The Mathematics Enthusiast

Volume 13
Number 1 *Numbers 1 & 2*

Article 7

2-2016

Why Defining the Construct Matters: An Examination of Teacher Knowledge Using Different Lenses on One Assessment

Chandra H. Orrill

Allan S. Cohen

Why Defining the Construct Matters: An Examination of Teacher Knowledge Using Different Lenses on One Assessment

Chandra Hawley Orrill (corrill@umassd.edu)
University of Massachusetts Dartmouth

Allan S. Cohen (acohen@uga.edu)
University of Georgia

Abstract: What does it mean to align an assessment to the domain of interest? In this paper, we analyze teachers' performance on the Learning Mathematics for Teaching assessment of Proportional Reasoning. Using a mixture Rasch model, we analyze their performance on the entire assessment, then on two different subsets of items from the original assessment. We consider the affordances of different conceptualizations of the domain and consider the implications of the domain definition on the claims we can make about teacher performance. We use a single assessment to illustrate the differences in results that can arise based on the ways in which the domain of interest is conceptualized. Suggestions for test development are provided.

Keywords: assessment development, teacher knowledge assessment

Introduction

While test performance is generally reported as if the score assigned a participant were the goal of the assessment, the actual interest is the inferences that can be made about a learner based on that score. It is critical to ensure an assessment is measuring what it is intended to measure if such inferences are to be accurate. Thus, assessments must be written in a way that allows accurate inferences to be drawn. If we are to make claims about quantities of or changes in participants' knowledge, alignment between the content and the underlying assumptions of the domain is critical. Further, if instruction is to be impacted in positive ways by assessment data, we need to ensure that scores accurately report knowledge of the intended construct. Thus, defining the construct one is interested in measuring is vital to the assessment process.

One particularly complex domain from a measurement perspective is that of teacher knowledge. This is complex because teacher knowledge is multidimensional. The specialized knowledge teachers need for teaching (SKT) necessarily includes content knowledge, pedagogical knowledge, and understandings of how students learn (e.g., Ball, Thames, & Phelps, 2008; Baumert et al., 2010; Manizade & Mason, 2011; Shulman, 1986; Silverman & Thompson, 2008). Measuring such knowledge requires adherence to a set of beliefs about the specific construct and how it is best tested, for example, using a paper-based assessment or using feedback in a video-based open-response system (e.g.,

The Mathematics Enthusiast, **ISSN 1551-3440, vol. 13, no. 1&2**, pp. 93–110

Kersting, Givvin, Sotelo, & Stigler, 2010). Despite the challenges of this complex knowledge domain, if we care about whether a teacher has the knowledge necessary to support student learning, assessments need to be written to address the domain.

In the case of SKT, we are faced with not only the complexity of the domain, but also the ambiguity of what it means in practical terms for a teacher to have or exhibit particular kinds of knowledge. In our work on proportional reasoning, we have chosen to focus on how teachers understand the mathematics of proportions rather than on their pedagogical understandings related to teaching such content. However, we also acknowledge that teachers need to be able to use the content in the process of teaching, thus our interest in assessment focuses on the knowledge teachers need as it is situated in tasks that ask teachers to make sense of student thinking, analyze multiple approaches to problems, or other authentic teaching activities. Fully defining the SKT in which we are interested is outside the domain of this article. However, we rely heavily on the work of Lamon (2007) and Lobato and Ellis (2010) in our definition. For example, we know that teachers need to have the ability to conceptually connect the two values in a ratio and to understand that a third, abstractable quantity results from that connection (e.g., Lobato & Ellis, 2010). Teachers also need to understand the multiplicative relationships inherent in proportional relationships (e.g., the constant of proportionality is the multiplicative relationship of one value in the ratio to another). And, they need to understand that this corresponds to the unit rate. Further, we assert that teachers should understand how proportional reasoning connects to other areas of mathematics such as fractions and geometric similarity (Lamon, 2007; Pitta-Pantazi & Christou, 2011). We rely on research on teacher knowledge that shows that teachers struggle to use unit rate when faced with values less than one (Harel & Behr, 1995; Post, Harel, Behr, & Lesh, 1988). And, we know from a series of small studies (e.g., Riley, 2010; Son, 2010) that teachers tend to rely on cross-multiplication, which seems to obscure the breadth of knowledge they may have about proportional relationships.

Considering only the domain of proportional reasoning for teachers, one could take a number of approaches to measuring different aspects of SKT. In this study, we considered one assessment's approach to measuring the construct of proportional reasoning knowledge for teaching through the lens of our emerging definition. By undertaking this effort, we were able to further define the domain of SKT for proportional reasoning and to consider valid measurement of that domain. We relied on an approach that searched for latent classes in the data and what those latent classes might reveal about the nature of teachers' knowledge. In this setting, latent classes are statistically determined groupings of participants who shared particular aspects of patterns in their responses to the assessment items. We suspected that the data available for this study might contain latent classes given results from previous research by Izsák, Orrill, Cohen, & Brown (2010) on teachers' understanding of rational numbers. That research indicated that in a similar assessment, a two-class model fit the data better than the one-class model.

Mixture Rasch Model

For this analysis, we used the mixture Rasch model (described below) as a means of further examining the data provided through Item-Response Theory (IRT) by an assessment of teacher knowledge. The standard Rasch model is useful for tests that are designed to assess single categories of knowledge. With respect to measures of teachers' proportional reasoning, the standard model constrains the information about a teacher to a single estimate of proportional reasoning. As such, the standard model will not detect differences in the ways that examinees respond to individual items on the test. The Rasch model assumes that all examinees are drawn from a single population. When it is suspected that this may not be the case, such as when different groups of teachers have different patterns of responding to the items, then a mixture Rasch model (Rost, 1990, 1997) may be more useful.

The mixture Rasch examines patterns in responses to the assessment items that allow participants to be placed into latent classes. Latent classes are not determined a priori nor are they typically determined by more apparent commonalities such as ability, gender, or race. Members in a latent class are homogeneous on the characteristic that caused the latent class to form. Previous research with mixture IRT models in general, of which the mixture Rasch model is one example, has demonstrated that these models can address whether or not examinees exhibit the same response characteristics or whether there are groups of examinees that are latent and can only be identified by examining homogeneities in their response patterns. The groups are termed latent since they are not immediately visible simply by examining their responses. Observable characteristics like gender, height, and ethnicity are considered manifest. In contrast, characteristics such as differences in use of cognitive strategies for answering test items are considered latent until subsequent analysis can make them manifest. Previous research, for example, has found that latent classes of students that differ in their use of cognitive strategies for answering test questions can be detected (Bolt, Cohen, & Wollack, 2001; Embretson & Reise, 2000; Mislevy & Verhelst, 1990; Rost, 1990, 1997).

In a recent study of fraction knowledge that used the mixture Rasch analysis, we found that teachers in one latent class attended to the referent unit more than those in the other class (Izsák et al., 2010). These fine-grained understandings are important for mapping a domain of knowledge that teachers should have and for understanding what learning means within the domain. For example, we saw that teachers sometimes transitioned between latent classes following an intervention, albeit without exhibiting growth in their test scores. Similarly, some had significant changes in test scores without changing latent classes (e.g., Izsák, Jacobson, de Araujo, & Orrill, 2012). In our previous work, the latent classes were found to align with aspects of knowledge that are critical for meaningful understanding of fractions. We assert, therefore, that understanding latent class membership can provide important insights into the nature of teacher knowledge that ability scores alone cannot.

Methods

Sample

For the purposes of this analysis, we used the same dataset for all three analyses, but included different subsets of items for each version. While determining the latent class membership, item difficulty, and ability scores relied on the combined datasets (25 teachers from our sample plus 351 teachers from the LMT-PR sample), the analysis of the latent classes relied solely on an analysis of the 25 teachers participating in a larger study in which this work is situated. The 351 teachers in the national sample included teachers in grades 4-8 (Hill, 2008). The 25 teachers we analyzed in depth here were all teachers in grades 6-8. These 25 participants represented a convenience sample of practicing mathematics teachers drawn from across three states. They represented an array of schools including public, private, and charter schools situated in urban, suburban, and rural settings. Because the data are presented and analyzed in aggregated form throughout this study, we have not provided in-depth information about each teacher in the study.

Instrument

The instrument of interest was the Learning Mathematics for Teaching form for Proportional Reasoning (LMT-PR: Learning Mathematics for Teaching, 2007), which included 73 unique items. Consistent with all LMT assessments, the LMT-PR seeks to measure a particular construct known as mathematical knowledge for teaching (MKT; Ball, et al., 2008). MKT emphasizes the use of knowledge in and for teaching rather than focusing on the broader body of knowledge teachers may have developed (Ball et al., 2008). Items on the LMT were designed to measure common content knowledge (CCK; math knowledge used outside of teaching) and specialized content knowledge (SCK; math knowledge specifically used in teaching) (Hill, 2008). The assessment was not created using a systematic conceptualization of the domain of proportional reasoning. Rather, the development team was opportunistic, using items that captured MKT across an array of topics with some breadth and an array of mathematical tasks of teaching (Thames, personal communication). The item development team aligned items to the *Principles and Standards for School Mathematics* (NCTM, 2000). The LMT-PR form included a wide range of questions including those asking teachers to solve proportions, to select harder or easier examples for students, to explain particular proportions ideas to students, to interpret a variety of graphs, and to determine whether given situations were directly/inversely proportional. As with other LMT forms, LMT-PR included questions that asked teachers to make sense of students' work, to interpret representations, and to select items appropriate for their students. It is the need for this array of understandings that makes the construct of teacher knowledge unique and challenging to measure.

We selected the LMT-PR both because it measures the content in which we are interested and because it situates many of the tasks in the work that teachers do, as described above. Our analysis was meant to highlight the importance of aligning constructs of interest to assessments and should not be interpreted as detracting from the important role the LMT has played in the assessment of teacher knowledge. The LMT has been one of the most widely used assessments of teacher knowledge and has been used as the basis for important research showing the alignment between teacher knowledge and student performance on standardized assessments (e.g., Hill, Rowan, &

Ball, 2005). Further, the LMT paved the way for the increasingly rich discussion of the construct of teacher knowledge in mathematics.

Table 1. *Items included in each version of the LMT-PR in this analysis.*

	LMT-PR 73	LMT-PR 60	LMT-PR 54
Items Removed from full version of LMT-PR	All items included	3a-c 10 15 24a-d 32a-d	3a-c 5 8 10 12a, 12d 14b 15 18 24a-d 32a-d

As shown in Table 1, each version of the LMT used in this study was created from the full assessment. Version 1 (LMT-PR 73) for this analysis considered all 73 items in the LMT-PR[1]. The second analysis (LMT-PR 60) removed 13 items that did not specifically measure proportional reasoning concepts. These items included several that asked teachers to interpret nonlinear graphs (see Figure 1 for an example). While these items featured covariational reasoning, they measured mathematics outside our construct of interest. If we wanted to carefully measure the domain to make inferences about teachers' proportional reasoning as we conceptualized it, then including items that did not specifically address proportional reasoning could create noise in the analysis.

[1] While the LMT is logically organized into sets of questions that appear to be testlets, in our past experience, those related questions have not had interdependent responses like testlets should. Participant responses on related questions appeared to be independent. Thus, it is appropriate to treat every item as a stand-alone item rather than considering them as testlets for the purposes of this analysis.

Ms. Reese and Mr. Ward celebrate student success by allowing students to eat small bags of popcorn at the end of the day. Describe what has happened over the course of the last six days in each of their classes.

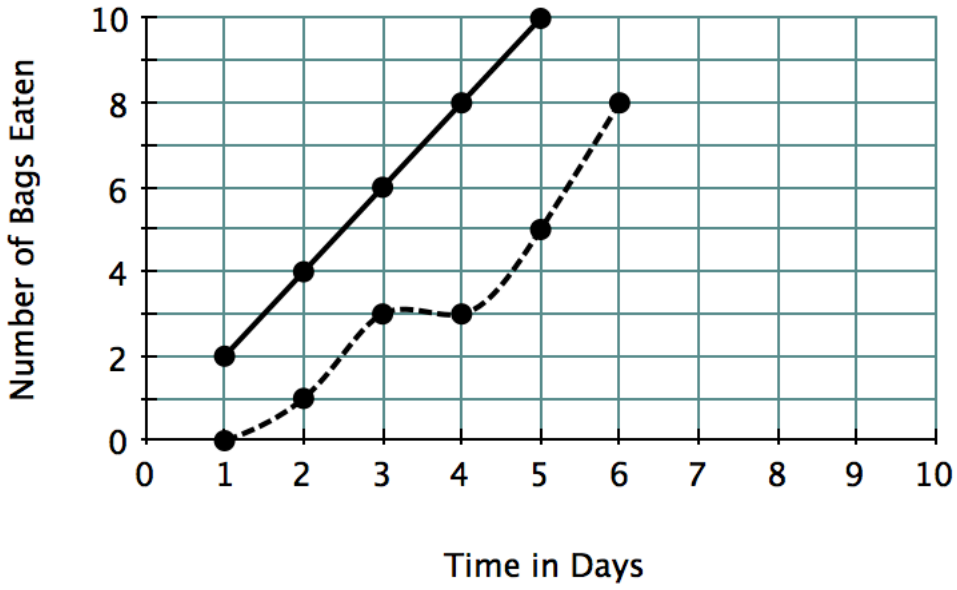

Figure 1. A task stem similar to those removed to create LMT-PR 60 for this study. (Note that actual LMT items are secure.)

For the third version, which included 54 items, we worked from LMT-PR 60 and removed those items reviewers suspected would lead to errors in measurement of the proportional reasoning domain based on our interpretation of the specialized knowledge teachers need for this domain. In particular, we were concerned with items that would lead to the right answers using incorrect reasoning or incorrect answers for reasons other than limitations in understanding.

One example of such an item was the single item from the LMT-PR for which we had item response interview data from a previous study. In that study, we asked 13 middle school teachers to think aloud on an item that asked them to explain why the cross-multiplication algorithm works. In our analysis of those interviews, we found that about half of the participants who answered the item incorrectly actually understood the intended correct answer as being the most mathematically precise. However, they chose a different response driven by their interpretation of the mathematics through the eyes of a teacher, even though the question did not necessarily ask them to. That is, they chose the less accurate answer because it more precisely reflected the way they would teach students about cross multiplication. This becomes particularly relevant to the discussion of our construct because the LMT team viewed this item as working appropriately because their definition of MKT is tightly tied to the ways in which teachers use their knowledge to teach (Thames, personal communication). Thus, the item would be seen as operating appropriately from the perspective that the teachers were not applying the most appropriate understandings in their teaching situations. However, from our perspective, which is grounded in knowledge in pieces (diSessa 1988; 2006), we assert that teachers

may have a number of knowledge resources that are connected in ways that cause them to be invoked in some situations, but perhaps not in others. Working from this perspective, we interpreted items like the cross multiplication item as not working because we saw evidence that some teachers understood the mathematics of interest, but did not invoke that mathematics in expected ways in the situation presented in the item. By not using their understanding in expected ways, the participants appeared not to have the understanding, but based on their explanation, that is not an accurate assumption.

Based on our understanding and experiences working with teachers around proportional reasoning, we removed six potentially problematic items to create the 54-item version of the assessment.

Data Analysis

As mentioned above, data analysis was done using a mixture Rasch model (Rost, 1990, 1997). In the standard Rasch model (e.g., Hambleton & Swaminathan, 1985; Lord, 1980), item responses are typically scored dichotomously (e.g., 1 for a correct response and 0 for an incorrect response). The resulting data are then used to estimate parameters that describe ability for each examinee (θ_j) and difficulty for each item (b_i). Ability (θ_j) describes the amount of proportional reasoning knowledge possessed by person j. The scale for the model is centered at 0, and if θ_j = 0, then person j is assumed to have the average amount of knowledge of proportional reasoning. The difficulty of the item, b_i, is expressed on the same scale as ability. The standard form of the Rasch model is given as

$$P_i(\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \qquad (1)$$

The item difficulty, b_i, indicates the point on the scale at which persons with ability equal to the item difficulty have a 50-50 chance of answering that item correctly. Items that are more difficult have higher difficulty parameter estimates. Similarly, items that are easier have lower difficulty parameter estimates. As an example, for item 1a (Figure 1), the item difficulty score is -2.26 for Class 1; therefore, someone whose ability is -2.26 has a 50% likelihood of answering the question correctly. In contrast, item 8 shows that a member of Class 2 to have a 50% chance of answering the item correctly, the participant would have to have an ability score of 2.736.

In our study, we used the mixture Rasch model (Rost, 1990) as estimated using a Markov chain Monte Carlo algorithm as implemented in the computer software WinBUGS (Spiegelhalter, Thomas, Best and Lunn, 2003). The mixture Rasch model can be given as

$$P_i(\Theta_j) = \frac{\exp(\Theta_{jg} - b_{ig})}{1 + \exp(\Theta_{jg} - b_{ig})}$$

where the subscript *g* is used to indicate latent class. In this equation, it can be seen that ability for person *j* and the difficulty for item *i* differ depending on the latent class.

For each item, the mixture Rasch model estimated a separate item difficulty for each latent class, a separate probability of belonging in each latent class, and a separate ability estimate for each examinee. To determine the number of latent classes, we used the BIC (Bayesian Information Coefficient, Schwartz, 1973). This is a standard approach to determining fit for mixture IRT Models (Li, Cohen, Kim & Cho, 2009).

It is a well-known caution that that a statistical result does not necessarily indicate a meaningful result. In latent class analysis, for example, researchers have found that spurious latent classes may be found (Alexeev, Templin, & Cohen, 2011). In this study, the three different solutions detected in the different mixture Rasch model analyses were each composed of different test content and thus had different interpretations. Given that the mixture Rasch model was of the same family as the Rasch model used to originally calibrate the LMT items, that the latent classes detected in this study had a meaningful interpretation based on the membership of the different classes and that there were differences in performance on the content by members of each class on the different tests, it seems reasonable to infer that the latent classes were not spurious. Further, the classes we found had a clear meaning based on the membership of the classes and the differential responses of each latent class to the questions on the test. The membership of latent classes changed when the content of the test changed, which is consistent with the assumptions underlying latent class analyses. Finally, the responses offered by the members of the latent classes were helpful in determining how to characterize the different classes and the differences in performance on each of the test questions were useful in helping to characterize latent classes.

Once latent classes had been identified, we undertook analysis of the latent classes by considering those places where item difficulties varied significantly by class as well as those places where members of the lower-scoring class found items to be easier than those of the higher-scoring class. As shown in Figure 2, item difficulties were reported in standard deviation units, with easier items showing lower scores. Analyzing these items for trends in the knowledge used helps us better understand what makes the knowledge of the members of one class different from that of the other class. In our previous research (Izsák et al., 2010) we found that having item interview data substantially enhanced our ability to discriminate between classes. However, interview data were not available for the present study.

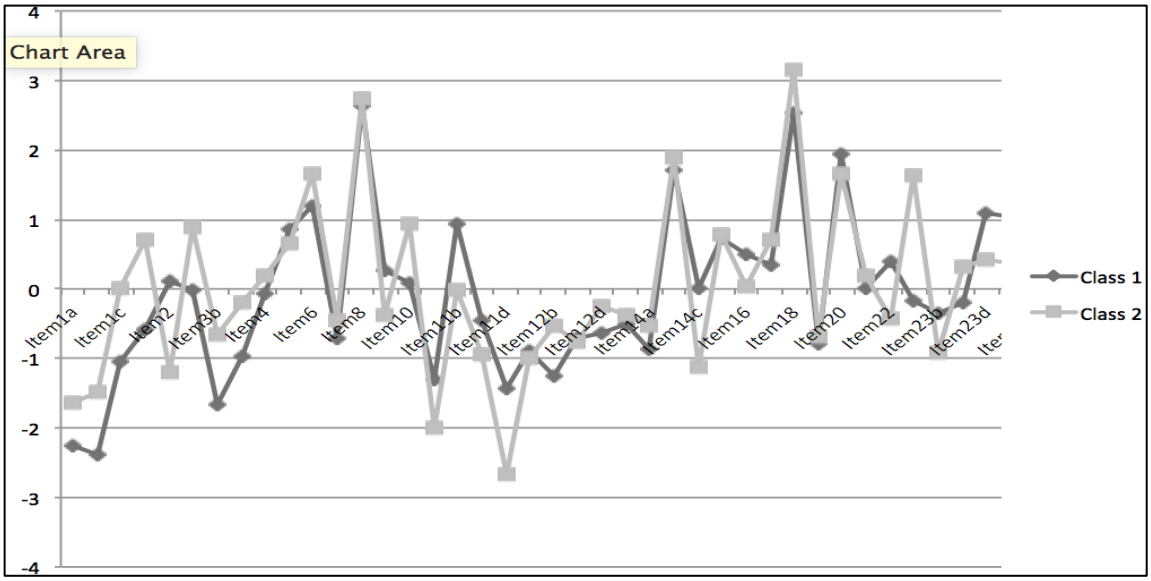

Figure 2. Plots of item difficulties for Class 1 and Class 2 in our analysis of LMT-PR. Overall, Class 2 scores are higher, but note there are items Class 1 members found easier.

In our initial analysis of the LMT-PR data from the combined dataset, we noted that the higher scoring of these classes seemed to find items focused on using cross-multiplication and other algorithms to be easier than their counterparts in the lower-scoring class. However, members of the lower-scoring class found items focused on understanding proportions and reasoning about them in a number of ways to be easier than those in the higher-scoring class. That is, the members of the higher-scoring class did better on items that allowed algorithmic thinking rather than solely focusing on reasoning about proportional relationships. These results suggested that the assessment itself may privilege teachers who are more facile with algorithms despite the field-wide emphasis on the importance of being able to reason about proportions (e.g., Lamon, 2007; Lobato & Ellis, 2010; NCTM, 2000). This finding raised interesting questions about the nature of the LMT-PR and what it might be measuring. This, in turn, led us to subsequent analyses focused primarily on 25 participants from whom we have collected completed LMT-PR forms, a prompted interview relying on a think-aloud protocol, and a face-to-face clinical interview. In the current study, we present our findings of the analysis of subsets of items on the LMT based on the performance of these 25 participants.

Results

Our aim in conducting this analysis was to consider the implications of construct definition on inferences that can be made about teachers' knowledge. Specifically, we sought to use this case as an illustration of the importance of mapping an assessment to the construct of interest. In this results section, we present the findings of our analysis of each of the three LMT-PR versions, relying on our analysis of the sample of 25 teachers. We follow with a discussion of two key issues: the impact of the definition of the domain on the overall ability scores of participants and the need to constrain assessments to

robust understandings of the mathematics of interest. In the Conclusions section, we highlight approaches that may support better assessment creation.

Ability Scores

Consistent with our position outlined at the outset, determining the construct to be measured is important for making claims. As shown in Table 2, removing the items that did not fit our definition of proportional reasoning led to a significant increase—where significance is defined as 0.3 or greater—in the scores of three participants (12%). By removing the items that did not measure proportional reasoning and those that were determined to be problematic (LMT-PR 54), seven of the 25 participants (28%) had significant increases in their ability scores. This changed the nature of our interpretations of these teachers' knowledge. This was particularly important for Bridgette, Kelly, and Kathleen who were all within one standard deviation of the mean at the outset, but ended well over one standard deviation above the mean. Interestingly, all of the significant changes in scores were increases rather than decreases.

Table 2. *Ability scores of the 25 participants on the three versions of LMT-PR. Each ability column shows the participants' abilities along a single continuum. Scores are reported as logits. All names are pseudonyms.*

	Name	Ability LMT-PR 73	Ability LMT-PR 60	Ability LMT-PR 54
1	Autumn	1.03	1.17	1.30
2	David	0.72	0.58	0.76
3	Alan	1.88	2.20	2.74
4	Ella	2.56	2.55	3.01
5	Mike	2.12	1.90	1.97
6	Bridgette	0.50	0.68	0.86
7	Allison	1.03	0.97	1.07
8	Larissa	2.12	2.20	2.51
9	Tori	0.87	0.87	0.97
10	Matt	0.87	0.77	1.07
11	Greg	2.73	2.75	2.74
12	Meagan	0.65	0.68	0.86
13	Brianna	0.15	0.22	0.36
14	Felicia	1.88	1.76	1.97
15	Eileen	1.47	1.51	1.55
16	Kelly	0.50	0.97	1.19
17	Todd	1.57	1.51	1.68
18	Kathleen	0.80	0.97	1.30
19	Patricia	1.67	1.76	1.82
20	Robyn	1.67	1.63	1.82
21	Peter	-0.84	-0.83	-0.58
22	Nancy	2.73	2.55	2.74
23	Diana	2.26	2.20	2.31
24	Christine	-0.34	-0.30	-0.48

25	Heather	2.40	2.99	3.34

Table 3. *Participants' abilities for each version of the LMT-PR. The Ability scores reported here are for each latent class, but have been scaled to be comparable across latent classes.*

	Name	Mixture Ability LMT-PR 73	Class LMT-PR 73	Mixture Ability LMT-PR 60	Class LMT-PR 60	Mixture Ability LMT-PR 54	Class LMT-PR 54
1	Autumn	1.12	2.00	1.13	1.00	1.26	1.00
2	David	0.80	2.00	0.61	1.00	0.77	1.00
3	Alan	1.91	2.00	1.95	1.00	2.60	2.00
4	Ella	2.53	2.00	2.26	2.00	2.80	2.00
5	Mike	2.16	2.00	1.72	2.00	1.85	1.00
6	Bridgette	0.58	2.00	0.70	1.00	0.85	1.00
7	Allison	1.10	2.00	0.84	2.00	1.05	2.00
8	Larissa	2.15	2.00	1.96	1.00	2.40	2.00
9	Tori	0.98	2.00	0.85	1.00	0.96	1.00
10	Matt	0.96	2.00	0.77	1.00	1.05	1.00
11	Greg	2.66	2.00	2.38	1.00	2.57	2.00
12	Meagan	0.78	1.00	0.70	1.00	0.87	1.00
13	Brianna	0.34	1.00	0.30	1.00	0.43	1.00
14	Felicia	1.94	2.00	1.62	1.00	1.90	2.00
15	Eileen	1.54	2.00	1.41	1.00	1.49	1.00
16	Kelly	0.59	2.00	0.94	1.00	1.16	1.00
17	Todd	1.62	2.00	1.42	1.00	1.60	1.00
18	Kathleen	0.88	2.00	0.95	1.00	1.25	1.00
19	Patricia	1.74	2.00	1.61	1.00	1.75	2.00
20	Robyn	1.73	2.00	1.49	1.00	1.73	2.00
21	Peter	-0.56	1.00	-0.63	1.00	-0.45	1.00
22	Nancy	2.65	2.00	2.28	2.00	2.56	2.00
23	Diana	2.27	2.00	1.95	1.00	2.19	2.00
24	Christine	-0.13	1.00	-0.16	1.00	-0.34	1.00
25	Heather	2.39	1.00	2.60	2.00	3.02	2.00

Latent Class Combined with Ability Scores

When we added latent class membership to the analysis, we were able to see that these three versions yielded different results (Table 3). In essence, they measured the construct differently. First, for each of the three versions, Class 2 was the higher scoring class (Table 4). This was true despite the fact that on LMT-PR 60 80% of the participants were in Class 1 whereas in LMT-PR 73 80% of the participants were in Class 2. While we initially suspected that the computer may have switched labels in the analysis (which does happen), our analysis of the responses from the members of these classes suggested

that was not the case. In LMT-PR 73, our analysis showed that for the 25 participants, membership in Class 1 indicated more facility with ratio tables and combining ratios, which includes knowing that it is okay to add ratios to each other. Class 1 teachers also found easier those items that asked them to explain why particular relationships did or did not work as proportions (e.g., using scale factor, equivalence, division, or additive reasoning as rationales). Class 1 teachers found determining particular instances of inverse proportion to be easier than Class 2 teachers and they were better at setting up proportions for simple word problems (e.g., if two tickets cost $5, how much do 10 tickets cost). In contrast, Class 2 was better at items that involved algorithms, including those using unit rate, scale factors, and cross multiplication. They found complex equations to be easier to set-up and verify (e.g., situations in which two people cut grass at different rates and one needs to set up an equation to determine how long it will take to mow a particular number of lawns) as well as showing more flexibility in acceptable proportion set-ups (e.g., lbs/lbs, $/lbs, and $/$ are all acceptable). If the class labels had just been flipped, the groupings described for LMT-PR 73 could be reversed for LMT-PR 60, however, that was not the case. Instead, for LMT-PR 60, Class 1 seemed to be identified through items that involved explanations of why a situation is proportional, much like Class 1 for LMT-PR 73. In addition, Class 1 for LMT-PR 60 was better able to identify numbers that would make problems easier or harder for students, to identify situations in which additive reasoning was being improperly used, and to correctly interpret a variety of ways to solve proportions without a standard algorithm or equation. In contrast, Class 2 found easier those items that asked them to identify situations that were related in ways that were not proportional (e.g., linear relationships), those that asked them to model complex situations, and those that asked them to use scale factor reasoning for within measure space scale factors (e.g., given $3/17 liters = $5/$x$ liters, one can divide $5 by $3 to see how many times larger 5 is than 3. Then multiply that result by 17 to determine x).

Class 1 and Class 2 meant different things in these analyses despite Class 2 including 80% of the participants in LMT-PR 73 and Class 1 including 80% in LMT-PR 60. In short, while it was clear that the class separations were unique to each version, the latent classes did not clearly organize the teachers in ways that might help us make inferences beyond whether the teacher was likely to be comfortable with algorithms.

Table 4. *Mean logit scale scores for each latent class for each version of LMT-PR*

	LMT-PR 73	LMT-PR 60	LMT-PR 54
Class 1 mean	0.565 n=20	1.096 n=5	.909 n=14
Class 2 mean	1.595 n=5	1.939 n=20	2.233 n=11

More interesting was LMT-PR 54 because more teachers scored significantly higher on it and because it seemed to separate the teachers into latent classes in ways that might be more useful for measuring our construct given the more balanced distribution of

teachers between the classes. Closer examination of LMT-PR 54 class membership highlighted more mathematical sensitivity in the separation of classes. Whereas LMT-PR73 and LMT-PR 60 both had clear separations, one of the clear distinguishing features between classes was that one class found using algorithms and algebraic approaches (such as modeling equations) to be easier than the other class. In LMT-PR 54, that distinction disappeared, which suggested that the privileging of algorithmic reasoning might have been mediated in this version of the assessment. Instead, we saw finer-grained and more conceptually grounded mathematical ideas separating the classes. For example, Class 1 found items that relied on scale factors determined by the within measure space relationship to be easier, whereas Class 2 found items related to scale factors determined between measure space to be easier (e.g., given \$3/17 liters = \$5/x liters, one can divide 17 liters by \$3 to determine the constant relationship, then multiply 5 by that value to answer for x). While both classes found particular items involving the combination of ratios to be easy, Class 2 seemed more able to both break down ratios, combine ratios, and to make sense of the ratio table representation. Class 2 found items that focused on why one cannot use addition to maintain equivalent ratios to be easier than Class 1. However, Class 1 was still better able to select explanations for why particular relationships did or did not work as proportions (e.g., using scale factor, equivalence, division, or additive reasoning as rationales). This was the one set of questions that Class1 consistently found easier across all versions. Class 1 in LMT-PR 54 also found the identification of easier or harder numbers (e.g., which set of numbers would make this problem harder for students to solve?) to be easier than Class 2 did. Interestingly, in LMT-PR 54, we see that Class 1 participants had an easier time with percentages as they related to proportions (for example, a sale price of 40% off is not the same as taking an additional 10% off of a 30% off price). This suggests that Class 1 might be more sensitive to questions grounded in the work of teaching (e.g., making pedagogical decisions) versus the work of solving problems.

In summary, LMT-PR 73 and LMT-PR 60 were more consistent with our initial analyses of LMT results. The class membership indicated that the higher scoring class was the one that found algorithms easier. It was not until we used the most narrowed form—the form that included only the 54 items related to proportional reasoning and removed items that seemed potentially problematic—that we were able to see differences in how participants reasoned about and with proportions and we were able to start seeing some separation tied to pedagogical concerns. This suggests that an assessment more focused on the construct of interest may yield more sensitive results.

Discussion

In this paper, we used the case of the LMT-PR to highlight the importance of aligning the construct of interest to the assessment being used. Our analysis was aided by the use of psychometric models that allow us to vary from the unidimensional analysis to which traditional IRT scores are constrained. This allowed us to look at the same participants' performance on different versions of the same assessment.

Our findings indicated that the version of the assessment that was most tailored to our definition of the construct yielded clearer information about teachers' understandings of proportional reasoning and the higher overall scores for our participants than the less focused versions. Given the emergence of high stakes testing for teacher hiring and

evaluation, being able to make clear, strong claims about teacher understanding is critical. For over one-quarter of the 25 participants in this study, scores varied significantly according to the particular items included in the assessment raising questions about what the versions of the assessment tell us about individual teacher knowledge. In contrast, at the group level (which is what the LMT is designed to measure) there was less variation in means. The group mean for the 25 participants on LMT-PR 73 was 1.39 and LMT-PR 54 was 1.49. We speculate that some of the limitation in variation is due to the overall skewing of our scores. The 25 participants we worked with were mathematically stronger than the national sample used to create the scale on which the scores were based.

In our analysis, we presented evidence that considering the assessment items' alignment to the construct of interest matters to the outcomes of that assessment and the inferences that can be made. The claims we can make about the participants changed based on the items considered, and the guidance provided by an analysis of a particular set of items could lead to potentially different implications for further learning opportunities for these participants.

Importance of Aligning Assessments to the Defined the Construct

This study demonstrated that aligning an assessment as closely as possible to the defined construct can lead to significantly different interpretations of teachers' knowledge than those assessments that are less aligned. Thus, it is important for assessments to be as well aligned to the construct as possible. The LMT-PR was created with the NCTM standards at its heart (Hill, 2008), however, those standards allowed a broad interpretation of the domain of proportional reasoning, which led to item development that spanned a variety of topics. This led to both limits in the number of items focused on key proportional ideas and the number of items focused on the mathematics to which proportions are connected. For example, there are no questions linking proportions to slope and only two questions that link proportions to similarity. There are also no questions that ask teachers about the definition of a ratio and how that definition might be the same as or different from a fraction. This is a limitation in the LMT-PR's alignment to our construct of interest.

Of course, alignment also relies on robust definitions of the construct to be measured. This is problematic in domains like teacher knowledge where the constructs are not yet well defined. It is also difficult when using existing assessments that may not fit a construct as tightly as necessary.

Our approach of constraining a pre-existing assessment to key mathematical ideas related to the construct yielded more useful information for guiding subsequent instruction and/or making claims about participants' understandings. LMT-PR 54, the version most aligned with our conception of the specialized knowledge teachers should have of mathematics, yielded information at a finer grain size than the less focused versions. In general, it seems that Class 1 found pedagogically focused questions easier (such as selecting easier or harder numbers), whereas Class 2 had more facility with manipulating equivalent ratios. This information is at a fine-enough grain size to help a professional developer plan subsequent instruction.

Additional work would need to be done to translate the findings of this study into a format that would support instructional decision-making for professional development.

But, we assert that the analysis presented here suggests there is promise in using latent class measures, combined with using focused assessments, to provide results that can provide the basis for instructional decision making.

Limitations

As with all studies, this one has a number of limitations. Here we present three major limitations. First, the national dataset included a number of 4th and 5th grade teachers for whom proportional reasoning is not in the content they teach. This may have led to very different results for the latent class analysis than if the teachers had all been from middle school. Further, our findings are limited by our lack of interview data with members of the latent classes. We could make stronger claims and, perhaps, find more similarities and differences between the latent classes were we able to hear why the teachers selected particular responses. Finally, we were limited in that the definition of the construct measured in the LMT is different from our own definition. This study sought to find the best fit between our construct definition and that LMT-PR, but that is still not as well aligned to our definition as if we had developed an assessment ourselves.

Conclusion

Clearly, the development of assessments for measuring teacher knowledge is an area in need of much more consideration (Orrill & Cohen, in press). In the domain of proportional reasoning, for example, additional work is needed to define the specific construct of interest: the knowledge teachers should reasonably be expected to know.

Test developers and users could ensure alignment between their construct of interest and the assessment using one of many available techniques that rely on mapping the assessment to the domain of interest. For example, the construct of proportional reasoning as we defined it and the assessment used may have been achieved through reliance on systematic identification of the subconstructs of interest and intentional spread of items across the subconstructs (e.g., Izsák et al., 2010).

Another approach to mapping the domain would be through the use of a Q-matrix, which provides a confirmatory approach to measuring a domain. Q-matrices, critical for cognitive diagnostic models (e.g., Izsák & Templin, in press), require the research team to identify the subconstructs to be measured and indicate which of those subconstructs each item addresses. This allows tracking of all of the subconstructs intended to be measured, thus ensuring not only that the relevant subconstructs are being measured but also that they are being paired in multiple ways so that one idea does not obstruct the other. For example, if we are interested in the use of representations and the teachers' understandings of combining ratios, we would not want all of the combining ratios tasks to include ratio tables because that representation may be unfamiliar to teachers, thus masking the participants' actual knowledge of combining ratios.

In the end, careful consideration of the alignment of the assessment to the construct is important for making claims of validity. Given the widespread move toward using assessments of teacher knowledge for high stakes decision-making, this becomes even more important. The study presented here shows that the same teachers have the perception of more knowledge or less simply based on the items included in the analysis.

Acknowledgements

Work on this paper was supported by the National Science Foundation through grant number DRL 1054170. The opinions expressed here are those of the authors and do not necessarily reflect the views of the NSF. The authors wish to thank Heather Hill for sharing the LMT-PR data with us as well as Mark Thames for his conversations about MKT. We also wish to thank Eric Gold, Dave Kamin, Tim Marum, Rob Nanna, Dennis Robinson, Ryan Robidoux, and Kaitlyn Walsh Rodrigues for their assistance in the initial analysis of the LMT-PR assessment.

Works Cited

Alexeev, N., Templin, J., & Cohen, A. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement, 48*, 313–332.

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

Baumert, J., Kunter, M., Bum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., …, & Tsai, Y. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

diSessa, A. A. (2006). A history of conceptual change research: Threads and fault lines. In R. K. Sawyer (Ed.). *The Cambridge handbook of the learning sciences* (pp. 265–282). New York: Cambridge University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mawhah, NJ: Lawrence Erlbaum.

Hambleton, & Swaminathan, H. (1985). *Item response theory: Principles and application*. Boston, MA: Kluwer-Nijhoff Publishing.

Harel, G., & Behr, M. (1995) Teachers' solutions for multiplicative problems. *Hiroshima Journal of Mathematics Education, 3*, 31–51.

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.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.

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., Jacobson, E., de Araujo, Z., & Orrill, C. H. (2012). Measuring growth in mathematical knowledge for teaching fractions with drawn quantities. *Journal for Research in Mathematics Education, 43*(4), 391–427.

Izsák , A., & Templin, J. (in press). Coordinating descriptions of mathematical knowledge with psychometric models: Opportunities and challenges. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. xxx–xxx). Journal for Research in Mathematics Education monograph series. Reston, VA: National Council of Teachers of Mathematics.

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.

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.

Li, F., Cohen, A.S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for dichotomous mixture IRT models. *Applied Psychological Measurement, 33*(5), 353–373.

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. N. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Manizade, A. G., & Mason, M. M. (2011). Using Delphi methodology to design assessments of teachers' pedagogical content knowledge. *Educational Studies in Mathematics, 76*, 1883–207. DOI: 10.1007/s10649-010-9276-z

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association & Council of Chief State School Officers (NGA & CCSSO) (2010). *Common core state standards* mathematics. Washington, DC: Author.

Orrill, C. H., & Cohen, A. (in press). Purpose and conceptualization: Examining assessment development questions through analysis of measures of teacher

knowledge. To appear in *Journal for Research in Mathematics Education Monograph.*

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

Post, T., Harel, G., Behr, M., & Lesh, R. (1988). Intermediate teachers' knowledge of rational number concepts. In Fennema, et al. (Eds.), *Papers from First Wisconsin Symposium for Research on Teaching and Learning Mathematics* (pp. 194–219). Madison, WI: Wisconsin Center for Education Research.

Riley, K. R. (2010). Teachers' understanding of proportional reasoning. In P. Brosnan, D. B. Erchick, & L. Flevares (Eds.), *Proceedings of the 32nd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1055–1061). Columbus, OH: The Ohio State University.

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

Rost, J. (1997). Logistic mixture models. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 449–463). New York: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

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

Silverman, J., & Thompson, P. W. (2008). Toward a framework for the development of mathematical knowledge for teaching. *Journal of Mathematics Teacher Education, 11*, 499–511. DOI: 0857-008-9089-5

Son, J. (2010). Ratio and proportion: How prospective teachers respond to student errors in similar rectangles. In P. Brosnan, D. B. Erchick, & L. Flevares (Eds.), *Proceedings of the 32nd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 243–251). Columbus, OH: The Ohio State University.

Spielgelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007). *WinBUGS with DoodleBUGS.* Medical Research Council, Imperial College and MRC, UK.