Revisiting Purpose and Conceptualization in the Design of Assessments of Mathematics Teachers' Knowledge

Chandra Hawley Orrill
STEM Education & Teacher Development
University of Massachusetts Dartmouth
Dartmouth, MA USA
corrill@umassd.edu
ORCID: 0000-0002-3907-554X

Yasemin Copur-Gencturk
Rossier School of Education
University of Southern California
Los Angeles, CA USA
copurgen@usc.edu
ORCID: 0000-0002-4470-9615

Allan Cohen
Department of Educational Psychology
University of Georgia
Athens, GA USA
acohen@uga.edu
ORCID: 0000-0002-8776-9378

Jonathan Templin
Department of Psychological and Quantitative Foundations
University of Iowa
Iowa City, IA USA
jonathan-templin@uiowa.edu
ORCID: 0000-0001-7616-0973

Corresponding Author: Chandra Orrill, Department of STEM Education & Teacher Development, UMass Dartmouth, 285 Old Westport Road, Dartmouth, MA 02747; 774-929-3052; corrill@umassd.edu

Revisiting Purpose and Conceptualization in the Design of Assessments of Mathematics Teachers' Knowledge

Abstract

In this paper, we focus on the design of assessments of mathematics teachers' knowledge by emphasizing the importance of identifying the purpose for the assessment, defining the specific construct to be measured, and considering the affordances of particular psychometric models on the development of assessments as well as how they are able to communicate learning or understanding. We add to the literature by providing illustrations of the interactions among these critical considerations in determining what inferences can be drawn from an assessment. We illustrate how the considerations shape assessments by discussing both existing and ongoing research projects. We feature discussion of two projects on which the authors of this paper are collaborating to demonstrate the affordances of attending to all three considerations in designing assessments of mathematics teachers' knowledge to provide readers with opportunities to see those considerations in use.

Keywords: assessment design, pedagogical content knowledge, content knowledge

Orrill and Cohen (2016a) described how the *purpose* of an assessment and the *conceptualization of the domain* to be measured shape the design of assessments of mathematics teacher knowledge. This paper extends the discussion of conceptualizing the domain and identifying the purpose of the assessment as they relate to developing an assessment by adding a third consideration: how particular psychometric models help with understanding learning. We explore these three key ideas—conceptualizing the domain, identifying the purpose, and selecting psychometric models that support meaningful sensemaking of what is being measured—and their practical implications. To explore these ideas, we present pilot work in two projects, one focused on creating an assessment aligned to a particular psychometric model and the other a project focused on understanding how teachers understand the mathematics being taught.

Domain and Purpose

Summarizing the earlier paper, it was asserted that the two questions any assessment development team should tackle first are about the domain to be measured and the purpose of the assessment. When thinking about teacher knowledge, specifically, domain becomes critical because teacher knowledge is multifaceted. For example, in Ma's (1999) seminal work comparing Chinese teachers to American teachers, she focused on what she described as "knowledge of mathematics for teaching" (Ma, 1999, p. xvii). In her interviews, she asked teachers questions such as:

> Let's spend some time thinking about one particular topic that you may work with when you teach, subtraction with regrouping. Look at these questions ($\begin{array}{r} 52 \\ -\,25 \\ \hline \end{array}$, $\begin{array}{r} 91 \\ -\,79 \\ \hline \end{array}$, etc.). How would you approach these problems if you were teaching second grade? What would you

> say pupils would need to understand or be able to do before they could start learning subtraction with regrouping? (p. 1)

The answers to this question could uncover understandings that a teacher may have about how to teach subtraction with regrouping as well as the teacher's understandings of the prerequisite knowledge children may need to learn subtraction with regrouping. However, a teacher could provide a very detailed answer to this question without demonstrating the skills and knowledge to solve the task in multiple ways; for example, by using a variety of strategies or a variety of representations.

In contrast, the Integrating Mathematics and Pedagogy (IMAP) assessment was developed to "assess respondents' beliefs about mathematics, beliefs about learning or knowing mathematics, and beliefs about children's learning and doing mathematics" (IMAP Research Team, 2003, p. 4). Their items, designed for a different purpose, measure teacher knowledge in a different way. For example, in an item from IMAP also focused on subtraction with regrouping and the ways students might approach such a problem (Figure 1), the IMAP team showed examples of student thinking instead of asking about teacher moves. The IMAP team asked the teacher to evaluate that thinking rather than focusing the teacher on prerequisite knowledge. Thus, to answer the question in Figure 1, a teacher would need to be able to make sense of students' work and situate it in an imagined or known hierarchy of student development of subtraction and related number concepts. In contrast, to answer the question from Ma, the teacher would need to situate the mathematics in a larger system of mathematical ideas.

While the two tasks by Ma and IMAP arguably focus on the same mathematics (subtraction with regrouping), they test different aspects of teachers' understandings. Although an assessor might learn something about a teacher's understanding from either item, each was

developed for a particular purpose and is designed to meet that purpose. Further, neither question explicitly examines how a teacher thinks about subtraction in terms of solving these specific tasks. Thus, using it to determine whether teachers can subtract with regrouping would likely yield incomplete information. Measuring exactly what matters is critical for creating assessments. Further, we assert that measuring what matters can only happen at the intersection of defining the construct, identifying the purpose of the assessment, and designing the assessment to work with psychometric models that provide the feedback desired.

INSERT FIGURE 1

Defining the construct of interest is a key element of understanding what one wants to measure. The two tasks above work precisely as intended because their authors specifically defined their construct. For example, neither research effort was focused on how teachers use representations, therefore, neither question focused on that. Neither research effort was focused on how many ways a teacher may have to solve the task for herself, therefore, they did not focus on that. Instead, each project designed tasks to uncover the understandings that were important for their purposes. In the case of IMAP, the researchers explicitly defined their domain, as noted above, adding, "This survey is not designed to explicitly assess respondents' beliefs about teaching mathematics…" (IMAP Research Team, 2003, p. 4). Thus, the research team mapped a domain that includes certain beliefs related to teaching, but not others. Without a clear definition of the construct to be measured, it would be very easy for an assessment development team to write a variety of items that, in the end, did not measure any coherent constructs.

The more well-defined the domain, the better the information obtained from the measure. In a recent study of the LMT for Proportional Reasoning (Learning Mathematics for Teaching, 2007), the scores of participants were examined as the task was refined to fit a clearer definition

of the construct being measured (Orrill & Cohen, 2016b). A mixture Rasch model was used. This made it possible to see not only how well participants scored as compared to the original national sample (Hill, 2008), but also to place participants into groups, based on patterns in their responses, that would have otherwise remained unnoticed. In that analysis, we looked at the full assessment as well as two subsequent task sets in which tests were created by systematically removing items from the assessment to make them better fits for our definition of proportional reasoning. That definition is focused on understandings teachers should have. It is grounded in a structural understanding that $\frac{a}{b} = \frac{c}{d}$ (e.g., Lamon, 2007) and focuses on the various ways in which teachers can apply that structure to a variety of proportional situations. For example, one could reason about the multiplicative relationship between values (e.g., in a 2c sugar : 5c flour ratio, there is 2/5 as much sugar as flour, also expressed as $y = kx$). One could also reason about the scaling of the relationship, for example, they could scale both x and y by doubling it or halving it. This definition draws heavily from frameworks provided by Lobato and Ellis (2010), Pitta-Pantazi and Christou (2011), and Lamon (2007). By narrowing the item pool to fit this well-defined construct, we not only saw considerable increases in mean scores (0.5 standard deviations or more in each latent class), but also were able to determine the reasons for class membership change. That is, in the latent classes identified by the psychometric model based on teachers' response patterns, we saw a shift in the underlying reasons for class membership. In the full assessment (73 items), class membership was explained by participants' facility with algorithmic and symbol-based algebraic approaches to proportions. Thus, the higher-scoring class was also the class whose members were better at recognizing, making sense of, and applying algorithms in a variety of situations as well as using algebra symbols. The lower-scoring group, however, was better at identifying whether a situation was actually proportional.

In contrast, the narrowest task set, which was also the one most closely aligned to our definition of proportional reasoning for teaching (i.e., 54 items from the original pool of 73 items) highlighted important differences in the ways teachers reasoned about the items. The lower-scoring class remained better at identifying proportional situations in this narrower problem set. We also saw differences, however, in the ways the teachers reasoned about proportions. Specifically, the lower-scoring class was better at reasoning with scaling than reasoning multiplicatively. We assert this was because they were more adept at reasoning about within measure-space situations (e.g., in a proportion of red paint to yellow paint to make a specific orange paint, the relationship of yellow paint to yellow paint is the within measure-space relationship) than reasoning about between measure-space situations (between measure-space relationships would compare yellow paint to red paint rather than yellow paint to yellow paint). Consistent with this observation, the higher-scoring class was more confident with between measure-space relationships, which require multiplicative reasoning, as well as combining ratios and making sense of ratio tables. Interestingly, the lower-scoring class seemed to find tasks grounded in the work of teaching (e.g., making sense of a students' thinking or identifying which numbers would be harder for students to work with) to be easier than their higher-scoring counterparts. By understanding these differences, which are clearly tied to our definition of proportions (our construct), we are better able to understand how these teachers understand the content and what kinds of understandings they may need to more thoroughly develop.

How Psychometric Models Help Us See Learning

In addition to the need to focus on purpose and the domain of interest, a third major factor should be considered: the psychometric model that is best suited for the assessment. In other words, the psychometric models that help illuminate learning. In our current efforts and the

work we have done previously, our team has developed a clear sense that different models afford different opportunities. While psychometric models should not limit assessment efforts, knowing how different models can help us 'see' learning can support assessment developers to better think about what is possible to measure, what kinds of questions can be asked, and what kinds of results can be reported out. Further, knowledge of which models to be used helps in building an assessment that is well-aligned with the model, providing enough items and tasks to ensure model results have a high degree of reliability.

As one example of how psychometric models shape the assessment and the information it can provide, consider the currently-popular item response theory (IRT) models. These models typically rely on dichotomous data that are analyzed to create a continuum of performance and a continuum of the relative difficulty of questions. The models provide information about how participants scored relative to other people in the sample as well as how difficult each item is for the participants in the sample. These models are both popular and practical because they provide a lot of information in understandable ways. These models are an important part of current efforts to develop learning trajectories (e.g., Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Confrey, Gianopulos, McGown, Shah, & Belcher, 2017).

As mathematics educators and psychometricians who are interested in supporting teachers' learning, we would argue that the usable data for assessing learning or understanding from unidimensional IRT analysis is limited. After all, it is not designed to provide fine-grained guidance about what the teachers may or may not have known within the larger domain. Moreover, scores from IRT are normative, not criterion-based. That is, we can treat 0 as being the mean score, but that is based on actual respondent's performance on the assessment and not an externally identified amount of knowledge an expert might demonstrate or that one might

expect of a particular participant. If we were to write assessments specifically for IRT, we would need to consider what data can be provided and develop items accordingly.

Fortunately, there are emerging psychometric models that provide other kinds of information. Specific to the efforts we discuss in this paper are diagnostic classification models (DCMs; e.g. Rupp, Templin & Henson, 2010) and Topic Analysis Models (aka Topic Models; e.g., Blei, 2012). DCMs are designed to enable researchers to identify and estimate a multidimensional set of conceptually linked ideas (called attributes in the DCM literature) which an assessment is developed to measure. In the case of our work, the domain of the assessment is teachers' knowledge of proportional reasoning. One measurable attribute of that knowledge is 'appropriateness' (e.g., Weiland, Orrill, Brown, & Nagar, 2019, accepted), which focuses on whether a participant is able to differentiate situations that can be reasoned about proportionally from those that cannot be reasoned about proportionally. Rather than providing assessors with continuous scores for each participant, DCMs provide a probability that a given participant has mastered each attribute. In trading continuous scores for categorical attributes, DCMs provide information about a profile of competencies of participants with a reliability higher than what is achieved with IRT models (Templin & Bradshaw, 2013). We assert that this kind of finer-grained information may be helpful to professional developers as they develop more tailored instruction for teachers.

Topic models (Blei, 2012) do not return a score at all. Instead, these models consider all of the words in a corpus of documents to determine latent themes (called topics) that are present in the documents. These models have been applied to open-ended assessments to analyze the themes students use in constructing their answers to test questions (Kim, Kwok, Cardozo-Gaibisso, Buxton & Cohen, 2017). Once these topics are identified, a researcher or professional

developer can look for changes in frequencies of the use of topics across a group of teachers as a means for measuring growth. As an illustration, in one recent application of topic models, students' written responses to a series of science prompts were considered across a unit of instruction (Kim et al., 2017). In this study, three main themes appeared, the first involved the use of middle-school appropriate technical terms (e.g., change, variable, independent, cause, dependent, effect, etc.), discipline-specific terms (e.g., energy, increase, population, decrease, different, amount, kinetic), and everyday language (e.g., put, stronger, big, think, little, bigger, etc.). Across four measurements, middle school students in the study sample, who were all English Language Learners, shifted as a result of an instructional intervention. They moved from using language in the everyday language topic to predominant use of words in the discipline-specific language and academic language topics in their responses. That is, this use of topic models reflected the instruction students received in learning about and explaining the process of scientific inquiry. In this way, topic models offer an interesting and innovative way to look at learning as reflecting the language one uses to communicate about an issue or problem. We assert that a professional developer who has information about how teachers are using language and the ways that might intersect with how they understand their content would have a powerful tool for supporting teachers' learning.

Applying Purpose, Construct, and Seeing Learning to Two Projects

Two Projects – Two Purposes

To illustrate our three key ideas in action, we turn to two projects on which this team of authors is currently working. We describe the constructs on which we are focusing in these projects. We then explore how each of the assessments helps us understand what teachers are learning as a product of the interaction between the purpose for which it was written, the

conceptualization of the domain and the psychometric models selected. Because these projects are both still in the development phase, we report on pilot work and work that is related, but outside the scope of these projects, as appropriate.

The first project is a professional development project, 'Advancing Understanding', funded by the Institute of Education Sciences (Copur-Gencturk, Nye, Orrill, & Cohen, 2018). For this project, we are developing an online, personalized professional development system focused on developing middle school teachers' understandings of proportions and how to teach proportion concepts identified in the Common Core for State School Mathematics (Common Core State Initiative, 2010) to middle grades students. The underlying goal of this project is to provide just-in-time professional development that is available anytime and anywhere. In the United States (our context), there are numerous reasons why teachers may find themselves in need of online PD. For example, teachers in rural districts may only have access to online PD. Similarly, teachers who are struggling or who need a refresher after being away from the mathematics for a period of time may need something right away to support their instruction for the near future.

The Advancing Understanding online course includes two modules—one for content knowledge (CK) development and the other for pedagogical content knowledge (PCK) development (Shulman, 1986). The module for content knowledge (CK) development is comprised of six submodules: what is ratio?; solving ratio problems with different representations; fraction and ratio relationship; understanding proportionality; geometric similarity and proportionality; and putting it all together. The second module focuses on pedagogical content knowledge (PCK) and includes five submodules (what is the plan-implement-reflect approach to teaching; planning; implementing; assessing and reflecting; and

putting it all together). To ensure that teachers' time is well-spent in this PD, each submodule starts with an assessment to determine whether the teacher has already mastered the content of that submodule. Passing the assessment allows the teacher to skip the submodule, otherwise, she is enrolled in the self-paced submodule. At the end of the submodule, the teacher again completes an assessment. If additional work is needed in the content of the submodule, the teacher returns to that content. However, if the teacher has mastered the content, she moves on to the next submodule. Thus, the purpose of the assessment for this project is diagnostic in that it places teachers into the submodules most relevant for them and determines the extent to which they have successfully mastered the content. To this end, the assessment items are closely aligned to the content of the modules, measure specific understandings about content or about teaching that content, and show whether the teacher has mastered the specific content of interest.

The second project, 'Usable Measures', is an assessment development project funded by the National Science Foundation (Copur-Gencturk, Cohen, Templin, & Orrill, 2018). Usable measures focuses on the development of a measure of teachers' knowledge of CK and PCK before and after any Common Core-aligned professional development focused on proportions. Unlike the previous project, in this project there is not a single intervention of focus, instead, we aim to measure, more broadly, the domain of proportional reasoning as it is defined by the Common Core (Common Core State Initiative, 2010) and as professional developers teach it. Here, we have chosen to use emerging psychometric models to provide useful feedback to professional developers both before their teaching begins and after they have completed work with a group of teachers. Assessment results are designed to focus the instructional experience on topics with which a group of teachers may need additional support. The same models also provide feedback on growth through the professional development experience. Thus, the purpose

of this assessment is to provide actionable feedback for designing instruction as well as to provide information about changes in teachers' understanding over time across a set of proportional reasoning concepts likely to be the focus of any given proportional reasoning professional development experience in the United States. In this way, we see the assessment as providing information to improve instruction while acknowledging the political reality that funding agencies and other stakeholders need evidence of the value of the program as shown in teachers' growth during participation.

As one might imagine, these two purposes, while related, lead to some important differences in the assessments. The purposes also led to differences in the kinds of psychometric models selected for the assessments because of the differences in the nature of the information needed between the projects.

Defining the Construct

For both of these projects, we focus on two constructs: content knowledge (CK) and pedagogical content knowledge (PCK; e.g., Shulman, 1986). Working from Shulman's initial definitions, we view CK as the participants' understanding of the mathematics and PCK as the knowledge of how to teach the content knowledge and how to assess learning. We have chosen these constructs because a growing body of literature suggests these are measurable constructs within the domain of teacher knowledge (e.g., Baumert et al., 2010) and that participants have been found to show growth in these constructs in professional development settings (e.g., Carney, Brendefur, Thiede, & Sutton, 2016; Copur-Gencturk & Lubienski, 2013; Copur-Gencturk, Plowman & Bai, 2019). Further, there is growing evidence that professional development focused on both CK and PCK is more effective than professional development

focused only on one or the other (e.g., Kennedy, 1998; Scher & O'Reilly, 2009). Thus, considering both and expecting professional development to attend to both is reasonable.

We work from the conception of CK outlined in the four strands of mathematical proficiency as defined in *Adding it Up* (National Research Council (NRC), 2001). Specifically, we take the position that CK relates to being able to use adaptive reasoning, strategic competence, conceptual understanding, and procedural understanding to reason about proportional situations. These strands include procedural and conceptual knowledge (e.g., Hiebert & Lefevre, 1986) as well as problem-solving and adaptive reasoning skills relevant to proportional situations (NRC, 2001). CK can be assessed using traditional mathematics problems, but it can also be measured by applying a solution strategy to another, similar, task (i.e., adaptive reasoning). The CK to be taught and measured in the Advancing Understanding project is shown in Table 1.

INSERT TABLE 1 ABOUT HERE

In the same way we must limit the content domain being measured, it is necessary to tightly define the aspects of PCK to which we intend to attend. For the purposes of our projects, we look at PCK in a matrix that includes phases of instruction (planning, implementation, and reflection) and approaches for structuring classroom discussion in ways that promote mathematical growth. Specifically, we draw from the *5 Practices for Orchestrating Productive Mathematics Discussions* (Smith & Stein, 2011). We have chosen to operationalize PCK in this way for two main reasons. First, the kinds of moves outlined in the *5 Practices* are well-known and considered to have a positive impact on the instructional environment because they provide clear guidance to teachers about how to support student learning in discussion-based mathematics classrooms (Moschkovich, 2013; National Council of Teachers of Mathematics,

2014; Stein, Engle, Smith, & Hughes, 2008). Second, the aspects of PCK shown in Table 2 are measurable using a paper-and-pencil assessment. Because we are limited in our format (e.g., we cannot conduct performance assessments or think-alouds), we had to identify a body of practices that benefit from the synergistic interplay of content and pedagogy and that are observable in questions that we can ask with the kinds of assessments we are developing. Then, we craft questions from each of the phases of instruction that are focused on particular PCK competencies to ensure that teachers have a holistic understanding of each of those competencies.

We conceptualize the difference between CK and PCK in terms of being able to perform pure mathematical tasks and being able to identify qualitative differences in students' thinking and make instructional decisions based on a given situation. If the intent of a task is to determine whether the respondent can apply a certain line of reasoning to solving the task, we consider the task to be measuring CK. However, if we are drawing on patterns of thinking to make instructional decisions or to assess the quality of a student's thinking, the task is measuring PCK. The first example is an application of understanding of mathematics (CK), whereas the latter is a combination of understanding mathematics and making instructional decisions about a student's thinking (PCK).

INSERT TABLE 2 ABOUT HERE

Seeing Learning – The Power of Psychometric Models

Elsewhere, there are discussions of the different ways in which psychometric models allow us to understand what a person knows and the extent to which there is change in what they know (e.g., Izsák & Templin, 2016; Orrill & Cohen, 2016b). For example, as noted above, IRT models (Hambleton & Swaminathan, 1985, 2013; Lord, 1980; Lord & Novick, 1968, 2008) can be used to examine how a given teacher compares to a larger group of teachers based on the

number and difficulties of items that were answered correctly. With these models, a change of 0.3 standard deviations is typically considered to be statistically significant growth. Thus, learning is operationalized as the learner answering more questions correctly.

Mixture Rasch models (Rost, 1990) separate the participants into latent groups based on patterns of responses. In our work on mixture Rasch findings (Izsák, Orrill, Cohen, & Brown, 2010; Orrill & Cohen, 2016b), we have found that these groups (also known as latent classes) can be analyzed to determine reasoning strategies used by members of the class. Thus, learning can be seen both in the change in scores and in potential movement between classes. Because the classes are determined by patterns in participants' answer choices without regard to levels of performance (scores), movement between classes is independent of the absolute score on the assessment.

In our current work, we are exploring new ways to see learning. In Advancing Knowledge, we have adopted the technique of relying on a Q-Matrix (e.g., Tatsuoka et al., 2016) to map the items to the relevant content to be certain there are adequate items in the pool to ensure that teachers have an opportunity to demonstrate their mastery of the content of each submodule. Our purpose in using the assessment is to place teachers into appropriate submodules to make their professional learning experiences as efficient as possible. Our constructs of interest are CK and PCK as defined in Tables 1 and 2 above. Learning in the professional development environment is demonstrated by teachers achieving a predefined mastery score on content for which they had previously been unable to demonstrate mastery.

For the Usable Measures project, we are applying two models: a psychometric model and a statistical model of text analysis. The first model is a DCM and the second is Topic Modeling. Using a Q-Matrix, consistent with all DCM efforts, we define how each item maps onto each

construct of interest to ensure there are adequate items to measure all of the attributes in which we are interested. For example, we are using three key attributes for CK: covariation and invariance; fraction and ratio connection; and appropriate use of proportional reasoning. All of our items are categorized into one or more of these categories. This approach is similar to that used by Bradshaw, Izsák, Templin, & Jacobson (2014). They found that using this approach differentiated teachers in fine-grained ways. They offer as an example, the mastery probabilities for three teachers who each scored 11 out of 27 on their fractions assessment. While these teachers would receive the same score using many psychometric models, DCMs allowed a different lens for highlighting the differences and similarities in those teachers' understandings. In their paper, Bradshaw et al. focused on four main attributes related to reasoning with fractions: referent units, partitioning and iterating, multiplicative comparison, and appropriateness. Their findings highlighted the value of using DCMs when they presented three profiles for teachers who had each answered 11 of 27 items correctly. Because DCMs report on the probability of mastery of each attribute, it was possible to gain considerable information about what differentiated the teachers. Teacher A was particularly unlikely to have mastered referent unit and partitioning and iterating, but had a strong probability of having mastered appropriateness and multiplicative comparison. Like Teacher A, Teacher B's probability for referent unit was very low, but her scores for appropriateness and multiplicative comparison were also very low, while her probability of mastery for partitioning and iterating was quite strong. The third teacher, also featured a relatively low likelihood of having mastered referent units and appropriateness, but was highly likely to have mastered partitioning and iterating and multiplicative comparison. Given these data, a professional developer would be able to see a trend across all three teachers that focusing on referent unit would be a good place to focus

Use of Topic Models enables us to incorporate open-ended responses into our assessments. Reports of teacher knowledge include the probability that a teacher has mastered each attribute as well as information about their frequency of use of language from each identified latent topic. In a test-retest situation, this allows us to see learning as growth in one or more attributes (e.g., probability of mastery is increased). It also allows us to consider learning as moving from using less precise to more precise mathematical language in item responses. Thus, allowing many more data points to be provided about any given participant's strengths and weaknesses.

Specifically, to capture different levels of teachers' proportional reasoning, we presented participants with the dimensions of four rectangles, all of which had a three-unit difference between the length and width. We then asked them to identify and explain which one of the rectangles looked more like a square. This question enabled us to investigate the extent to which teachers noticed the proportional relationship between the length and width of a rectangle rather than the constant difference between the length and width. Using data collected from 246 teachers in Grades 3–7 who were participating in a different project, we conducted two separate analyses: assigning scores to teachers' responses and using topic modeling to determine the underlying themes in teachers' responses.

The results from topic modeling indicated that a four-topic model best fit to the data. We interpreted each topic by examining the most frequent words that occur for that topic and by analyzing the kinds of reasoning used by teachers who made the most use of the topic. The most commonly used words in the first topic were *all*, *three*, *difference*, *length*, and *width*, and the responses that used this category the most included correct and incorrect responses focusing on the three-unit difference. The second category included words such as *percent*, *side*, *proportion*,

larger, *dimension*, and *closer*, and the responses that used this category the most included the percentage of sides and focused on the dimensions of the given rectangles. The third topic category included words such as *ratio*, *side*, *closer*, *one*, *most*, and *square*, and the responses that were best fit for this category the most clearly indicated that the ratio of sides for a square would be 1; therefore, when the ratio of sides became closer to 1, it would look more like a square. The final category included words such as *look*, *answer*, *think*, *not*, and *square*, and the responses that used this category the most gave incorrect responses for a variety of reasons.

As shown in Table 3, the strongest correlation was between Topics 1 and 3 ($r = -.60$), suggesting that the proportion of use of Topic 1 tended to decrease as the proportion of use of Topic 3 increased. Recall that responses in Topic 3 seemed to focus more on how close the ratio of sides for the given rectangles was to a square, which would be 1, whereas those in Topic 1 seemed to focus more on the differences between the sides; therefore, it is not surprising that these two topics were negatively related That is, the proportion of use of Topic 1 tended to decrease as the proportion of use of Topic 3 increased. Topic 3 also seemed to be low but negatively correlated with Topic 2 ($r = -0.28$) and with Topic 4 ($r = -0.25$), indicating that Topic 3 might be drawing on different interpretations of the proportional nature of the situation than the other two topics.

In addition to conducting topic modeling, as mentioned, we scored teachers' responses based on whether they correctly identified which rectangle looked more like a square (1 = yes; 0 = no) and the extent to which they provided a correct explanation (0 = incorrect explanation, such as "1 foot by 4 feet; both numbers are even"; 1 = partially correct explanation, such as "They are all three feet away from being square. I think that would look the most square due to the size"; 2 = correct explanation, such as "D–A square's side lengths need to be a ratio of 1 to 1,

and 27 to 30 is closer to 1 to 1 than the others"). As shown in Table 4, the explanation scores and correct answer scores were highly correlated (r = .77). More interesting was the fact that teachers who gave correct responses and who used correct explanations also tended to use Topic 3. This result suggests that the use of Topic 3 may indicate an important kind of reasoning related to correctness and level of reasoning.

Both Topic 1 and Topic 4 were moderately and negatively related to the correct answer and explanation scores, suggesting that the use of these two topics tended to be associated with incorrect answers and inaccurate explanations. It is interesting that the correlation between Topic 2 and the correct answer and explanation variables was essentially zero, meaning that the use of Topic 2 seemed to be unrelated to the selection of the correct answer or the correct explanation.

Conclusion

This paper extends our previous work focused on the importance of identifying the purpose of an assessment and defining the constructs to be measured by including a discussion of psychometric models and the ways they work in concert with the purpose and construct to allow us to 'see' learning. We then illustrated these three themes as they exist in our current assessment development efforts. The contribution we aim to make to the literature is in providing illustrations of the power of selecting various psychometric models, paired with having a clear purpose and construct, in helping make sense of learning and understanding. While there are many efforts highlighting various assessments, we seek to attend to how the design of the assessment impacts what researchers are able to learn from that assessment.

Too often, in the field of teacher assessment, assessments are used without knowing answers to questions about the construct or the purpose of the assessment. Then, instructional decisions are made based on data gathered from an instrument that may not be a good fit for the

purpose. Similarly, the instrument may provide limited guidance for shaping instruction. Thus, understanding the construct and purpose are not only relevant needs for assessment developers, but also users. Similarly, understanding the affordances of a variety of psychometric models can help the field move forward in developing assessments that both support meaningful instruction and are sensitive enough to show growth.

In conclusion, we want to address the complexity of identifying the construct when working with a multifaceted construct such as teacher knowledge. While we present our particular constructs here, we acknowledge that there are multiple ways to conceptualize the knowledge teachers need to use in the classroom to support student learning. Further, we acknowledge that the way in which this knowledge is conceptualized has serious implications for assessments developed to measure it. Consider, for example, the work of Kersting and her colleagues (Kersting, Givvin, Sotelo, & Stigler, 2010; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012). They have conceptualized their construct, which is focused on usable knowledge for teaching (Kersting et al, 2010), to only include knowledge as it is implemented. This is grounded in their assertion that some teacher knowledge may be inert except in the act of teaching. Thus, their assessment engages teachers in making decisions about students' understandings based on video clips, as that is as close to practice as they could get in a testing situation. In this way, the construct drives not just the questions in the assessment, but the entire format of the assessment.

In the field of teacher assessment there are numerous constructs that attempt to capture the knowledge teachers need - ranging from Shulman's seminal work that yielded PCK and CK (Shulman, 1986), to Ball and colleagues' mathematical knowledge for teaching (e.g., Ball, Thames, & Phelps, 2008), to the knowledge quartet (e.g., Rowland, 2013). To move forward

seriously in characterizing the knowledge teachers need to be successful at supporting students' learning, the field needs to take seriously the work of measuring the constructs of teacher knowledge precisely and intentionally.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Acknowledgement

The work reported here was supported in part by the National Science Foundation under grants DRL-1813760, DRL-1751309, and DRL-1054170 as well as the Institute of Education Studies grant number R305A180392. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

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. doi: 10.1177/0022487108324554

Baumert, J., Kunter, M. Blum, W., Brunner, M., Voss, T., Jordan, A., …Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133-180. doi: 10.3102/0002831209345157

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Bradshaw, L., Izsák, A, & Templin, J., Jacobson, E. (2014). Diagnosing teachers' understanding of rational number: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33 (1)*, 2-14.

Carney, M. B., Brendefur, J. L., Thiede, K., Hughes, G., & Sutton, J. (2016). Statewide mathematics professional development: Teacher knowledge, self-efficacy, and beliefs. *Educational Policy*, *30*(4), 539–572. doi: 10.1177/0895904814550075

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education, 42*(2), 127-166. doi: 10.5951/jresematheduc.42.2.0127

Common Core State Standards Initiative (2010). *Common core state standards*. Retrieved from http://www.corestandards.org/the-standards

Confrey, J., Gianopulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments

designed about mathematics learning trajectories. *ZDM, 49*(5), 717-734. doi: 10.1007/s11858-017-0869-1

Copur-Gencturk, Y., Cohen, A., Templin, J., & Orrill, C. (2018). Usable measures of teacher understanding: Exploring diagnostic models and topic analysis as tools for assessing proportional reasoning for teaching. Grant funded by the National Science Foundation (Award No. DRL-1813760)

Copur-Gencturk, Y., & Lubienski, S. T. (2013). Measuring mathematical knowledge for teaching: A longitudinal study using two measures. *Journal of Mathematics Teacher Education, 16*(3), 211-236. doi: 10.1007/s10857-012-9233-0

Copur-Gencturk, Y., Nye, B., Orrill, C., & Cohen, A. (2018). Advancing middle school teachers' understanding of proportional reasoning for teaching. Grant funded by the Institute of Educational Sciences. (Award No. R305A180392)

Copur-Gencturk, Y., Plowman, D., & Bai, H. (2019). Mathematics teachers' learning: Identifying key learning opportunities linked to teachers' knowledge growth. *American Educational Research Journal*,.doi: 0002831218820033.

Hambleton, R.K., and Swaminathan, H. (1985, 2013). *Item response theory: Principles and applications*. Springer Science + Business Media LLC.

Hiebert, J. & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hill, H. C. (2008). *Technical report on 2007 proportional reasoning pilot Mathematical Knowledge for Teaching (MKT) measures learning mathematics for teaching*. Ann Arbor, MI: University of Michigan.

Integrating Mathematics and Pedagogy (IMAP) Research Team (2003). *IMAP web-based beliefs survey manual*. Available http://www.sci.sdsu.edu/CRMSE/IMAP/Beliefs-Survey_Manual.pdf

Izsák, A., Orrill, C. H., Cohen, A., & Brown, R. E. (2010). Measuring middle grades teachers' understanding of rational numbers with the mixture Rasch model. *Elementary School Journal, 110*(3), 279-300.

Izsák, A., & Templin, J. (2016). Coordinating descriptions of mathematical knowledge and psychometric models: Opportunities and challenges. In A. Izsák, J. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 5-30). Journal of Research in Mathematics Education Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Kennedy, M. M. (1998). Education reform and subject matter knowledge. *Journal of Research in Science Teaching*, *35*(3), 249–263.

Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Using video to predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education, 61*(1-2), 172-181.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*(3), 568-589. doi: 10.3102/0002831212437853

Kim, S., Kwak, M., Cardozo-Gaibisso, L.A., Buxton, C.A., & Cohen, A.S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics, 1*(1), 82-102.

Lamon, S. J. (2007). Rational numbers and proportional reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 629–667). Charlotte, NC: Information Age Press.

Learning Mathematics for Teaching (2007). *Survey of teachers of mathematics: Form LMT PR-2007.* Ann Arbor, MI: University of Michigan.

Lobato, J., & Ellis, A. B. (2010). *Essential understandings: Ratios, proportions, and proportional reasoning.* In R. M. Zbieck (Series Ed.), *Essential understandings.* Reston, VA: National Council of Teachers of Mathematics.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum Press.

Lord, F.M., & Novick, M.R. (1968, 2008). *Statistical theories of mental test scores*. NY: Academic Press.

Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States.* Mahwah, NJ: Lawrence Erlbaum Associates.

Moschkovich, J. (2013). Principles and guidelines for equitable mathematics teaching practices and materials for English language learners. *Journal of Urban Mathematics Education, 6*(1), 45-57. http://ed-osprey.gsu.edu/ojs/index.php/JUME/article/viewFile/204/135

National Council of Teachers of Mathematics (2014). *Principles to actions: Ensuring mathematical success for all.* Reston, VA: Author.

National Research Council (2001). *Adding it up: Helping children learn mathematics*. National Academies Press. Washington, DC: National Academy Press.

Orrill, C. H., & Cohen, A. (2016a). Purpose and conceptualization: Examining assessment development questions through analysis of measures of teacher knowledge. In Izsák, A., Remillard, J. T., & Templin, J. (Eds)., *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 139-153). Journal of Research in Mathematics Education Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Orrill, C. H., & Cohen, A. S. (2016b). Why defining the construct matters: An examination of teacher knowledge using different lenses on one assessment. *The Mathematics Enthusiast, 13*(1 & 2), 93-110.

Pitta-Pantazi, D., & Christou, C. (2011). The structure of prospective kindergarten teachers' proportional reasoning. *Journal of Mathematics Teacher Education, 14*(2), 149–169.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rowland, T. (2013). The knowledge quartet: The genesis and application of a framework for analysing mathematics teaching and deepening teachers' mathematics knowledge. *Journal of Education, 1*(3), 15-43.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. New York: Guilford.

Scher, L., & O'Reilly, F. (2009). Professional development for K-12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness, 2*(3), 209-249. doi: 10.1080/19345740802641527

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Smith, M. S., & Stein, M. K. (2011). *5 practices for orchestrating productive mathematics discussions.* Reston, VA: National Council of Teachers of Mathematics.

Stein, M. K., Engle, R., Smith, M.S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. *Mathematical Thinking and Learning, 10*(4), 313-340. doi: 10.1080/10986060802229675

Tatsuoka, C., Clements, D. H., Sarama, J., Izsák, A., Orrill, C. H., de la Torre, J., Tatsuoka, K. K., Khasanova, E. (2016). Developing workable attrtibutes for psychometric models based on the Q-matrix. In A. Izsák, J. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 73-96). Journal of Research in Mathematics Education Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. Journal of Classification, 30, 251-275.

Weiland, T., Orrill, C. H., Brown, R. E., & Nagar, G. G. (2019). Mathematics teachers' ability to identify situations appropriate for proportional reasoning, *Research in Mathematics Education*. doi: 10.1080/14794802.2019.1579668

Weiland, T., Orrill, C. H., Nagar, G. G., Brown, R. E., & Burke, J. (accepted). A framework for a robust understanding of proportional reasoning for teaching. *Journal of Mathematics Teacher Education*.

Table 1. CK submodules and mathematical ideas being measured in the Advancing Understanding project

Submodule	**Mathematical Ideas**
What is a Ratio	- Coordinating two quantities - Comparing quantities multiplicatively - Determining when a relationship is proportional
Solving Ratios Using Different Representations	- Highlighting the ratio of two quantities remains constant as the corresponding values change - Understanding how to move between representations - Attending to the emerging third quantity (e.g., miles-to-hours yields "speed")
Fraction & Ratio Relationship	- Understanding that equivalent ratios and equivalent fractions can be created in similar ways - Combining ratios is not the same as adding fractions
Rates & Proportions	- Seeing proportions as infinitely many equivalent relationships of equivalent ratios - Recognizing the constant relationship between two ratios that yields the third quantity (e.g., speed)
Similarity & Proportion	- Understanding that similar shapes have the same ratio of lengths for corresponding sides - Understanding that volume and area of similar shapes are not proportional
Putting It All Together	- Using mathematical structures to determine whether a situation is proportional - Solving multistep ratio problems

Table 2. Framework for Pedagogical Content Knowledge adopted by both projects

Plan	Select an activity that promotes conceptual understanding and aligns to learning goals
	Identify key mathematical ideas targeted in an activity
	Know common solution strategies for a variety of proportional reasoning tasks
	Understand strengths and weaknesses of representations, how they are related, and how to use them to support learning
Implement	Make sense of students' work including identifying conceptual errors
	Know mathematics to highlight of focus on in students' work
	Select and order students' work based on sophistication of strategies
	Explain and use examples, models, representations, and arguments to support students' sensemaking
	Know how to connect students learning of proportions to prior fraction knowledge and upcoming linear function knowledge
Assess/ Reflect	Assess instructional strategies and representations to identify strengths and weaknesses of those strategies
	Use formative assessment approaches to determine whether student learning has occurred
	Use formative assessment data to make decisions about assessment items and further instructional items to promote students' learning

Table 3. Correlations among the four topic-modeling categories

	Topic 1	Topic 2	Topic 3	Topic 4
Topic 1	1			
Topic 2	–0.345***	1		
Topic 3	–0.604***	–0.275***	1	
Topic 4	–0.303***	–0.122	–0.254***	1

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 4. Correlations between the topics and the correctness of the answer and explanation

	Correct answer	Correct explanation
Correct answer	1	
Correct explanation	0.777***	1
Topic 1	–0.220***	–0.301***
Topic 2	0.034	–0.047
Topic 3	0.372***	0.537***
Topic 4	–0.248***	–0.283***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Figure 1. Example of subtraction item from IMAP Beliefs Survey

4. Here are two approaches that children used to solve the problem 635 – 482.

Lexi	Ariana
$^{5}\cancel{6}^{1}3\ 5$ $-\ 4\ 8\ 2$ $1\ 5\ 3$ Lexi says, "First I subtracted 2 from 5 and got 3. Then I couldn't subtract 8 from 3, so I borrowed. I crossed out the 6, wrote a 5, then put a 1 next to the 3. Now it's 13 minus 8 is 5. And then 5 minus 4 is 1, so my answer is 153."	635 – 400 = 235 235 – 30 = 205 205 – 50 = 155 155 – 2 = 153 482 Ariana says, "First I subtracted 400 and got 235. Then I subtracted 30 and got 205, and I subtracted 50 more and got 155. I needed to subtract 2 more and ended up with 153."
4.1 Does Lexi's reasoning make sense to you? ◯ Yes ◯ No	**4.2 Does Ariana's reasoning make sense to you?** ◯ Yes ◯ No

4.3 Which child (Lexi or Ariana) shows the greater mathematical understanding? Why?
4.4 Describe how Lexi would solve this item: 700 – 573.
4.5 Describe how Ariana would solve this item: 700 – 573.