12-1-2017

Assessing Teacher Attentiveness to Student Mathematical Thinking

Michele B. Carney
Boise State University

Laurie Cavey
Boise State University

Gwyneth Hughes
University of Wisconsin-Madison

ASSESSING TEACHER ATTENTIVENESS TO STUDENT MATHEMATICAL THINKING

Validity Claims and Evidence

ABSTRACT

This article illustrates an argument-based approach to presenting validity evidence for assessment items intended to measure a complex construct. Our focus is developing a measure of teachers' ability to analyze and respond to students' mathematical thinking for the purpose of program evaluation. Our validity argument consists of claims addressing connections between our item-development process and the theoretical model for the construct we are trying to measure: attentiveness. Evidence derived from theoretical arguments in conjunction with our multiphased item-development process is used to support the claims, including psychometric evidence of Rasch model fit and category ordering. Taken collectively, the evidence provides support for the claim that our selected-response items can measure increasing levels of attentiveness. More globally, our goal in presenting this work is to demonstrate how theoretical arguments and empirical evidence fit within an argument to support claims about how well a construct is represented, operationalized, and structured.

Michele B. Carney
Laurie Cavey
BOISE STATE UNIVERSITY

Gwyneth Hughes
UNIVERSITY OF WISCONSIN–MADISON

THERE have been recent calls to better document how teacher preparation and professional development programs prepare and support teachers (e.g., CAEP Commission on Standards and Performance Reporting, 2013; Grossman, Hammerness, & McDonald, 2009; Lampert, 2009). One of

THE ELEMENTARY SCHOOL JOURNAL
Volume 118, Number 2. Published online October 30, 2017
 0013-5984/2017/11802-0005$10.00

many tools that might help in this endeavor are large-scale assessments for the purpose of program evaluation. Although assessments like the Praxis (Educational Testing Service, 2016) have long been used for individual teacher evaluation of content knowledge, there is growing interest in developing assessments for program evaluation of constructs related to practices central to teaching mathematics (e.g., LMT, 2005). Such assessments enable the examination of changes within a single program or comparison across multiple programs. This article recounts our efforts to create an assessment to measure teachers' attentiveness, which we define as teachers' ability to analyze and respond to students' mathematical thinking, to assist in program evaluation and research. But given such a complex practice, how does one argue that the assessment captures the core elements of that construct?

The goals of this article are twofold. First we seek to describe our Mathematical Attentiveness in Pedagogical Practice (MAPP) item-development process for teacher educators and researchers who may use this specific assessment. A second, broader goal is to illustrate an argument-based approach to providing evidence of the relationship between a construct posited to be assessed and the items intended to measure that construct. To set the stage, we situate the need to assess the particular construct under investigation—attentiveness—within the teacher and mathematics education literature.

Teacher education leaders are calling for an end to organizing programs in ways that dichotomize teaching practices and the disciplinary knowledge needed to teach (Darling-Hammond & Bransford, 2007; Grossman, Hammerness, et al., 2009; McDonald, Kazemi, & Kavanagh, 2013). Ball, Sleep, Boerst, and Bass (2009) hypothesize that a focus on integrating knowledge and practice in teacher education will lead to better preparation for the complex work of teaching. Likewise, we see the need to integrate knowledge and practice in the evaluation of our teacher education programs. The integration of knowledge and practice around teachers' ability to attend to students' mathematics thinking from a progressive formalization perspective is a particularly important construct within the literature on mathematics teacher education (Freudenthal, 1973; Gravemeijer & van Galen, 2003; Lampert et al., 2013; Pierson, 2008; Treffers, 1987) and a critical component of teacher education programs. It is our aim that teachers exit programs with a better understanding of how students' ideas develop and how to productively work with those ideas in the classroom. We use the term *attentiveness* to refer to the construct of interest throughout this article. Because our work involves both preservice and in-service teachers, we use the term *teacher learner* to encompass both groups; and because we examine both initial licensure and professional development programs, we use the term *teacher education program* to encompass both.

Because of the important role of attentiveness in classroom instruction, it is necessary to develop assessments that enable the examination of how the construct builds over time and is influenced by specific teacher preparation activities or professional development programs. Because resources for program evaluation are often limited, we have also incorporated the additional design constraint of developing a measure that can be administered on a large scale, without a tremendous scoring burden. Given the complexity of our theoretical construct and the desire for items that are easy to administer and score, a critical first step with respect to validity is establishing that item responses are linked to our theoretical model for

attentiveness. That is, we need to provide evidence of construct representation—evidence that we have operationalized the construct in a manner representative of its depth and breadth.

In this article, we provide an illustration of an argument-based approach to presenting specific claims and associated validity evidence as it relates to construct representation (further details are provided below). We do this by describing the construct attentiveness (the basis of the assessment) through a theoretical model situated in the research literature (Claim 1) followed by an explanation of the MAPP item-development process in relation to this model (Claim 2). We then provide evidence (for Claims 3, 4, and 5) that item responses are reflective of the theoretical model. For the purposes of this article, we use an example item, designed to assess attentiveness in middle-grades data analysis and statistics. The careful illustration of these components of assessment development further clarifies the meaning of the construct while providing an example of construct representation validation efforts for others involved in assessment development. Next, we situate construct representation within the larger measurement validity literature and describe how it relates to our assessment development work.

Validity and Construct Representation

Historically, validation efforts around assessments have referred to validity as a property of the assessment (i.e., language referring to a valid and reliable assessment). Despite numerous recommendations to the contrary (e.g., Kane, 2006; Messick, 1995; Newton, 2012), this language is still prevalent in the research literature (Cizek, Rosenberg, & Koons, 2008). Validity theory experts such as Michael Kane and Samuel Messick, and more specifically, the last two versions of the Standards of Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014), call for validation efforts to focus on proposed interpretations and uses of test scores. Relatedly, Kane (2001, 2006) calls for an "argument-based" approach to validation, in which the assessment developer states claims related to test score interpretation or use through an interpretive argument and then provides evidence to justify these claims through a validity argument. Despite the consistent articulation of this approach in the standards, its application in practice is not widespread (Cizek et al., 2008). There are high-quality examples in the literature (Bell et al., 2012; Schilling & Hill, 2007); however, these are generally written for individuals with expertise in measurement.

The shift from validity of an assessment to validity of the interpretation and use of assessment scores largely transfers the burden of determining the appropriateness of a test for a particular situation from the developer to the user (Newton, 2012). Therefore, an argument-based approach places additional onus on the assessment developer to present validation work in a way that is accessible to assessment users (who may or may not have measurement expertise). This validation approach consists of the explicit statement of assessment claims through an interpretive argument, presented with supporting evidence through a validity argument. This structure of claims and arguments provides a framework for presenting an argument-

based approach to validation work that is accessible to nonmeasurement experts. In this article, we present the construct representation aspect of our validation efforts related to an assessment of attentiveness through this organizational approach.

Validation efforts are an ongoing process. Arguments can be made related to specific aspects of score interpretations with the understanding that additional validation effort may be needed to further support score interpretations. For the purpose of examining the relationship between a construct posited to be assessed and the items intended to measure that construct, we have found it helpful to group validation efforts into two distinct but related sets of activities (Embretson-Whitely, 1983). The first, referred to as *construct representation*, establishes connections between the theoretical model for the construct and the assessment items posited to measure that construct (similar to elements involved in internal investigations as described by Lissitz & Samuelsen, 2007). Gathering validity evidence associated with construct representation is typically undertaken before and during item development and is particularly important for complex constructs. A second aspect of validation efforts (Embretson-Whitely's, 1983, "nomothetic span") examines the relationships between test scores and other variables of interest and thus is a future area of research for our assessment (Loevinger, 1957, also refers to this as "external validity"). This article presents an argument to support construct representation of attentiveness in the MAPP item-development process.

Messick (1995) highlighted six aspects of validity, three of which connect specifically to construct representation: content, substantive, and structural aspects of validity. Content validity involves providing evidence (typically theoretical in nature) that the content of the assessment is relevant to and representative of the construct under examination. Content validity for assessments, such as achievement tests, is often established in a test specifications document describing the content that "lies within" the domain, followed by expert analysis of how a particular idea or item is essential to the measurement of the construct.[1] However, for constructs that involve complex, cognitive processes, "domain theory, in other words, scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes" (Messick, 1995, p. 745) is particularly important. Content validity for teachers' cognitive processes related to attentiveness involves justifying (*a*) the relevance and representativeness of the construct's theoretical model from the research literature and (*b*) the operationalization of this model in the item-development process.

Substantive validity involves providing evidence that the theoretical model accounts for the responses or that the responses can be explained by the theoretical model. In other words, substantive validity focuses on the alignment between the theoretical construct posited to be assessed and the empirical results. Substantive validity related to attentiveness involves examining the alignment of hierarchical coding rubrics and scoring models (built from the construct's theoretical model) to empirical responses. Last, structural validity involves providing evidence that empirical responses are consistent with the structure of the construct domain. Structural validity is based on the idea that items should assess the construct under examination and not assess aspects unrelated to the construct. Structural validity related to attentiveness involves demonstrating that we are measuring a unidimensional construct and that there are not underlying relationships between items un-

explained by the theoretical model. Messick (1995) stresses the importance of using these aspects of validity within a comprehensive argument. Therefore, although these aspects should not be used as ad hoc elements to "justify the validity of an instrument," they can be used to assist in identifying the claims and necessary evidence that should be incorporated into a comprehensive argument for the validity of test-score interpretations and uses.

Our intent is to provide an illustration of an argument-based approach to validation addressing Messick's (1995) aspects of content, substantive, and structural validity as they relate to construct representation. We begin by providing an organizational overview in Table 1. The first column contains the five claims from our interpretive argument, and the second column presents the particular evidence for the validity argument. The remainder of this article is organized to address each of the five claims, in order, with their related evidence.

Claim 1: The Theoretical Model for Attentiveness Is Grounded in the Research Literature

We assert in Claim 1 that elements of mathematical knowledge for teaching (MKT; Ball, Thames, & Phelps, 2008), core practices (Ball et al., 2009; Grossman, Compton, et al., 2009), and progressive formalization (Freudenthal, 1973, 1991; Treffers, 1987) can be used to develop and describe a theoretical model for attentiveness. The evidence we present situates our theoretical model for the attentiveness construct

Table 1. Interpretive Argument Claims and Associated Validity Argument Evidence Related to Attentiveness Construct Representation

Interpretive Argument Claims	Validity Argument Evidence
Claim 1: Elements of mathematical knowledge for teaching, core practices, and progressive formalization can be used to develop and describe a theoretical model for attentiveness.	Theoretical justification of our model of attentiveness is situated within the research literature from mathematics and teacher education and includes the differentiation of our model from other models used to develop similar assessments.
Claim 2: Items developed through our multiphased process—that make use of authentic examples of student and teacher thinking—sufficiently represent the cognitive processes required for attentiveness.	Theoretical justification for the multiphased item-development process that explicitly makes use of authentic examples of student and teacher cognitive processes.
Claim 3: Coding rubrics developed to hierarchically categorize teachers' responses to students' mathematical thinking are reflective of the organization of the construct within our theoretical model.	Theoretical justification, in conjunction with empirical analysis, for the development of coding rubrics that operationalize increasing levels of attentiveness.
Claim 4: Teacher learners' responses to free-response items are consistent with the coding rubric structure, indicating consistency with the theoretical model.	Empirical evidence of the consistency of alignment between the coding rubric categories and empirical responses.
Claim 5: Teacher learners' responses to selected-response items are consistent with the scoring model, indicating consistency with the coding rubrics and our theoretical model of increasing levels of attentiveness.	Empirical evidence of the consistency of item responses to the scoring models through Rasch item fit and category threshold ordering.

within the research literature from mathematics and teacher education and differentiates our theoretical model from those used to develop related assessments.

The construct we are interested in measuring, attentiveness, is short for analyzing and responding to students' mathematical thinking from a progressive formalization perspective. We use the term *attentiveness* because it implies a strong focus on understanding and making use of students' thinking in teachers' practice. Progressive formalization is an aspect of the realistic mathematics education theory (Freudenthal, 1973, 1991; Treffers, 1987). Via progressive formalization, students initially apply their mathematical knowledge and intuition to informally model a problem mathematically, or to "mathematize" the problem (Freudenthal, 1991). As students continue to solve problems by using, refining, and formalizing models under the guidance of their teacher, they progressively formalize mathematical ideas and connect them to established conventions.

One can perceive of progressive formalization as a cycle. Students are presented with a task requiring them to model a problem based on their mathematical understanding and personal experience (informal). Then, teachers press students toward more efficient strategies and models (preformal), based on the original student thinking, while emphasizing key mathematical ideas. Finally, mathematical conventions are introduced and connected (formal). Once a concept has been formalized, it then represents part of the student's schema used to solve a new problem, beginning the process again. By tying together content in this way, teaching mathematics can be seen as supporting students' development, application, and formalization of a few key concepts over time rather than as conveying a set of disconnected terms and procedures.

Progressive formalization represents our instructional philosophy. We see teachers' ability to analyze students' work and respond in ways likely to move them to deeper mathematical understanding as paramount to instructional practice. This ability has also been recognized more broadly in mathematics education literature, beyond its specific relevance to progressive formalization. We briefly describe that literature to situate the construct of attentiveness within the broader context of the mathematics education community. We start by providing a description of MKT, which encompasses, among other things, aspects of the knowledge teachers need to carefully attend to student thinking.

Mathematical Knowledge for Teaching

Shulman (1987) made a case for the construct of pedagogical content knowledge as the integration of content knowledge and pedagogical knowledge, providing the impetus for mathematics educators to examine teacher knowledge as having aspects of subject matter content and teaching. Mathematics educators frequently use the construct MKT when referring to the integrated mathematical knowledge specific to the work of teaching mathematics (Ball et al., 2008). Ball and colleagues expanded on Shulman's work by describing MKT as interconnected components of knowledge that include mathematical ideas along with how those ideas relate to certain aspects of teaching, such as curriculum, teaching strategies, and students' ways of knowing. For example, knowledge of students' ways of knowing can be built

up over time by observing and reflecting on how students respond to certain classroom tasks related to particular mathematical ideas, what instructional prompts promote different types of mathematical thinking, what mathematical ideas students come with, and so on.

Assessing teacher knowledge related to the practice of attentiveness relates closely to MKT's specialized content knowledge and knowledge of content and students categories. According to Ball et al. (2008), specialized content knowledge is related to the mathematical work teachers do that is unique to teaching, which includes comparing different models or unpacking a complex concept such as rate. For our purpose of assessing attentiveness within the content area of data analysis and statistics (the content focus of our example item and assessment), we contend this is a domain in which there is clearly a difference between common knowledge about data concepts and the specialized knowledge needed to teach data concepts. As an example, consider how one might unpack knowledge associated with using a linear function to represent bivariate data. How does using a function as a representation of bivariate data connect to using the mean to represent univariate data? How does variability in the data affect how well the function "fits" the data? Considering these questions (and others) is part of the mathematical knowledge specific to teaching this concept that prepares teachers for the work of analyzing and responding to students' ideas.

MKT's knowledge of content and students category includes the ability to recognize common ways students conceptualize particular ideas (Ball et al., 2008, p. 401). Continuing with the data concepts example, recognizing how a student conceptualizes the mean can inform how the teacher responds to that student when she is learning about linear functions as models. That is, being able to quickly recognize student thinking and relate it to particular learning goals enables teachers to respond to students' ideas in ways that take both students' ideas and the learning goals into account. Thus, attending to and building on students' ideas in support of progressively formalizing mathematics requires the ability to recognize and interpret student thinking in relation to important mathematical ideas. This ability, in turn, requires knowledge of how students' ideas develop over time, how a formal conceptualization is related to less formal ideas, and how to effectively press students to consider their ideas in new ways.

Existing Assessments

Several assessments use MKT as a model for building instruments that intend to measure the knowledge teachers use in their practice. Although the different aspects of the MKT model are interlinked, the types of questions posed by an instrument can place emphasis on different parts of that model. For example, the content examinations in the Praxis series (Educational Testing Service, 2016) focus almost exclusively on mathematics content knowledge, ignoring other aspects of MKT. To narrow our review of prior work related to assessing teachers' responses to student thinking, we briefly discuss four instruments that share a common model for conceptualizing teacher knowledge (i.e., MKT) and format (i.e., large-scale administration with short or multiple-choice response) to our own.

The Learning Mathematics for Teaching (LMT) assessments, from the University of Michigan, represent a range of content-area domains and grade bands and are frequently used to evaluate programs designed to develop teachers' MKT (LMT, 2005). LMT items are multiple choice and tend to focus on a combination of mathematics subject matter and pedagogical knowledge. Items based on students' mathematical thinking often ask the examinee to identify a common student misconception or solution processes appropriate to the task. Extensive validity evidence has been provided related to the instruments (Hill, 2005; Hill & Ball, 2004; Hill, Ball, Blunk, Goffney, & Rowan, 2007; Hill, Sleep, Lewis, & Ball, 2007).

The Diagnostic Teacher Assessments in Mathematics and Science, or DTAMS, from the University of Louisville, were created in response to the need for greater middle-school-teacher mathematical content knowledge (Saderholm, Ronau, Brown, & Collins, 2010) and were constructed with a focus on both content and pedagogical knowledge. The items are multiple choice (content) and open-ended (content and pedagogy), the latter requiring either higher level problem solving based on a specific mathematics concept or response to a classroom situation that might arise.

The Knowledge of Algebra for Teaching, or KAT, instrument (McCrory, Floden, Ferrini-Mundy, Reckase, & Senk, 2012) is focused on secondary algebra topics. Questions range from mathematical content knowledge, such as identifying contexts for exponential functions, to specialized content knowledge, such as identifying how equivalent expressions are related to a given context.

Although all three assessments rely on theories of learning that assume students' conceptual understanding is important, they are essentially pedagogically agnostic—they were not built from a specific instructional theory. Any of these assessments might play an important part in program evaluation within the construct of teachers' MKT, particularly within the realm of content knowledge. But although all three assessments seek to measure teachers' pedagogical knowledge, they do not do so with a specific focus on progressively formalizing students' understanding based on their current conceptions.

Finally, the Teacher Analysis of Student Knowledge (TASK; Supovitz, Ebby, & Sirinides, 2013) presents teachers with authentic student responses to a problem and then asks a series of open-ended questions around that work. The TASK is designed to capture a similar aspect of the pedagogical knowledge that we are trying to measure, through a focus on learning trajectories, and could be very valuable at the individual-course scale. But because it requires a great deal of analysis and coding, it is not ideal for large-scale program evaluation.

Existing assessments address constructs that fall within the MKT framework and, often, specifically within common and specialized content knowledge. Although some items may address aspects of pedagogical content knowledge, such as items that ask teachers to identify appropriate contexts for a particular mathematical operation or to interpret multiple-solution strategies to a problem, they are primarily focused on mathematical knowledge. Attentiveness integrates mathematical knowledge with a progressive formalization perspective because it includes determining what a student understands and using that information to respond to the student in a manner that builds on his or her thinking. This requirement that mathematical knowledge is used to build on a particular student's understanding is a distinct pedagogical perspective and not a strong element of current MKT-based assess-

ments. Demonstrating MKT as described in the MKT framework is a necessary but not sufficient condition for demonstrating a high level of attentiveness. Therefore, although existing measures represent a valuable part of assessing MKT, we see a need for an assessment that captures evidence for teachers' ability to interpret and respond to authentic student work from a progressive formalization standpoint and yet does not require qualitative coding.

Attending to Students' Mathematical Thinking

Many constructs, such as decentering (Teuscher, Moore, & Carlson, 2015), Mathematically Significant Pedagogical Opportunities to Build on Student Thinking (MOST; Leatham, Peterson, Stockero, & Van Zoest, 2015), responsiveness (Pierson, 2008), and professional noticing (Jacobs, Lamb, & Philipp, 2010), exist within the teacher and mathematics education literature related to attending to students' thinking more generally. For example, professional noticing involves investigating the teacher's ability to recognize mathematical aspects of students' ideas along with knowledge of how certain pedagogical approaches may (or may not) support the advancement of student understanding (Mason, 2011). Jacobs et al. (2010) operationalized the noticing construct through a framework for prompting teacher reflection on the ideas expressed by a single student during a videotaped interview along with how teachers might respond to students.

Furthermore, recent teacher education efforts point to the need to identify "core" or high-leverage practices—regular routines that teachers use to engage students in productive intellectual work (Grossman, Hammerness, et al., 2009). Effective implementation of a core practice requires the enactment of integrated knowledge of teaching, subject matter, students, and curriculum (Grossman, Schoenfeld, & Lee, 2007). The Learning Teaching in, from and for Practice project (Lampert et al., 2013) has identified *eliciting and responding to students' contributions* as a significant aspect of teaching that can be developed in preservice teachers through their model. In our work with teacher learners, we emphasize careful attention to students' ideas while keeping the important mathematical goals of the task in mind, similar to the type of ambitious teaching described by Lampert and colleagues. In short, we see attentiveness as a core practice of teaching, closely aligned with eliciting and responding to students' contributions.

Summarizing Attentiveness

Attentiveness differs from other constructs (e.g., decentering and MOST) through the explicit couching of attentiveness within the progressive formalization instructional theory. We see attentiveness as the coordinated ability to employ MKT in conjunction with pedagogical moves associated with progressive formalization of mathematical ideas. The aspects of attentiveness in which we are primarily interested are those foundational to progressive formalization. These include being able (*a*) to identify and understand key mathematical ideas from a specialized content-knowledge perspective, (*b*) to recognize and interpret students' informal solution strategies in relation to more formal ways of reasoning, and (*c*) to respond to students in an effort to build on students' understanding in a way that challenges them to progressively formalize or reformalize their ideas.

Because of the interplay of the mathematical ideas, student thinking, and pedagogical moves involved, evidence of attentiveness can be exhibited at different levels of specificity. For example, a teacher's response to a student that is generic enough to apply to almost any student at any time (e.g. "Can you think of another way to solve this problem?") provides no information about whether that teacher is striving to teach in ways that align with a progressive formalization perspective. Conversely, a teacher's analysis of student work that consists only of an evaluation of correctness provides some information about the mathematical ideas to which the teacher is attending but no information about whether that teacher is attending to students' thinking. Thus, the greater the specificity in a teacher's pedagogical moves with respect to aspects of attentiveness, the better our position is for making reasonable inferences about the alignment of teacher responses with progressive formalization.

Thus far in our argument we have addressed how the attentiveness construct is related to progressive formalization, MKT, and the core practice of eliciting and responding to students' contributions (Claim 1). In the next section, we continue the argument by describing how the item-development process captures a range of teachers' cognitive processes associated with attentiveness (Claim 2). Later, when addressing Claim 3, we describe how we have operationalized attentiveness in a rubric consisting of four interrelated components of teacher knowledge, built from the professional noticing framework (Jacobs et al., 2010), at three increasing levels of specificity (focused on progressive formalization).

Claim 2: Items Approximate Cognitive Processes of Attentiveness

We assert in Claim 2 that items developed through the multiphased MAPP item-development process, which makes use of authentic examples of student and teacher thinking, sufficiently represent the cognitive processes required for attentiveness.

General Description of Development Phases

The MAPP item-development process culminates in Phase 4 with a selected-response item (see Fig. 1, right-side column, for example item) designed to measure teacher learners' abilities to analyze and respond to student thinking. Each item asks the teacher-learner respondent to rank a series of exemplar teacher-learner responses to prompts about a student's solution to a statistical task. By asking teacher learners to rank given responses to student work, we transform what typically requires extensive coding into an easy-to-administer selected-response-item format. The assessment items are designed with the intent of capturing evidence of knowledge that transfers to practice, although this is an empirical question to be addressed in a future study. Each item is based on important mathematical ideas and includes authentic examples of both student and teacher responses. The MAPP item-development process involves four sequential phases (summarized in the left-side column of Fig. 1).

In Phase 1, we develop a cognition matrix for student learners for a specific concept within a mathematical domain (in this case, statistics), focusing on progres-

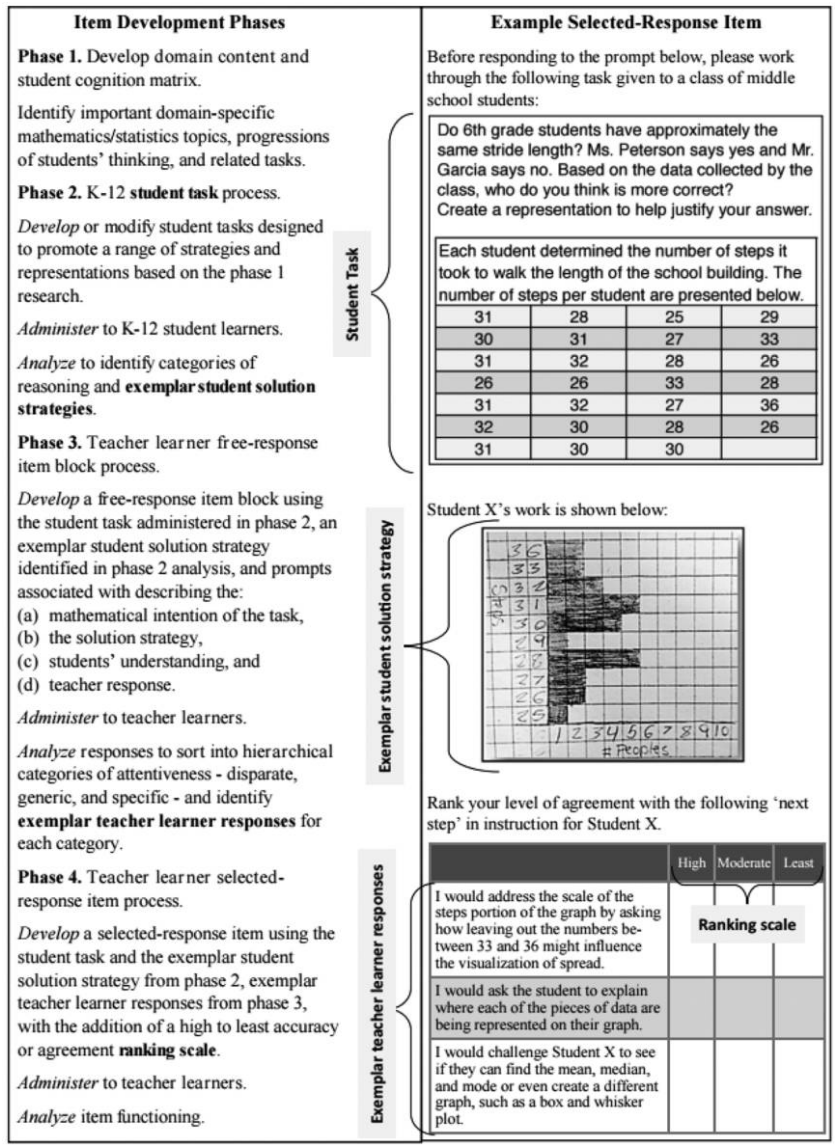


Figure 1. MAPP item-development phases (left column) and Phase 4 selected-response example item (right column).

sions of student thinking in line with progressive formalization. In Phase 2, we develop K–12 student tasks designed to promote a multiplicity of strategies and representations. Following administration to K–12 student learners, we analyze student solution strategies to identify common categories of student reasoning and exemplar student solution strategies. In Phase 3, we develop free-response item blocks for teacher learners using the student task and exemplar student solutions from Phase 2. The free-response prompts involve describing the mathematical intention of the task, the student's approach or understanding, and how the teacher learner would respond to the student. The exemplar student solution strategies are intended to garner a range of responses from teacher learners, reflecting various analyses and

pedagogical strategies. Phase 3 also includes analyzing teacher learners' responses in relation to categories of increasing levels of attentiveness—disparate, generic, and specific—and identifying exemplar teacher-learner responses for each category. The selected-response items developed in Phase 4 include exemplar teacher-learner responses, one from each hierarchical category of attentiveness, with the addition of a high to least accuracy or agreement ranking scale. The data from the Phase 4 selected-response items are then scored and analyzed in relation to increasing levels of attentiveness.

Through this process, our goal is to generate selected-response items from authentic teacher and student responses, representing a range of mathematical ideas embedded in instructional practice. We do this by administering the Phase 3 and 4 items to teacher learners with a variety of experience for a particular mathematical topic. This response variety can be achieved either by sampling across programs or, as in the exemplar below, by sampling from the same set of teacher learners before and after a professional development program. Likewise, student tasks (Phase 2) can be administered either before topic instruction or across grade levels to achieve the range of responses one might expect to see in practice.

Claim 2 Exemplar

The right side of Figure 1 provides an example Phase 4 selected-response item involving informal notions of variability in quantitative data. In Phase 1 (not depicted in Fig. 1), we examined literature on how students move from informal notions of center and variability toward deep conceptual understanding and fluency with standard representations (Langrall, Mooney, & Phillips, 2002; Reading & Shaughnessy, 2004). Based on this Phase 1 work, in Phase 2 we developed a student task related to distributions of children's stride length, designed to elicit a range of strategies and representations from students who had not yet been introduced to formal notions of center and variability. Teachers in a professional development program administered the item to two sixth-grade classes as part of their regular instruction. We analyzed the student solution strategies to identify work samples that captured students' informal ways of representing center and variability. One such exemplar strategy (identified as Student X's work in Fig. 1) was used in a Phase 3 free-response item block that was administered to 79 teacher learners. The teacher-learner responses were analyzed based on a rubric (described in Claim 3) from which exemplar teacher-learner responses were selected for use in development of the Phase 4 item. The final Phase 4 item intends to assess teacher learners' ability to respond to a student's relatively informal representation. We revisit different phases of the development of this item as we address the claims associated with our construct representation argument.

At this point in our argument, we have addressed aspects of content validity by establishing the theoretical basis for attentiveness (Claim 1) and sufficient representation of student and teacher cognitive processes through our MAPP item-development process (Claim 2). In the next section, we shift the focus of our discussion from content to substantive validity to justify the way we have operationalized attentiveness in the MAPP item-development process.

Claim 3: Attentiveness Coding Rubrics Are Reflective of the Theoretical Model

We assert in Claim 3 that the coding rubrics developed to hierarchically categorize teachers' attentiveness are reflective of the organization of the construct within our theoretical model. The evidence to support Claim 3 comes from the development of coding rubrics to describe increasing levels of each component of attentiveness through an iterative process of theory informing analysis and analysis informing theory.

We described the construct of attentiveness in Claim 1 as grounded in MKT, with direct ties to the practice of eliciting and responding to students' contributions through the instructional perspective of progressive formalization. We operationalized attentiveness through four components. The first component relates to teachers' ability to identify and describe the mathematical intention of a student task. This draws strongly from MKT, with a specific focus on specialized content knowledge. The next three components are motivated by the professional noticing construct operationalization (Jacobs et al., 2010): teachers' ability to describe the student's approach to solving a task (second component), teachers' ability to describe the student's likely underlying mathematical understanding (third component), and teachers' instructional response to the student (fourth component).

Phase 3 items are structured to address each component of our operationalization of attentiveness through the presentation of a student task and an associated exemplar student solution strategy, followed by free-response item prompts related to each component. The appendix provides an example Phase 3 item.[2] The resulting free responses from teacher learners must then be analyzed in relation to attentiveness. We needed to develop coding rubrics for this purpose that were reflective of our attentiveness construct, including its hierarchical structure. Following Clement, Chauvot, Philipp, and Ambrose (2003), our coding rubrics were developed through the interplay of our theoretical assumptions and empirical survey data. We describe our process for rubric development, followed by a description of the final version of the rubric.

To develop rubric descriptions for how each component could be hierarchically categorized into increasing levels of attentiveness, we began by sorting responses to a particular prompt (e.g., "If you were the teacher of Student X, please describe in detail how you would respond to Student X") into groups of high, medium, and low levels of attentiveness based on our theoretical description. We then examined these groupings to identify and describe commonalities in the responses. This began with item-specific descriptions of the groupings, but once this process was conducted across several items and mathematical topics, we were able to examine these groups for commonalities across items. The commonalities in responses across items but within a group were examined in relation to our theoretical model for attentiveness. This interplay between empirical data and theory led to the development of three categories that represent increasing levels of attentiveness: *disparate, generic*, and *specific*. The disparate category represents incorrect thinking or particularly tangential responses. The generic category includes responses that provide a related but relatively generic response or a partial analysis. The specific category indicates a high level of evidence for attentiveness through a targeted response to the

item that often includes a high level of detail. Table 2 provides an overview of the components of attentiveness in relation to the category descriptions of each component. Further articulation of the evidence for the relationship between the rubric and the attentiveness construct was provided by Cavey, Carney, and Hughes (2017).

Although the coding rubrics we developed to hierarchically categorize teachers' attentiveness provide a common tool for evaluating teachers' Phase 3 responses, specific item-coding rubrics are developed for individual items to ensure that the aspects of the four components are clearly articulated in relation to the task and student work samples. An example of a specific item-coding rubric is provided in Table 3 in the next section addressing Claim 4.

Now that we have addressed how the interplay between theory and empirical data resulted in the development of coding rubrics (Claim 3), we continue our argument by examining the alignment between an item-specific coding rubric and empirical responses from the third phase of the MAPP item-development process.

Claim 4: Free-Response Item Data Are Consistent with the Attentiveness Coding Rubric

We assert in Claim 4 that teacher learners' responses to free-response items are consistent with the attentiveness coding rubric structure, indicating consistency with the theoretical model. The evidence for Claim 4 comes from the examination of the frequency of responses within each category of attentiveness for a particular item.

Claim 4 Exemplar

Figure 1 presented the multiphased MAPP item-development process specific to an item that involves a student task (the stride-length task) and student work sample focused on analyzing center and spread for univariate data. In Phase 3 of the MAPP item-development process, the item block presented in the appendix was administered to teachers. The sample of responses to this Phase 3 item block came from two sections of a graduate-level course for in-service teachers focused on teaching data analysis and statistics concepts in kindergarten through grade 9. The 72 responses were drawn from 36 people, with each person having taken the online inventory before and after participation in the course. An additional seven responses (for 79 responses total) were from individuals who responded to either the pre- or postcourse inventory request (but not both). The pre- and postcourse sampling was conducted to provide a broad range of responses, with individuals at the end of the course ideally having a higher level of knowledge for teaching data analysis and statistics, including the ability to be more attentive to student thinking. The individuals in the course varied in teaching experience, although the majority had 6 or more years of experience. Throughout the course, participants indicated that they felt relatively weak in the area of knowledge on data analysis and statistics.

We present the analysis of the teachers' responses to the item prompt, "If you were the teacher of Student X, please describe in detail how you would respond." The 79 responses were coded using the rubric described in Claim 3, with further

Table 2. Coding Rubric Relating Components of Attentiveness to Categories of Increasing Levels of Evidence

	Categories and Their Descriptions for Attentiveness Evidence		
Components of Attentiveness	Disparate	Generic	Specific
Mathematical intention	The disparate mathematics category involves responses that go well beyond the focus of task, are particularly tangential, or incorrect.	The generic mathematics category involves responses that are related to the mathematical intention but do not focus on the mathematical intention of the task.	The specific mathematics category involves responses that are closely related to the mathematical intention of the task.
Student process or approach	The disparate process category involves responses that are extremely generic, incorrect, or only tangentially related to the student work sample.	The generic process category involves responses that represent relatively generic or only partial analysis of the student work sample.	The specific process category involves responses that appropriately and fully describe the student's work sample.
Student understanding	The disparate understanding category involves responses that over extrapolate a student's understanding of the task, provide an unspecific description of their understanding, or are incorrect.	The generic understanding category involves responses that provide a generally correct but relatively underspecified description of the student's understanding.	The specific understanding category involves responses that provide a specific description of student understanding, oftentimes in relation to the mathematical intention of the task.
Teacher response	The disparate response category involves responses that praise work, are overly prescriptive, redirect (rather than build upon) student thinking, or are incorrect.	The generic response category involves responses that are generically appropriate across multiple student work samples (e.g., explain your work), or attends to the mathematical intention or student's understanding (but not both).	The specific response category involves responses that attend to both the mathematical intention and the student's understanding in a manner likely to press student thinking.

Table 3. Specific Item-Coding Rubric and Category Frequencies for the Phase 3 Item in the Appendix for the Prompt Related to the Teacher's Response to the Students

Response Coding	Frequency (%)	Category Description	Example Responses
Disparate	25	The disparate response category involves responses that praise work, are overly prescriptive, redirect (rather than build upon) student thinking, or are incorrect. For the stride-length task, this often involved a response focused on procedural aspects of data analysis, such as calculating the mean and median without explicit focus on these calculations as a tool to respond to the original prompt.	"I would challenge Student X to maybe see if she can find the mean, mode, median, and range or even create a different graph, such as a box and whisker plot." "Looks good. Can you go tell that group what your thinking was."
Generic	59	The generic response category involves responses that are generically appropriate across multiple student work samples, or attend to the mathematical intention or student's understanding (but not both). For the stride-length task, responses in this category often involved redirecting the student to respond to the original task prompt but without a focus on the representation created by the student.	"I think I would first ask the student to explain where each of the pieces of data are being represented on the graph." "I would ask Student X what he thought of his findings and see which teacher he agreed with and why."
Specific	11	The specific response category involves responses that attend to both the mathematical intention and the student's understanding in a manner likely to press students' thinking. For example, asking a student how her graph addressed the question and then address a specific aspect of the student's graph (e.g.. the scaling of the *y*-axis).	"I would address the scale of the steps portion of the graph. I would ask if she agrees with Ms. Peterson or Mr. Garcia and why." "I would ask the student to explain how the graph justifies the answer. I would ask why she chose to represent the data this way. I might ask why she left out numbers between 33 and 36."
Other	5	The other response category involves responses that indicated respondents were not sure of the answer or did not respond to the item.	"This students graph shows a clear relationship between the amount of steps taken with the frequency that they occurred. It shows that the majority of data is clustered together."

category description added specific to the item's mathematical intention and student work sample (see Table 3).[3] In addition, we identified example responses within each category to assist in coding and as potential selections for the development of a selected-response item. The category frequencies were disparate (25%), generic (59%), specific (11%), and other (5%). Based on these percentages, there is a strong association between the rubric categories and the sample of responses, indicating that responses are consistent with the rubric structure and theoretical model. If a large percentage of responses had been coded as *other*, this would potentially indicate a lack of consistency between the coding rubric structure and the theoretical model. However, the relatively small percentage of responses in this category supports the claim of consistency between the coding rubric and the theoretical model. As demonstrated with this exemplar, an examination of the consistency of the response to the coding category structure is used with each item in the MAPP item-development process to determine whether a Phase 4 selected-response item should be constructed from the Phase 3 responses. When the responses do not align with the coding rubric, they typically are extremely varied (numerous categories) or very narrow. This indicates an issue with the item structure and a need to either revise the item or remove it from further development.

At this point in our argument, we have presented evidence related to Messick's (1995) content validity (Claims 1 and 2) and substantive validity (Claims 3 and 4). Evidence for Claim 4 also addressed structural validity concerns, which we continue to address in the next section by examining the consistency of empirical responses from the fourth phase of the MAPP item-development process to the theoretical and scoring models.

Claim 5: Teachers' Responses to Selected-Response Items Are Consistent with Scoring Model

We assert in Claim 5 that teachers' responses to selected-response items are consistent with the scoring model, indicating consistency with the theoretical model of increasing levels of attentiveness. We analyze the selected-response item data (Phase 4) with the Rasch partial credit model (Masters, 1982) and present evidence for this claim through examination of Rasch item fit statistics and the ordering of the category thresholds.

The fourth and final phase of the MAPP item-development process involves developing selected-response rank-order items, which are similar in format to our free-response rank-order items but ask respondents to rank their levels of agreement with a list of available teacher responses instead of describing their own responses. Each selected-response rank-order item typically consists of three exemplar teacher-learner responses from Phase 3—one from each category of disparate (0), generic (1), and specific (2)—and three ranking scale categories, associated with different levels of agreement. The ranking scale categories differ depending on the prompt but typically contain classifications, such as high, moderate, and least agreement. The development of a block of selected-response rank-order items is followed by administration to teacher learners, scoring of rank-order responses, and subsequent analysis.

Scoring Model

When completing an item, teacher learners can select a ranking category only once, thereby forcing a most- to least-accurate (or high to low) ranking of the three teacher responses. Teacher learners' choices for each item are then scored such that a respondent can score 2 points (full credit), 1 point (partial credit), or 0 points (no credit). Within a selected-response item (e.g., see appendix), if a respondent selected the "correct" educator response for the most accurate or high agreement category, they received 1 point. Similarly, if a respondent selected the "correct" educator response for the least accurate or least agreement category, they also received 1 point. The middle category of moderately accurate or agree is not directly scored. Following the scoring, Rasch analysis is used to examine whether the scoring model is consistent with the theoretical model through examination of item fit and ordering of the category thresholds.

Rasch Model

When data meet Rasch-model requirements, the model transforms ordinal observations into an equal-interval scale, meaning differences in test takers' performance (termed *person ability*) and item difficulties (for dichotomous data) or response thresholds (for polytomous data) are represented as an interval relationship versus the traditional ordinal ranking received from totaling raw scores or calculating a percentage correct (Merbitz, Morris, & Grip, 1989; Wright & Linacre, 1989). Estimates obtained from a Rasch analysis situate person ability and item difficulty along a common equal-interval scale (Bond & Fox, 2013). As a result, person ability and item category thresholds can be interpreted in relation to one another probabilistically.

In a partial credit Rasch model—such as the one we are using for the conversion of our rank-order item responses to scores of 0, 1, or 2—category thresholds (Rasch-Andrich thresholds) are used to indicate where along the continuum the probability of a person receiving a score of 0 or 1 is equally probable (Threshold 1) and where the probability of a person receiving a 1 or 2 is equally probable (Threshold 2). Examination of the hierarchical relationship among person abilities and category thresholds on a common interval scale lends itself to validation efforts with respect to the consistency of the scoring model with the theoretical model (Wolfe & Smith, 2006). Andrich, De Jong, and Sheridan (1997, p. 59) state that "central to the understanding of what constitutes more or less of the property on the continuum is the definition of the successive categories which reflect successively more of the property. However, there is no guarantee that the categories will operate in the way intended. Therefore, this ordering must be treated as a hypothesis about the data and it is important that the statistical model applied has the property it can reject this hypothesis. That is, the ordering must be a property of the data themselves, not simply the model." In the next section, we examine whether our partial credit scoring system results in ordered thresholds, thus providing evidence to support the claim that the scoring model is consistent with the coding rubrics and the theoretical model for increasing levels of attentiveness.

Claim 5 Exemplar

The Claim 5 exemplar involves description of (*a*) the development and scoring for the Phase 4 example item; (*b*) an assessment (in which the Phase 4 example item is embedded), its administration, and the respondent sample; and (*c*) the Rasch analysis with a focus on examination of the fit statistics and category threshold ordering of the Phase 4 example item. In addition, we include the fit statistics and category threshold ordering of all items involved in the assessment administration to provide an illustration of how fit statistics and category threshold order can be used diagnostically to assist in the MAPP item-development process, with a particular focus on ensuring consistency with the theoretical model (Andrich et al., 1997).

Development and scoring of a selected-response exemplar item. The responses to the free-response item described in Claim 4 (and associated with Phase 3 of the MAPP item-development process) were used to identify three teacher responses for development of the selected-response rank-order item. This process requires multiple considerations. First, we wanted one response from each of the three coding categories representing increasing levels of attentiveness (disparate, generic, and specific). This was necessary to ensure that the potential selections were representative of a continuum of less to more attentiveness. For the disparate category, we selected a response that redirected the student to focus on more procedural aspects of data analysis and did not build on the graph created by the student: "I would challenge Student X to see if he or she can find the mean, median, and mode or even create a different graph, such as a box and whisker plot." We anticipated that teachers who primarily focused on procedural aspects of data analysis and were not focused on building on students' thinking were likely to select this response based on familiarity with the terms. For the generic category, we selected a response that could be used across almost any representation created by students for this particular task: "I would ask the student to explain where each of the pieces of data are being represented on the graph." We selected this prompt because it has the potential to build on students' thinking, depending on how the teacher follows up on the student's response but does not clearly focus on aspects of center and spread within the data set. For the specific category, we selected "I would address the scale of the steps portion of the graph by asking how leaving out the numbers between 33 and 36 might influence the visualization of spread," given its focus on the representation created by the student and how that particular representation could influence understanding of spread. Although the wording of the exemplar teacher responses are similar to those gathered in Phase 3, the exact wording is often slightly adjusted for ease of interpretation. The Phase 4 selected-response item is presented on the right side of Figure 1.

If an individual selected "high agreement" for the response, "I would address the scale of the steps portion of the graph by asking how leaving out the numbers between 33 and 36 might influence the visualization of spread," he or she received 1 point. If an individual also selected "least agreement" for the response "I would challenge Student X to see if they can find the mean, median, and mode or even create a different graph, such as a box and whisker plot," he or she would receive an additional point, resulting in 2 points or full credit. If an individual selected only one of these two options, he or she would receive 1 point or partial credit. If an in-

dividual did not select either of these options, he or she would receive 0 points or no credit. For this example item, 37 respondents received no credit (0), 32 respondents received partial credit (1), and seven respondents received full credit (2).

Exemplar assessment sample and administration. The example item was embedded in a 12-item assessment. All 12 items were developed through the multiphased MAPP item-development process, with two items focused on the mathematical intention, four items focused on the students' understanding, and six items focused on the teachers' responses. The assessment was administered online to a sample of 76 secondary-level teachers (grades 6–12) involved in a professional development project on data analysis and statistics instruction. Although the majority of the respondents taught mathematics in grades 6–8, 27 respondents taught mathematics in grades 9–12. The assessment was administered approximately 2 months after participation in a 4-day summer workshop but before follow-up support during the school year. Our goal was to administer the assessment to a sample that represented varying levels of attentiveness. By administering the assessment after the summer professional development workshop, we hoped to increase the level of attentiveness of some individuals to assist in creating a spread in respondents' ability.

Rasch analysis. We analyzed the data using the partial credit Rasch model in WinSteps Version 3.92.1 (Linacre, 2016). For the purposes of our initial item development, fit indices ranging from 0.7 to 1.3 for infit and outfit mean square (MNSQ) were considered acceptable (Bond & Fox, 2013). MNSQ outfit was outside this range for four items. Examination of response patterns for two of the items with high MNSQ outfit (>1.3) revealed one respondent who received a score of 2 on these two items but a score of 0 on all but one of the remaining items. Given the lack of consistency in the individual's responses, these two scores were converted to missing data before proceeding with data analysis (as recommended by Linacre, 2010). MNSQ outfit for two items remained just outside the acceptable range (CA_MathInt = 0.69 and BFMResp = 0.68); however, given that overfit is less of an issue of concern in terms of model fit (Bond & Fox, 2013, p. 240), we proceeded with analysis of the category thresholds.[4]

Table 4 provides the item labels and thresholds (with standard errors) for the 12 items administered in the assessment. Next, we provide an analysis of the data related to the example item (SRXResp), followed by analysis of a different item (not previously discussed) that has disordered category thresholds (SAHResp) to provide an illustration of how the Rasch model can serve as a diagnostic tool for improving consistency between scoring and theoretical models.

Item SRXResp (presented in Fig. 1) is the selected-response version of the exemplar discussed in Claims 2 and 4. An examination of the category threshold measures reveals that the first threshold measure of –0.37 (the location on the continuum where it is equally likely that a respondent with that person ability would score a 0 or 1) is 2.08 logits lower in difficulty than the second threshold of 1.71 (the location on the continuum where it is equally likely that a respondent with that person ability would score a 1 or 2). A plot of the probability of scoring within a particular category in relation to the person ability is provided in Figure 2*a* to assist in visualizing the meaning of the thresholds. The point of intersection between the black line (representing a score of 0) and the darker gray line (representing a score of 1) is Threshold 1, and the point of intersection between the darker

Table 4. Item Labels and Threshold Measures (Standard Error) of the 12 Phase 4 Selected-Response Items from Claim 5

	Measure	
Item Label	Threshold 1 (*SE*)	Threshold 2 (*SE*)
BoCResp	1.25 (.30)	1.59 (.56)
MaCUndSta	−.84 (.31)	−.70 (.27)
BS_MathInt	−.82 (.28)	.36 (.29)
BSRUndSta	−1.06 (.31)	−.19 (.27)
BSQResp	.15 (.27)	.33 (.32)
SRXResp	−.37 (.26)	1.71 (.42)
BFMResp	−1.36 (.33)	−.47 (.26)
CA_MathInt	−.03 (.26)	1.25 (.39)
FPVResp	−.67 (.28)	.40 (.30)
FPWUndSta	−1.16 (.30)	.37 (.29)
SAHResp[a]	.67 (.27)	−.06 (.32)
SDLUndSta	−.52 (.28)	.17 (.29)

[a] Item with disordered thresholds.

gray line and the lighter gray line (representing a score of 2) is Threshold 2. Looking at the graph, one can see that respondents with overall person ability scores less than −0.37 are most likely to score 0 on this item, with the likelihood decreasing as the person ability increases. Those respondents with overall personal ability scores between −0.37 and 1.71 are most likely to score 1, with the likelihood peaking at the midpoint. Respondents with overall personal ability scores greater than 1.71 are most likely to score 2, with the likelihood increasing as the person ability increases. The threshold ordering for this item—and all other items except SAHResp—provide evidence to support our claim that the scoring model is consistent with the theoretical model regarding successive levels of attentiveness.

The threshold measures for item SAHResp indicates a disordering of the scoring categories. The first threshold measure of 0.67 is higher in difficulty than the second threshold of −0.06. Consider Figure 2*b* to visualize the meaning of the disordered thresholds. By examining the dark gray curve, we see that there is never a person ability range for which the probability of a score of 1 is the greatest—that is, the likelihood of scoring a 0 or 2 is always greater than the likelihood of scoring a 1, regardless of the person ability. If this had occurred for the majority of the items on the assessment, it would call into question the claim related to the consistency of the scoring model with the hypothesis that successive scoring categories are indicative of higher levels of attentiveness. However, because it occurred only for this one item, it potentially indicates an issue with the particular item itself related to the responses we selected to represent the disparate, generic, and specific coding categories or poor estimates due to low category frequencies. Therefore, we use the threshold ordering not only as evidence to support our claims but also as a diagnostic tool to assist in item improvement. The selected-response options for item SAHResp will be modified before our next administration.

In Claim 5 we provided evidence for the consistency of empirical responses from the fourth phase of the MAPP item-development process to the theoretical and scoring models. This concludes our argument for construct representation of attentiveness.

Discussion

Our work is focused on developing an assessment to measure attentiveness, which we see as teachers' ability to analyze and respond to students' mathematical thinking from a progressive formalization perspective. We see three primary contributions of our work to the broader academic conversation around mathematics teacher education specifically and educational assessment in general. First, introducing and describing the construct of attentiveness draws together several strands of current mathematics and teacher education research. Second, the presented validity arguments, specific to construct representation for attentiveness through the MAPP item-development process, will aid teacher educators in evaluating the fit of assess-

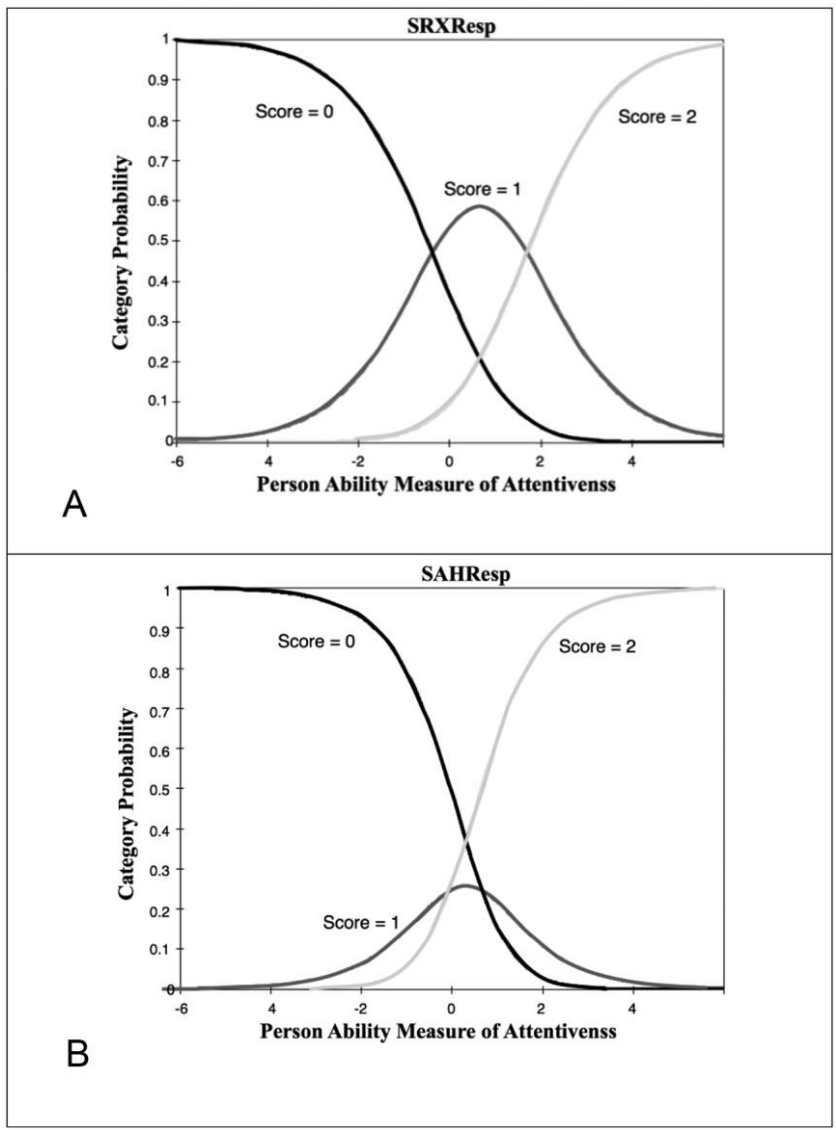


Figure 2. Category probability graph of scores of 0, 1, and 2 on two items representing ordered (*a*) and disordered (*b*) thresholds.

ments that include these items for program evaluation. Finally, we provide an example to the educational community of an argument-based approach through a claim-evidence model. Each of these areas is discussed in turn.

The Construct of Attentiveness

We have situated the construct of attentiveness within the larger body of literature addressing MKT. Being attentive to student thinking requires integrated knowledge of mathematics and students' mathematical thinking, similar to aspects of the MKT framework described by Ball et al. (2008). Specifically, attentiveness requires a specialized knowledge of mathematics that is unique to the work of teachers, who need to understand how particular complex mathematical concepts can build from a range of students' ideas. In this way, attentiveness connects to the literature on the progressive formalization of mathematical ideas (Freudenthal, 1991; Treffers, 1987). To press student thinking forward, teachers must also have the ability to recognize ideas expressed by students. This ability is related to several constructs in the research literature (e.g., MOST via Leatham et al., 2015) that address the MKT associated with recognizing opportunities so as to build on students' ideas to support the development of important mathematical ideas. Thus, the construct of attentiveness could serve to consolidate several areas of research into a single measurable construct, contributing to the research and evaluation of teacher education.

Program Evaluation and Improvement

As the teacher education field moves toward an increased focus on core practices (Ball et al., 2009; Grossman, Hammerness, et al., 2009), we need assessments designed to examine the effectiveness of such initiatives that draw on authentic examples of practice. Our four-phase MAPP item-development process, which incorporates authentic student and teacher work, is meant to create items that measure attentiveness and that can be implemented at scale for the purposes of program evaluation and comparison. By exploring the validity evidence associated with our construct, development process, and items, we go beyond presenting a new assessment by justifying claims about the construct representativeness of our items. Those claims are supported with evidence from the literature and data gathered in the MAPP item-development process.

Creating easily administered measures for program evaluation that draw from teachers' classroom experiences is arguably better for evaluating teacher education programs and initiatives than relying on large-scale content knowledge tests such as the Praxis (Hill, Umland, Litke, & Kapitula, 2012). Although measures such as observations or reflective portfolios can provide an important depth of understanding of teachers' practices, the ease of administration and analysis of our assessment items could enable programs to periodically assess their participants' attentiveness throughout the program without placing an undue burden in terms of resources.

In addition, if incorporated as part of a teacher education program evaluation framework, an assessment that specifically measures attentiveness will likely shape the content and focus of program activities. We see progressive formalization as an

important pedagogical philosophy that mathematics teacher education should be emphasizing because it aligns with constructivist notions of eliciting and building on students' thinking and emphasizes the importance of connecting to mathematical conventions. Because the attentiveness assessment is grounded in progressive formalization, creating an assessment that authentically measures attentiveness may help ensure that teacher education programs include instruction and preparation for this key aspect of teaching mathematics. We posit that easy-to-implement assessments with strong construct representation are more likely to positively drive instruction. Attentiveness is certainly not the only important aspect of teaching mathematics, and there are likely other key aspects of teaching in general not captured by existing large-scale measures, bringing us to our final reason for presenting a detailed validity argument and evidence.

Validity and Presenting Validity Evidence

Kane (2006) has called for an argument-based approach to examining the validity of proposed interpretations and uses of test scores that involves the presentation of claims and associated evidence to support the justification of such claims. Although there are few examples of this approach in the literature, Schilling and Hill (2007) provide one such example. However, if we are asking individuals, who may or may not have expertise in measurement, to evaluate the validity evidence for an assessment in relation to their particular use of that assessment, we need to consider presenting our arguments in ways that are accessible to this broader group. The more global goal of presenting our current work related to assessing attentiveness is to demonstrate a potential organizational format for presenting a validity argument and associated evidence that is accessible to the greater educational community. We have demonstrated how quantitative methods (often the primary focus of measure validation work) fit within a larger argument that provides evidence for claims about how well the construct is represented, operationalized, and structured (i.e., construct representation). By presenting such work, we aim to spur conversations in the measurement community related to the presentation of assessment validation efforts and in the larger education research community related to assessment validity.

Implications and Future Work

We plan to continue the iterative development of assessment items, centered on construct representation, by establishing connections between the theoretical model for attentiveness and the items themselves. Once we have a number of well-functioning items, we intend to undertake external validation efforts that examine the relationship between assessment scores and other variables of interest, similar to Embretson-Whitely's (1983) nomothetic span. It will also be important to examine the consequential validity of the test score uses and interpretations (Messick, 1995) by examining both the intended and unintended consequences of attempting to measure attentiveness for the purposes of program evaluation. For example, the LMT (n.d.) project website states that "LMT measures cannot be used for hiring, promotion, pay, or tenure. Our measures are not designed to make highly accurate

statements about individuals' mathematical knowledge. Instead, they can be used to compare groups of teachers' mathematical knowledge, or examine how a group of teachers' knowledge develops over time." Our assessment shares the same intended use, but it will be important to examine the consequences associated with attempting to measure attentiveness.

A possible limitation of the assessment is that certain teacher learners may be able to differentiate very accurately between given responses and statements but may struggle to find their own ways of responding when they are in front of students. We are thus particularly interested in looking at the relationship between teachers' abilities to respond to students in real time in the mathematics classroom and scores on this assessment. For example, examining the relationship between teachers' attentiveness scores to observations of their classroom practice would provide evidence of an attentiveness score's relationship to important core practices.

Future work is also needed to investigate the implicit claim that the ability to be attentive to students' thinking is domain dependent. Do teachers attend to students' thinking consistently across mathematics topics? For example, a middle school teacher may be able to successfully attend to students' ideas related to univariate data analysis and statistics but may be unable to adequately attend to an elementary student's ideas about fractions. This is an empirical question, and examination should include a focus on the level of domain specificity needed to assess attentiveness. Although others have begun to investigate different aspects of teacher cognition related to students' thinking (e.g., Jacobs et al., 2010), more work is needed regarding mapping teachers' cognition. Investigating whether attentiveness is domain dependent may further inform definitions for specialized versus common content knowledge (e.g., Speer, King, & Howell, 2015). We plan to use the MAPP item-development process in conjunction with Rasch analysis to inform our iterative process of item development and to build on the theory of what constitutes MKT (Bond & Fox, 2013).

Finally, through the MAPP item-development process, we do not mean to suggest that teaching can be boiled down to a set of selected-response items. Teachers' attentiveness is just one aspect of the complex work done in the classroom. We have chosen to focus on this construct because of its relationship to teachers' practice of building on students' thinking as a foundational aspect of classroom mathematics instruction. As a field, we have called for this type of practice for years but have failed to see large shifts in typical classroom instruction (Measures of Effective Teaching Project, 2012). We see the assessment of attentiveness as a tool to press for classroom instruction that recognizes the importance of eliciting students' ideas and then responding to students in ways that challenge students to progressively formalize or reformalize important mathematical ideas.

Appendix

Before responding to the prompt(s) below, please work through the following task given to a class of middle school students:

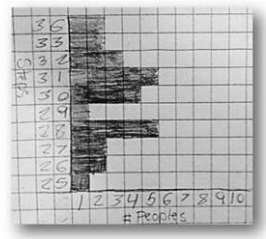
Do 6th grade students have approximately the same stride length? Ms. Peterson says, "Yes" and Mr. Garcia says, "No." Based on the data collected by the class, who do you think is more correct? Create a representation to help justify your answer.

Each student determined the number of steps it took to walk the length of the school building. The number of steps per student are presented below.

31	28	25	29
30	31	27	33
31	32	28	26
26	26	33	28
31	32	27	36
32	30	28	26
31	30	30	

1. What important mathematical idea(s) is this task targeting?

Student X's work is shown below:



2. Please describe in detail what you Student X's representation indicates to you about their mathematical understanding.

3. If you were the teacher of Student X, please describe in detail how you would respond to Student X.

Figure A1. Phase 3 free-response item

Notes

Michele B. Carney is an assistant professor of mathematics education in the Curriculum, Instruction and Foundational Studies Department at Boise State University; Laurie Cavey is an associate professor of mathematics education in the Department of Mathematics at Boise State University; Gwyneth Hughes is an outreach specialist in education outreach and partnerships at University of Wisconsin–Madison. Correspondence may be sent to Michele Carney at Boise State University, 1910 University Drive, Boise, ID 83725-1745; e-mail: michelecarney@boisestate.edu.

1. A test specifications document and expert review is essential to providing evidence of content validity for our assessments specific to particular mathematics domains (e.g., data analysis

and statistics for grades 6–12). However, the focus of this article is on construct representation as it pertains to teachers' cognitive processes.

2. The example item in the appendix does not include a prompt related to the student's approach because of the amount of overlap in responses to students' approach and understanding in previous administrations of data-analysis-related items.

3. Twenty percent of the responses were dual coded by a second researcher, with 80% agreement after the first round of analysis and 100% agreement following a brief discussion.

4. We conducted a principal component analysis of the residuals to look for secondary dimensions. The Rasch dimension explained 36.2% of the variance, with 13.4% explained by persons and 22.8% explained by items. The largest secondary dimension explained only 1.8% of the variance. The variance explained by the items is almost 13 times greater than that explained by the secondary dimension, indicating that the data structure can be considered unidimensional and meets local independence for further analysis.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D., De Jong, J., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 59–70). New York: Springer.

Ball, D. L., Sleep, L., Boerst, T. A., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *Elementary School Journal*, **109**(5), 458–474.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, **59**(5), 389–407.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, **17**(2/3), 62–87.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Hove: Psychology.

CAEP Commission on Standards and Performance Reporting. (2013). *CAEP accreditation standards and evidence: Aspirations for teacher preparation*. Washington, DC: Council for the Accreditation of Educator Preparation. Retrieved from http://caepnet.org/~/media/Files/caep/standards/commrpt.pdf?lapen

Cavey, L. O., Carney, M. B., & Hughes, G. (2017, April–May). *Understanding and responding to students' thinking: A study in measurement and theory building*. Paper presented at the American Educational Research Association annual meeting, San Antonio, TX.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, **68**(3), 397–412.

Clement, L., Chauvot, J., Philipp, R., & Ambrose, R. (2003). A method for developing rubrics for research purposes. *International Group for the Psychology of Mathematics Education*, **2**, 221–228.

Darling-Hammond, L., & Bransford, J. (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco: Wiley.

Educational Testing Service. (2016). *The Praxis tests*. Princeton, NJ: Author. Retrieved from https://www.ets.org/praxis

Embretson-Whitely, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, **93**(1), 179–197.

Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht: Reidel.

Freudenthal, H. (1991). *Revisiting mathematics education*. Dordrecht: Kluwer.

Gravemeijer, K., & van Galen, F. (2003). Facts and algorithms as products of students' own mathematical activity. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to* Principles and Standards for School Mathematics (pp. 114–122). Reston, VA: National Council of Teachers of Mathematics.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, **111**(9), 2055–2100.

Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching*, **15**(2), 273–289.

Grossman, P., Schoenfeld, A., & Lee, C. (2007). Teaching subject matter. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 201–231). San Francisco: Jossey-Bass.

Hill, H. C. (2005). Content across communities: Validating measures of elementary mathematics instruction. *Educational Policy*, **19**(3), 447–475.

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education*, **35**(5), 330–351.

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M., & Rowan, B. (2007). Validating the ecological assumption: The relationship of measure scores to classroom teaching and student learning. *Measurement*, **5**(2/3), 2–3.

Hill, H. C., Sleep, L., Lewis, J., & Ball, D. (2007). Assessing teachers' mathematical knowledge: What knowledge matters and what evidence counts. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. **1**, pp. 111–155). Charlotte, NC: Information Age.

Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education*, **118**(4), 489–519.

Jacobs, V. R., Lamb, L. L., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, **49**(2), 169–202.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, **38**(4), 319–342.

Kane, M. T. (2006). Valildation. In R. L. Brennan, National Council on Measurement in Education, & American Council on Education (Eds.), *Educational measurement* (pp. 17–64). Westport, CT: Praeger.

Lampert, M. (2009). Learning teaching in, from, and for practice: What do we mean? *Journal of Teacher Education*, **61**(1/2), 21–34. doi:10.1177/0022487109347321

Lampert, M., Franke, M. L., Kazemi, E., Ghousseini, H., Turrou, A. C., Beasley, H., . . . Crowe, K. (2013). Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, **64**(3), 226–243.

Langrall, C. W., & Mooney, E. S. (2002). *The development of a framework characterizing middle school students' statistical thinking*. Paper presented at the Sixth International Conference on Teaching Statistics (ICOTS6), Cape Town.

Leatham, K. R., Peterson, B. E., Stockero, S. L., & Van Zoest, L. R. (2015). Conceptualizing mathematically significant pedagogical opportunities to build on student thinking. *Journal for Research in Mathematics Education*, **46**(1), 88–124.

Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions*, **23**(4), 1241.

Linacre, J. M. (2016). Winsteps Rasch Measurement (Version 3.92.1) [Computer software]. Retrieved from http://www.winsteps.com

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, **36**(8), 437–448.

LMT (Learning Mathematics for Teaching). (n.d.). *Learning Mathematics for Teaching (LMT) project*. Retrieved from http://www.umich.edu/~lmtweb/

LMT (Learning Mathematics for Teaching). (2005). *Mathematical knowledge for teaching (MKT) measures*. Ann Arbor: University of Michigan.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, **3**, 635–694.

Mason, J. (2011). Noticing: Roots and branches. In M. Sherin, V. Jacobs, & R. Philipp (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 35–50). New York: Routledge.

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

McCrory, R., Floden, R., Ferrini-Mundy, J., Reckase, M. D., & Senk, S. L. (2012). Knowledge of algebra for teaching: A framework of knowledge and practices. *Journal for Research in Mathematics Education*, **43**(5), 584–615.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, **64**(5), 378–386.

Measures of Effective Teaching Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, **70**(4), 308–312.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**(9), 741–749.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement*, **10**(1/2), 1–29.

Pierson, J. L. (2008). *The relationship between patterns of classroom discourse and mathematics learning*. Ann Arbor, MI: ProQuest.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 201–226). New York: Springer.

Saderholm, J., Ronau, R., Brown, E. T., & Collins, G. (2010). Validation of the diagnostic teacher assessment of mathematics and science (DTAMS) instrument. *School Science and Mathematics*, **110**(4), 180–192. doi:10.1111/j.1949-8594.2010.00021.x

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement*, **5**(2/3), 70–80. doi:10.1080/15366360701486965

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, **57**(1), 1–22.

Speer, N. M., King, K. D., & Howell, H. (2015). Definitions of mathematical knowledge for teaching: Using these constructs in research on secondary and college mathematics teachers. *Journal of Mathematics Teacher Education*, **18**(2), 105–122.

Supovitz, J., Ebby, C. B., & Sirinides, P. (2013). *Teacher Analysis of Student Knowledge (TASK): A measure of learning trajectory-oriented formative assessment*. Retrieved from ERIC database. (ED547658)

Teuscher, D., Moore, K. C., & Carlson, M. P. (2015). Decentering: A construct to analyze and explain teacher actions as they relate to student thinking. *Journal of Mathematics Teacher Education*, **19**(5), 433–456.

Treffers, A. (1987). *Three dimensions: A model of goal and theory description in mathematics instruction—The Wiskobas project*. Dordrecht: Reidel.

Wolfe, E. W., & Smith, E. V., Jr. (2006). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, **8**(2), 204–234.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, **70**(12), 857–860.