their instantiation, explains why in some contexts even adults fall back to earlier levels (e.g., failing to conserve in certain situations).

Related to the second misconception is a third: the belief that LTs are insensitive to individual and cultural differences. Hierarchic interactionalism posits that different developmental courses are possible within constraints, depending on individual, environmental, and social–cultural confluences (Clements & Battista, 2001; Confrey & Kazak, 2006; Sarama & Clements, 2009b). Within any developmental course, children at each level of development have a variety of cognitive tools—concepts, strategies, skills, utilization, and situation knowledge—that coexist. The differences within and across individuals create variation that is the wellspring of invention and development. At a group level, however, these variations are not wide enough to vitiate the theoretical or practical usefulness of the tenet of developmental progressions; for example, in a class of 30, one might find only a handful of different solution strategies (cf. Murata & Fuson, 2006), most of which represent adjacent levels along the developmental progression. Further, environment and culture affect the pace and direction of the developmental courses. For example, the degree of experience that children have to observe and use number and other mathematical notions and to compare these uses affects the rate and depth of their learning along the developmental progressions. The degree to which children learn mathematical words, exposure to which varies greatly across cultural groups (Buss & Spencer, 2014), affects developmental courses. Words alert children to the class of related words that they must learn and to specific mathematical properties, laying the foundation for learning mathematical concepts and language (cf. Sandhofer & Smith, 1999) by providing a nexus on which to build their nascent constructs (Vygotsky, 1934/1986).

On one hand, no single absolute developmental progression, and thus LT, exists for a mathematical domain because other factors—such as environment, culture, and educational experiences—may affect a child's actual pathway for learning and development. On the other hand, a large number of pathways do not exist, because universal developmental factors interact with culture and mathematical content. Educational innovations and culturally specific educational environments may establish new and potentially more advantageous sequences, so work with LTs should recognize the culturally situated nature of children's learning and the need for a "critical view" (Wager & Carpenter, 2012). We concur that educators should examine generalized claims about children by including adequate samples of a broadly inclusive collection of children across cultural and economic constructs. Yet, this respect for varying populations of children need not imply that we as an educational community must expect that substantially different LTs exist for some groups of children. First, people often translate this expectation into "my children learn more slowly," an avoidable trap of low expectations disguised as sensitivity to cultural and individual differences. Second, we believe that the cognitive core of the LTs is valid within different cultural contexts and should be instantiated in

different ways in different sociocultural settings to promote learning and equity. Of course, researchers should test this belief in future research that uses such a cognitive core account in a range of settings.

In summary, we used LTs in our research because they address the components of mathematical goals, developmental progressions, and instructional tasks; and they are increasingly important in guiding the writing of standards (Maloney, Confrey, & Nguyen, 2014; NGA & CCSSO, 2010) and curricula (Clements, 2007; Maloney et al., 2014). Further, teachers who understand LTs understand the mathematics associated with the various levels of reasoning, the way that children think and learn about mathematics, and how to help children learn it better. LTs connect research and practice. They help teachers understand the level of knowledge and thinking of their classes and the individuals in their classes as important in serving the needs of all children. In this way, LTs serve as a basis for formative assessment, an essential component of high-quality teaching (see National Mathematics Advisory Panel, 2008). More refined and better validated LTs can contribute to improved standards, curricula, and teaching practices.

THE NATIONAL SCIENCE FOUNDATION'S CHILDREN'S MEASUREMENT PROJECT

The Children's Measurement research team built, refined, and tested a set of LTs for geometric measurement (Barrett et al., 2011; Barrett et al., 2012; Sarama, Clements, Barrett, et al., 2011; Sarama, Clements, Van Dine, et al., 2011) to represent children's developing measurement knowledge and their ways of learning measurement. To complement our emphasis on children's learning, we looked for ways of improving instruction, assessment, and curriculum development related to measurement concepts. A main goal of our team has been to address the educational need for measurement in the pre-K to Grade 5 mathematics curriculum by clarifying, extending, and improving LTs. We see LTs as tools with wide-ranging application in education. Because of our interest in addressing a range of educational issues, we work to improve these tools through design cycles of research. For example, after establishing a mathematical goal (e.g., NCTM, 2006; NGA & CCSSO, 2010), we began by reviewing the research and attempting to piece together a coherent longitudinal story, even from separate disparate studies (Sarama & Clements, 2009b). We assessed children at the beginning and end of the study by using clinical interviews with younger children and a written format with older children to clarify the cognitive attributes of the nascent levels. We conducted teaching experiments with individual children, as well as classroom-based teaching experiments with intensive qualitative analyses to revise and expatiate those levels and their interconnections and sequences.

The research in this monograph began with the LTs that we had created by using these procedures (Clements & Sarama, 2014a; Sarama & Clements, 2009b), recognizing that we needed to refine and extend them. We believe that the resulting LTs can strengthen the educational and curricular infrastructure needed for

implementation of the measurement standards within the CCSSM (NCTM, 2006; NGA & CCSSO, 2010) from kindergarten through Grade 8.

Our Research Goals

One of our central goals has been to delineate the development of children's knowledge and potential strategic competence on increasingly demanding sets of length, area, and volume measurement tasks through successive grade levels from pre-K to Grade 5. Another central goal has been to establish prototypical longitudinal stories of children's ways of reasoning through measurement activities. Such stories, which we based on increasingly complex tasks, influenced the design and implementation of instructional sequences, guiding teachers toward essential concepts relevant to children in their particular grade levels, in keeping with the CFP (NCTM, 2006). We therefore conducted longitudinal studies by following the developmental reasoning of selected children at two research sites over a 4-year span, from pre-K to Grade 2 at one research site and from Grade 2 to Grade 5 at the other (see Participants section for more detail).

We expect these LTs for measurement to help improve formative classroom assessment by teachers because they can more easily recognize opportunities to engage children proactively in conceptual change (Sztajn et al., 2012; Wickstrom, Baek, Barrett, Cullen, & Tobias, 2012; Wilson, 2009). We also expect the LTs to support formal assessment efforts and research on curriculum development (Clements, 2007).

Research Question

Our research focuses on supporting developmental coherence within the curriculum for elementary mathematics and within the instructional approaches that educators recommend for teaching. We are therefore interested in describing the growth of children as persons who reason about their world by using measurement tools to solve problems and organize quantitative thinking (K. F. Miller, 1989; P. W. Thompson, 1992). In particular, we examine central concepts in measurement learning, including the following: (a) identifying attributes of continuous dimensions to quantify and compare, (b) building appropriate units with conservation, (c) partitioning objects to generate quantity, (d) composing and coordinating units and groups of units in structures cumulatively, (e) iterating units in correspondence with number schemes including an assignment for zero, and (f) marshaling logical support and justification of measurement claims to guide scientific reasoning (cf. Clements & Stephan, 2004; Lehrer, 2003; Stephan & Clements, 2003). The main research question that motivated the Children's Measurement Project addresses these concepts across three spatial measurement domains (length, area, and volume): How do children's thinking and learning about each of the three domains of spatial measurement develop over time from pre-K through Grade 5?

Participants

We drew the children in our study from an urban parochial school in the Northeast (78% White, 9% Black, 9% Hispanic, 1% Asian, 3% two or more races) and from a suburban public school in the Midwest (68% White, 9% Black, 9% Hispanic, 9% Asian, 6% two or more races). Approximately 20 children were in each class at these schools, and all the children in pre-K through Grade 5 at both schools participated in a quantitative assessment conducted during Year 1 of the study. By using this assessment and collaborating with teachers, we selected a cohort of eight pre-K children at the Northeast site and eight Grade 2 children at the Midwest site. They became the focus children for our 4-year longitudinal study. Children at the Northeast site participated in the study during their 4 years of schooling from pre-K through Grade 2. Children at the Midwest site participated in the study during their 4 years of schooling from Grade 2 through Grade 5.

On the basis of our initial quantitative assessment and teacher input, we selected two low-performing children, four middle-performing children, and two high-performing children as a representative subset at both school sites. In addition, we strove for variation in socioeconomic status (SES) and cultural background that reflected the demographics of the schools. The selected sample also included equal numbers of male and female children. Finally, we used the same methods to select another eight children at each site to serve as "background" children. We used these background children to pilot tasks, gather additional information, and act as replacements for children who left the focus group. Attrition at the Northeast site resulted in three children leaving the study after the first year, one child after the second year, and two children after the third year. At the Midwest site, one child left the study after the first year, one child after the third year, and two children during the fourth year. Each time a child left the study, a child from the background group who had comparable characteristics replaced the child who left so that we could maintain a balanced focus group (in gender, cultural background, mathematics performance, and SES) at that site.

Project Design

The next two sections describe our qualitative and quantitative methods for data collection and analysis. Although our primary goal was to provide longitudinal accounts of children's developing understanding of measurement, we complemented these accounts by designing and checking assessment items at each level of the length, area, and volume LTs. We administered these assessment items to all the children at the Midwest site and selected classrooms of children at the Northeast site.

To produce longitudinal accounts of children's developing knowledge, we used teaching experiments to create descriptions of plausible sequences of growth along the LTs for length, area, and volume. These teaching experiments functioned in a supportive role alongside regular classroom instruction based on existing

curriculum materials in the schools. We recognize that a number of instructional factors may have led to the documented growth through extended sequences of teaching episodes throughout the study. Our primary goal, however, was to validate and refine the developmental progressions within the LTs instead of focusing on the instructional tasks.

Qualitative Methods

At the Northeast site, three children remained in the study for the 4-year duration, and they were the main focus of our qualitative analysis:

- Edith was a high-performing female with a desire to provide the correct answer always. She sought approval and positive feedback from the researchers, so that the researchers had to exercise vigilance to avoid guiding her through their vocalizations and behavior. Edith often commented that her mother regularly did mathematics problems with her at home.
- Ryan was a middle-performing male who was often quick to answer and usually took the lead in any group setting. He had a strong number sense and a willingness to take risks, which often resulted in repeated attempts to work through problems until they were complete.
- Lia was a low-performing female with a strong sense of spatial structuring: She was the first to correctly construct a three-dimensional cube object from a two-dimensional representation. However, she struggled to estimate and often lost track during long multistep problems. Lia remained in kindergarten for 2 years because of weak recognition of letters.

To furnish additional details, we turned to two children for whom we had between 2 and 3 years of longitudinal data: Zola, a middle-performing female and Marina, a low-performing female. Although the school retained Lia in kindergarten, the other four children progressed yearly through the grades.

At the Midwest site, five children remained in the study for the 4-year duration, and we focused on them for our qualitative analysis. We selected the children from two classes of children at one school.

- Arielle was a high-performing female with a tendency to use arithmetic operations to operate on problems posed in geometric space rather than operating with spatial operations and images directly to solve problems. She also had a tendency to work quietly until reporting her response to tasks.
- Abby was a middle-performing female with a tendency to explore task situations with a variety of strategies, and she often appealed to spatial actions or objects to support her reasoning about tasks. She was easy to engage and was willing to explain her thinking while she was working.

- Owen was a middle- to high-performing male who often worked quietly but with keen energy to examine problems in measurement. He balanced his reasoning between spatial and arithmetic operations.
- Drew was a middle-performing male with an articulate sense of mathematics, a capacity to imagine space and operations in space, and a willingness to use novel approaches to problems of measurement. However, sometimes his novel approaches took him off task and he required extra work to regain the task topic and goal.
- Anselm was a middle-performing male oriented primarily toward reasoning about spatial objects, with a secondary emphasis on arithmetic operations with measured values. We noticed that Anselm often focused on the instructions and questions of his peers or teacher to such an extent that he would prefer to follow another person's line of questions or thinking and develop alternative approaches to problems rather than proceed with his own initial approach.

To provide additional analysis at various stages of our study at the Midwest site, we included further data from two children for whom we had between 2 and 3 years of longitudinal data: Danny, a low-performing male, and Randy, a middle-performing male. Finally, we include occasional analysis from sessions with one female child, Sara, for whom we had little more than 1.5 years of data. Sara was a low-performing female who was unable to complete the second year of this research program when her family moved away from the school.

Our qualitative data sources included assessments, as well as individual and classroom-based teaching experiments. We administered assessments to children during 2008 and 2011. These two sets of assessments complemented the teaching experiments. The assessments included modifications of initial tasks and novel tasks designed to elicit specific thinking and behaviors indicative of particular levels of the developmental progression of the LTs (see www.childrensmeasurement.org).

To further evaluate each LT beyond the assessments, we conducted two types of teaching experiments (Cobb & Gravemeijer, 2008; Steffe & Thompson, 2000). First, we conducted individual teaching experiments, which included teaching episodes that occurred approximately every 4 weeks with individual focus children. Second, classroom-based teaching experiments occurred approximately once a year; during these experiments, we took the lead in the measurement instruction when it occurred in the classroom curriculum (see Figure 1.1). Previous observations and interpretations influenced each subsequent teaching episode. We designed the classroom-based teaching experiments with the classroom teachers. We also observed the classroom teachers' lessons on measurement. Data sources included video records and field notes. This combination of individual and classroom settings addressed our desire to study children's behaviors in naturalistic contexts. We analyzed these behaviors and compared them with the LTs to identify

the level at which a child was demonstrating knowledge or to highlight gaps or inconsistencies in the LT description, and we tentatively revised LTs when a preponderance of evidence favored revision. The most demanding checks were across investigators within and especially between sites. We had to confirm any consistent behavior or revision across sites and incorporate or explain any disconfirming evidence before we could consider that the theoretical assertion was warranted.

	Northeast Site	**Midwest Site**
Year 1 2009–2010 school year	• Initial assessment • 4-day CTE focusing on length • 4 teaching episodes: 4 involving length and 1 also involving area	• Initial assessment • 4 teaching episodes: 3 involving length and 1 involving area
Year 2 2010–2011 school year	• 4-day CTE focusing on area and perimeter • 13 teaching episodes: 11 involving length, 4 involving area, and 7 involving volume	• 4 CTEs: 3 focusing on length and 1 focusing on area • 9 teaching episodes: 4 involving length, 2 involving area, and 3 involving volume
Year 3 2011–2012 school year	• 5-day CTE focusing on length • 14 teaching episodes: 8 involving length, 5 involving area, and 6 involving volume	• 5 CTEs focusing on area • 12 teaching episodes: 8 involving length, 7 involving area, and 6 involving volume • Grade 4 assessment
Year 4 2012–2013 school year	• 9 teaching episodes: 2 involving length, 6 involving area, and 4 involving volume • 8 Measurement Club meetings: 5 involving length, 5 involving area, and 6 involving volume • Final assessment	• 2 CTEs involving length, area, and volume • 9 teaching episodes: 8 involving length, 6 involving area, and 3 involving volume • Final assessment

Figure 1.1. Data collection cycles and data sources across the 4-year longitudinal study. (A CTE is a classroom-based teaching experiment.)

Quantitative Methods

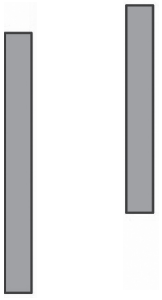
To provide triangulation on the qualitative findings and to ascertain whether we could generalize those findings to other children, we created an assessment instrument and administered it to children ($n = 258$) in pre-K through Grade 5 at the two participating schools during spring 2011. These children were a representative

sample of students at their respective schools. The total sample included 25 pre-K, 32 kindergarten, 29 Grade 1, 42 Grade 2, 41 Grade 3, 44 Grade 4, and 45 Grade 5 children. This sample included the cohorts of children that we followed in our teaching experiments.

We developed the instrument analyzed here to align with the levels of our measurement LTs, specifically those for length, area, and volume. Tasks included the measurement items from a previously developed and validated assessment, as well as tasks from previous empirical studies (Clements & Sarama, 2007). We used two items to assess each level within each of these LTs (for a total of 52 items; see Figures 1.2, 1.3, and 1.4 for sample length, area, and volume items, respectively), with the understanding that a correct response on an item demonstrated that a child was at least at that level. In addition, for the lower levels of each LT, we designed items for presentation through interview to limit the confounding effect of reading ability for younger children. We assessed children in pre-K, kindergarten, and Grade 1 entirely through interview; we assessed children in Grades 2 and 3 through a combination of interview and written items; and we assessed children in Grades 4 and 5 entirely through written items. We videotaped all interviews and conducted them one-on-one with an assessor and child. We presented all written items to children in their classrooms and collected all work and written responses.

Text continues on page 20

Length Direct Comparer



Place the two strips of paper (one longer and one shorter) on the desk in front of the child. Place them about 10 to 12 inches apart in the orientation shown in the illustration. Allow the child to move them.

Ask: *Which strip of paper is longer (or taller)? How do you know?*

End-to-End Length Measurer

Put out a 12-inch strip and a pad of 1.5-by-2-inch sticky notes. Place the pad of sticky notes on the tabletop at an angle of about 45 degrees with respect to the table edges.

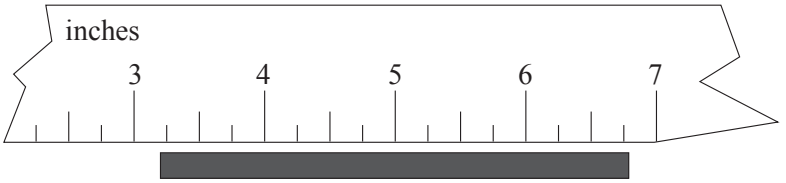
Ask: *How many sticky notes will fit along this strip?*

Figure 1.2 continues on next page

Length Unit Relater and Repeater

Say: *I measured this long strip* (hold up 12-inch green strip) *with this* (hold up 1-inch yellow strip) *and found that it was 12 yellow strips long. If I measure the long strip with this blue strip* (hold up 2-inch blue strip), *how many of these blue strips will I need?*

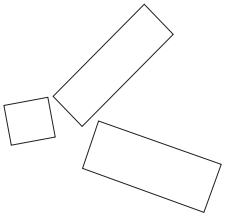
Consistent Length Measurer



Say: *This is a picture of a rod just below a broken section of a ruler. Use this picture to measure the length of the rod. How long is the rod?*

Figure 1.2. Sample length items from the assessments.

Area Simple Comparer



Scatter the rectangle cutouts in front of the child. (Make sure that they do not line up.) Allow the child to move the rectangles.

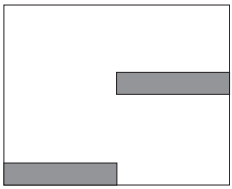
Ask: *Which piece of paper will let you paint the biggest picture?* (If the child wants to fold or cut, ask, *Can you do it without folding or cutting?)*

Side-to-Side Area Measurer

Say: *I wanted to cover this rectangle* (trace around the boundary of the larger rectangle) *with these squares* (point to one of the square-inch units). *I started drawing them in. Please finish the drawing by completely covering the rectangle.*

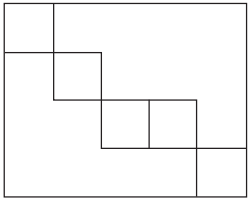
Figure 1.3 continues on next page

Area Unit Relater and Repeater



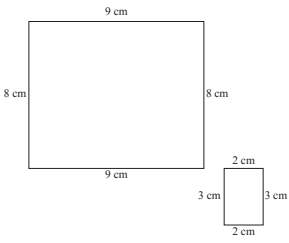
Ask: *How many tiles like this one*  *would cover the larger rectangle? Please include the two tiles already drawn.*

Partial Row Structurer



Ask: *How many squares would completely cover this rectangle? Please include the five squares already drawn.*

Area Row and Column Structurer)



Ask*: How many of the small rectangles would cover the large rectangle?*

Figure 1.3. Sample area items from the assessments.

Capacity Direct Comparer

Show the child the two containers as shown below:

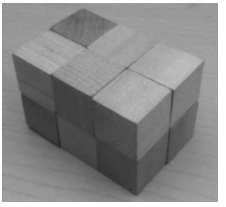
Say: *Pay attention because I am going to ask you a question about these two containers in a minute.*

(Point to the two containers. Completely fill one of the containers with water. Pour the water from container into the other.)

Ask: *Which of these two containers can hold more water?*

(Point to the two containers again.)

Primitive 3-D Array Counter

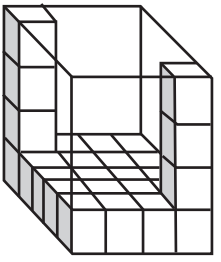


Place a 2 × 3 × 2 solid ("glued together") on the table. Show a separate inch cube.

Say: *This is a cube* (show child individual inch cube).

Ask: *How many cubes like this do you think that you would need to make this?* (gesture to solid)

Partial 3-D Structurer



Ask: *How many cubes altogether will it take to fill the box?*

Figure 1.4 continues on next page

3-D Row and Column Structurer

Tell students: *This net* (on the left) *folds up to make the box* (in the middle).

Ask: *How many cubes* (on the right) *would you need to fill the box?*

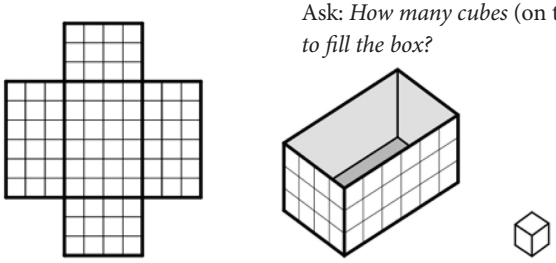


Figure 1.4. Sample volume items from the assessments.

We pooled these data across both the Northeast and Midwest sites and coded each item for correctness according to the following rubric: 0 = incorrect, 1 = correct with prompt (all prompts given according to the assessment protocol), 2 = fully correct. We considered any item coded with 1 as partially correct in the Rasch modeling. During the coding process, we included an additional code of 999 to identify any item for which we could not attribute a child's response directly to the child's understanding (e.g., assessor error affected the response of the child, child's response was unintelligible or unreadable, or child did not provide a response for the item). We treated any item coded as 999 as missing data in the analysis.

In our analysis of these data, we used item response theory (IRT), which allows researchers to create an interval scale of scores for both the difficulty of items and the ability of the persons assessed. The Rasch model is the simplest and most efficient IRT model. Use of the Rasch model furnishes evidence of both the validity and the reliability of an assessment. In addition, it allows mathematical estimation of both the probability that a person will answer an item correctly (person ability), as well as the probability that an item will be answered correctly by a person (item difficulty). Rasch modeling yields an ability score on an interval scale with a consistent, justifiable metric, thereby allowing accurate comparisons, even across ages, as well as meaningful comparison of change scores (Wright & Stone, 1979). Rasch modeling reports all ability scores in units called *logits*. Just as two inches is twice as long as one inch, two logits represent twice as much ability as one logit. This scoring differs from other scores such as percentile ranks, for example, because we cannot say that a person at the 50th percentile has twice as much ability as a person at the 25th percentile.

One underlying assumption of Rasch modeling is that we are measuring a unidimensional construct or latent trait. For our analysis, we defined *measurement competence* as that latent trait (Bond & Fox, 2007; Linacre, 2014; Watson, Callingham, & Kelly, 2007). To measure measurement competence, we sequenced the items, strictly maintaining the order within each measurement domain (length, area, volume) but intermingling items across domains according to the available developmental evidence, including age specifications from the literature and difficulty indices from our pilot testing. We therefore posited that items were organized according to increasing order of difficulty across domains, but our theoretical claims that this sequencing represented increasingly sophisticated levels of mathematical thinking were only for items within a given domain. We submitted the results of administering this revised instrument to the Rasch model.

Rasch analyses estimate the distance between items and between persons on a single scale. That is, item difficulty and each person's underlying competence are on the same equal-interval scale, indicating the theoretical latent trait. We used Winsteps (Linacre, 2014) to estimate fit statistics, reliabilities, separation indices, and item difficulties. Fit statistics (infit and outfit) are estimates of the degree to which responses show adherence to the expectations of the Rasch model. They indicate how well the model empirically supports the assumption of unidimensionality; that is, whether it is measuring a single attribute (a critical characteristic of fundamental measurement). In addition, the mean square (MNSQ) statistic is a transformation of the difference between the predicted and the observed scores (residuals) that indicates the degree of fit of an item or a person. Its expected value is 1, with values between 0.5 and 1.50 regarded as productive for measurement (Wright & Linacre, 1994). The Z-statistic (ZSTD) is a standardized fit statistic with a mean of 0 and variance of 1. For ZSTD, the range of acceptable values for a 95% confidence interval is between -2 and 2 (Bond & Fox, 2007; Linacre, 1994).

Item reliability is an estimate of the replicability of item placement within the hierarchy of items along the measured trait and is similar to Cronbach's alpha (Bond & Fox, 2007; Linacre, 1994). Item separation indices are assessments of the ability of the measure to differentiate items along the scale. Because we had employed qualitative analyses to refine the instrument in several previous cycles of formative testing and revision, we performed qualitative examination of video only on items with poor item characteristics for this study.

Chapters 2, 3, and 7 of this monograph give a description of specific Rasch analyses of the separate dimensions for length, area, and volume, respectively. These specific analyses provide further quantitative support in the validation and refinement of the developmental progressions in each measurement domain.