Using Psychometrics to Advance Assessment in Mathematics Education:  
The Does it Work Project

Chandra Hawley Orrill  
University of Massachusetts Dartmouth

The Does it Work: Building Methods for Understanding the Effects of Professional Development (DiW) project is focused on understanding what teachers learn in professional development, how they use that learning in their classrooms, and whether the teachers’ learning has impact on student performance. To this end, DiW has developed an item pool to help assess teacher learning from the professional development experience.

How are theoretical perspectives on knowledge reflected in the assessment?

The DiW project, and the assessment we developed for it, uses two perspectives in the design of the instrument that are unique. First, we focus on fine-grained understandings of teacher knowledge and, second, we applied a systematic mapping of the domain as the organizational structure for thinking about both the content we were interested in and what it means to know this mathematics. Further, we moved beyond IRT models based solely on scaling and used the mixture Rasch model that classifies and scales examinees.

While there are a number of studies that have considered teacher knowledge, too often, those studies focus on deficits in understandings. For example, Ma’s (1999) work compares Chinese teachers’ knowledge to American teachers’ knowledge to determine the places where the knowledge is lacking about particular concepts. Similarly, in Borko et al. (1992), the authors followed a single preservice teacher as she attempted to teach fraction division. They found that she lacked the conceptual understandings to provide answers to students’ questions about why the algorithmic approach to solving such problems works.

In contrast, research on students’ reasoning about rational numbers has long been focused on understanding how knowledge develops over time and how various schemes support the development of more robust understandings of these concepts (e.g., Mack, 1990, 1995, Olive, 1999; Steffe, 2001, 2003; Tzur, 1999, 2000). In this work, there is often, though not always, a developmental aspect but, more important for our study, there is an effort to gauge what the students do understand and where that understanding breaks down. To this end, the researchers have tended to develop tasks and methods that help to better delineate what students do and do not understand about fine-grained mathematics rather than clustering topics together into a coarser division of the content.

---

1 Presented as part of Izsák, A., Confrey, J. E., Orrill, C., McCrory, R., & Kelly, A. (April, 2010). Using psychometrics to advance assessment in mathematics education. Symposium presented at the Research Presession of the 88th Annual Meeting of the National Council of Teachers of Mathematics: San Diego. The work reported here is supported by the National Science Foundation under grant number REC-0633975. The results reported here are the opinions of the author and may not reflect those of NSF. Special thanks goes to the Does it Work research team for their support in every aspect of the reported work.
In DiW, we have sought to build from the fine-grained analyses of student reasoning in some important ways while working within the limitations of the multiple-choice test format. First, we used research on students’ understanding of mathematics, as well as the coarser-grained teacher research, to help us conceptualize the domain that we wanted to measure. Rather than simply conceiving of a set of items as measuring “rational number for middle school teachers” we mapped the domain in particular way. As shown in Figure 1, we created a domain map. The first separation of content was a coarse one as we separated fractions from decimals. Then, we drew from literature about how students come to understand fractions to identify that one of the major issues students face in developing a conceptual understanding of fraction operations is identifying and reasoning about the referent unit. With that literature in mind, we subdivided each category into smaller categories:

- **Part Whole Comparisons** is a category that looks at whether the participant can norm (e.g., name a fraction in relation to a particular whole) at the most basic levels.
- **Addition and Subtraction** items are straightforward in that all of the values in the situation refer to the same whole.
- **Multiplication**, in contrast, requires learners to understand two different wholes. The solver must consider how to isolate a portion of a portion of a whole. So, \( \frac{1}{3} \times \frac{4}{5} \) could be conceptualized as locating and naming an area that is \( \frac{4}{5} \) of a whole and identifying the subsection that is \( \frac{1}{3} \) of that \( \frac{4}{5} \). The result is expressed in terms of the original whole.
- **Division** also has different wholes. In division items, however, the referent unit shifts in the response so that the final answer refers to the divisor. In a problem such as \( \frac{2}{3} \div \frac{1}{2} \), the question is considering how many \( \frac{1}{2} \)s of a whole are in \( \frac{2}{3} \) of the whole. The quotient, \( \frac{4}{3} \), indicates that there is one entire \( \frac{1}{2} \) plus another \( \frac{1}{3} \) of \( \frac{1}{2} \) in the original \( \frac{2}{3} \).
- **Proportion and Ratio** were included under fractions because they are often conceived of and operated on in terms of fractions.

<table>
<thead>
<tr>
<th>Fractions and Percent</th>
<th>Numeric</th>
<th>Verbal</th>
<th>Drawing</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare/Part Whole</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Addition and subtraction</td>
<td>1</td>
<td>4</td>
<td>(2)</td>
<td>7</td>
</tr>
<tr>
<td>Multiplication</td>
<td>2</td>
<td>2</td>
<td>2 (5)</td>
<td>11</td>
</tr>
<tr>
<td>Division</td>
<td>1</td>
<td>4</td>
<td>(3)</td>
<td>8</td>
</tr>
<tr>
<td>Ratio and proportion</td>
<td>1</td>
<td>1 (4)</td>
<td>(4)</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decimals</th>
<th>Compare/Place Value</th>
<th>1 (1)</th>
<th>–</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addition and subtraction</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Multiplication</td>
<td>1</td>
<td>–</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Division</td>
<td>1</td>
<td>4</td>
<td>–</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>Counts</strong></td>
<td><strong>12</strong></td>
<td><strong>26</strong></td>
<td><strong>26</strong></td>
<td><strong>64</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1: Domain Map for DiW items.*
Once we had created an organization of the domain centered on referent units, we further refined our conceptual map by considering the representations of mathematics that we wanted to include in our assessment. Because ours is a professional learning agenda aimed at supporting teachers in becoming better able to implement mathematics in ways that are consistent with the NCTM standards, we chose to include three representations: those in which quantities were presented numerically, those in which numbers referred to quantities presented verbally (e.g., word problems), and those in which numbers referred to quantities presented visually (e.g., drawings of lengths and areas).

With this detailed mapping of the domain in mind, we were able to begin constructing our assessment. We began by identifying as many items from the middle grades LMT assessment of MKT (Hill, 2007) as appropriate to populate our assessment. Then, we constructed additional items to create a set of items that was more balanced across all cells of our domain map. (Note that in Figure 1, the items developed by the DiW team are in parentheses while those from the LMT are not.)

Our goal in the creation of the assessment in this way was to look beyond simply whether teachers understand fraction division or multiplication. We aimed to create an assessment that would provide teachers with multiple ways to demonstrate their knowledge of referent units in multiple contexts (e.g., fractions versus decimals or drawings versus numerically). And, we aimed to create an assessment that would allow us to see both the kinds of content teachers seemed to understand as well as the kinds of representations of that content that were more accessible for the teachers.

We matched our assessment to the mixture Rasch model, which detects subgroups of examinees by identifying patterns of responses across all items on the measure. The mixture Rasch model then scales examinees within each subgroup. Unlike other kinds of class separation, this analysis does not rely on prespecification of variables on which to separate participants such as educational background, certification, or years of practice. Instead, the statistical model analyzes the participants’ responses to determine if there are distinct patterns. In the case of this assessment, we found a two-class solution. This means that teachers are separated into one of two classes as well as being placed along a traditional IRT continuum. In this way, we are provided with additional information about how the teachers approach this mathematics. (See Izsák, Orrill, Cohen, & Brown, 2010 for elaboration on the mixture Rasch model and our analysis of the two-class solution.)

What research questions is DiW answering by pairing descriptions of mathematical knowledge with psychometric models in different ways?

One of the driving goals of DiW was understanding what teachers learn. Part of the goal of the project, in some sense, is to identify what it means for teachers to learn. Because we are using the mixture Rasch models and because we have an extensive collection of qualitative data collected from the professional development class sessions and through interviews, we have three definitions of learning that we can consider. First, we can consider gain scores to be a sign that learning occurred. This is the simplest measure in that it simply considers whether a teacher was able to respond to more questions correctly on the posttest than on the pretest. Because our
scores are reported as z-scores, a gain score, in this case, indicates that the participant moved to a new location along the scale of ability created by all teachers\(^2\).

The second measure of learning we are considering is whether there is movement from one latent class to the other. In general, one of our latent classes, Class 1, includes teachers who are stronger overall than in Class 2. This does not imply that all members of Class 2 score below those in Class 1, rather, the overall means of class 2 are lower. Further, we have conducted an extensive analysis of the latent classes to determine what understandings seem to characterize each. This is explained in detail in our recent *Elementary School Journal* article (Izsák, Orrill, Cohen, & Brown, 2010). In that analysis, we found that Class 1 teachers were better able to reason about referent units than Class 2 teachers and that they seemed to have a better understanding of multiplying by 1 and its effect in fraction multiplication and division situations. Therefore, having teachers move from Class 2 to Class 1 as a result of the professional development, regardless of gain score, would be considered a sign that the teachers had learned.

Finally, learning can be measured from careful analysis of qualitative data. Because we have the assessment data, we are able to pinpoint those points in the qualitative data that might be the most fruitful for analysis to understand what teachers learned through their participation in InterMath. In our current work, we used patterns in the quantitative data to highlight those places in the teachers’ reasoning that seemed to account for movement from Class 2 to Class 1. We are currently conducting an analysis of this kind to understand learning from a qualitative perspective.

**How is DiW leveraging empirical results from the mathematics education research base?**

This is addressed in the first question.

**How did DiW develop and validate the assessment items?**

To establish the construct validity of items that we developed, we administered a pilot version of our instrument to 13 teachers. Within 2 to 3 days of their completion of the instrument, we followed up with a one-hour interview with each teacher that was primarily focused on the items we had created. Consistent with the goal of understanding how they approached the items, we asked questions such as “Why did you choose choice A?” and “What about the wording of this question was confusing?” We did not reveal to teachers whether their answers were right or wrong.

The interviews helped us ensure that teachers were selecting correct answers for ways that demonstrated the intended mathematics—for instance, identifying appropriate referent units for numbers—not because they guessed or selected the correct choice based on faulty reasoning. The interviews also helped confirm that when teachers chose incorrect answers it was because they did not understand the intended mathematics, not because they misinterpreted the question.

Once we had completed these interviews, we edited or removed items to create an instrument that consisted of 64 items, 45 from the middle grades LMT assessment of MKT item pool and 19 that we developed for DiW. We also included the same survey of professional development.

---

\(^2\)While the mixture Rasch model produces a separate scale for each latent class, our psychometric team calculates the necessary adjustments to allow us to treat the z-scores as if they are on the same scale.
experiences used in the middle grades LMT assessment instrument. This survey collects data about grade levels taught, number of courses taken in mathematics, and certification information.

Once the instrument was fully developed, we administered it to a convenience sample of 201 teachers spread across 13 districts in 4 states. The teachers came from urban, suburban, and rural schools. We selected 16 teachers who were located in the Southeast to interview about the final instrument to solidify our construct validation work and to help us interpret results of fitting the mixture Rasch model to our data.

We were allowed to use data collected by the University of Michigan as part of their LMT efforts as well. These data allowed us to determine the extent to which our sample seemed to mirror their randomized national sample.

Is the mixture Rasch models adequate for assessing knowledge that matters?

The mixture Rasch model is a step in the right direction for better supporting professional developers in understanding both the needs of their participants and in determining how effective the professional development was. Like the traditional IRT model, the mixture Rasch model provides information about how well a set of teachers perform compared to the “average” and the z-scores generated as results of the analysis can be useful for interpreting whether the change from pre to posttest is significant (like the University of Michigan, we use .3 standard deviations as a measure of significance). However, the mixture Rasch provides additional information in the form of class membership analysis. Admittedly, the test developer has to do an initial analysis of class membership in order to determine what those classes mean in terms of the content of interest. But, once that analysis has been conducted, class membership has considerable value both in terms of thinking about learning and in terms of thinking about the design of instruction. For example, if a professional developer were able to administer our pretest before a professional learning experience, s/he could determine whether the group needed additional support in the area of referent unit reasoning or whether they were more solid in that area. In this way, the professional development experience could be somewhat customized for the participants. The professional developer would also be able to determine the effectiveness of the professional development by considering whether there were increases in z-scores and whether teachers moved from one latent class to another. However, mixture Rasch models fall short of providing professional developers with detailed information about particular aspects of understanding.
Works Cited


