**Psychometric Models and Assessments of Teacher Knowledge**

Andrew Izsák
Joanne Lobato

San Diego State University

Chandra H. Orrill
Allan S. Cohen
Jonathan Templin

The University of Georgia

**Abstract**: We discuss three ways to coordinate descriptions of mathematical knowledge with psychometric models (i.e., statistical models for tests) when building assessments. The three examples are sequenced to move from more coarse-grained to more fine-grained descriptions of mathematical knowledge and to move from models that simply scale to those that both scale and classify. We use tests on rational number developed for in-service middle school teachers to illustrate each combination. Our primary goal is to describe and compare several approaches to assessment that can inform research in undergraduate mathematics education as well, not to describe results of research on teachers in detail.

In this paper we examine different ways that the research base in mathematics education can be leveraged to build tests that can be used with large samples. Much of the research on mathematical thinking has relied on case-study methods that provide insight into how individuals access and use knowledge in the course of reasoning about problem situations. The time and effort required to use these methods renders them impractical with large samples. Large-scale tests, especially multiple-choice tests, are relatively easy to administer to large samples, but are usually insensitive to aspects of reasoning that case-study work has identified as essential for understanding what people know and can do mathematically. Recent developments in the fields of mathematics education and psychometrics, however, are creating new opportunities for building innovative assessments that can be used with large samples and that are more sensitive to aspects of reasoning highlighted in case-study work.

We illustrate three approaches to coordinating descriptions of mathematical knowledge with psychometric models using multiple-choice tests developed for in-service middle school teachers. We focus on tests for teachers because several research projects have been coordinating descriptions of knowledge with psychometric models in this area. All of the projects are investigating knowledge that teachers need to enable their students' learning, and all of the tests include items about rational numbers as treated in the middle school. The three examples are sequenced to move from more coarse-grained to more fine-grained descriptions of mathematical knowledge and to move from psychometric models that scale based on scores to those that both scale and classify teachers into groups. The examples illustrate the range of possibilities for working at the intersection of mathematics education and psychometrics. Our primary goal in this report is the coordination of descriptions of knowledge with psychometric models. Rather than describe results of research on teachers in detail, we focus on approaches to assessment that can inform research in undergraduate mathematics education as well.

**Example 1 (Using Item Response Theory models to measure "amounts" of mathematical knowledge):** The first and best known of our three examples comes from the work of Ball and colleagues (e.g., Ball & Bass, 2000; Ball, Thames, & Phelps, 2008; Hill, Sleep, Lewis, & Ball, 2007) who have developed a construct termed *mathematical knowledge for teaching* (MKT). In so doing, they have refined Shulman's (1986) conceptualization of teachers' knowledge in terms of subject-matter knowledge, pedagogical knowledge, pedagogical content knowledge, and other categories. One main refinement has been the further subdivision of subject-matter knowledge into common content knowledge and specialized content knowledge.

*Common content knowledge* is knowledge of mathematics that many educated adults have and use in a variety of professions—for instance, knowledge of procedures for computing with fractions. *Specialized content knowledge* is knowledge of mathematics that is used specifically in the work of teaching—for instance, knowledge that would support teachers' efforts to analyze students' novel approaches to computations and judge whether those approaches would generalize to other examples.

Ball and colleagues have also developed several multiple-choice instruments for elementary and middle grades teachers. Much of their effort has gone into developing items that measure common content knowledge and specialized content knowledge as defined above. In one report, Hill (2007) described a multiple-choice instrument for middle school teachers (hereafter referred to as the Learning Mathematics for Teaching, or LMT, instrument) designed to measure common content knowledge and specialized content knowledge in number and operations and in prealgebra/algebra. When describing the content of the LMT instrument, Hill reported:

> The number and operation category includes whole number operations, rational number characteristics and operations, integers, ratio and proportion, percent, and radicals. In the area of prealgebra/algebra, we included items designed to measure teachers' knowledge of linear, quadratic, and exponential functions; algebraic expressions and simple equation solving; inequalities; an absolute value with unknowns. (Hill, 2007, p. 99).

Hill also reported an item domain map with four cells reproduced below in Table 1 (from Hill, 2007, p. 100). The numbers in Table 1 indicate the number of items on the LMT test for each cell of the matrix. In preparation for the later part of our report, we point out that a large swathe of mathematics has been packed into each of the four cells. For this reason, we characterize this approach as one based on describing *broad categories* of knowledge.

TABLE 1: *Item domain map for the MKT/LMT instrument*

| | Common Content Knowledge | Specialized Content Knowledge | Total |
|---|---|---|---|
| Number & Operations | 22 | 22 | 44 |
| Prealgebra/Algebra | 17 | 31 | 48 |
| Total | 39 | 53 | 92 |

Ball and colleagues have coordinated broad categories of knowledge with traditional item response theory (IRT) models (e.g., Hill, 2007; Hill, Ball, & Schilling 2008; Hill, Schilling, & Ball, 2004). In so doing, they have measured overall "amounts" of knowledge that teachers possess either within one or across several such categories. For our purposes here, discussing the most basic item response theory model suffices. (Item response theory encompasses a family of related models.)

Using traditional item response theory models, these researchers administer tests like the one described by Hill (2007) and discussed above. Responses are scored dichotomously (e.g., 1 for a correct response and 0 for an incorrect response). The data are then used to estimate a set of parameters that make the observed responses most likely. The parameters include an "ability" parameter for each teacher and a "location" parameter for each item. The ability parameters, $\theta_j$, are standard scores describing the overall "amount" of mathematical knowledge possessed by each teacher. The scale is centered at 0: $\theta_j = 0$ can be interpreted to mean that teacher $j$ possesses the average amount of knowledge. Teachers with above average amounts of knowledge have positive ability ($\theta_j > 0$). Teachers with below average amounts of knowledge have negative ability ($\theta_j < 0$). Units of measurement are in standard deviations. IRT models are based on scaling because they order teachers along a common scale represented by the number line. For instance, these models allow researchers to say that a given teacher is a certain number of standard deviations above or below the mean on the number and operations scale or along the prealgebra/algebra scale.

The location parameters describe the relative difficulty of each item. Each item has its own item characteristic curve that models the probability that teachers will get the item correct as a function of their "ability," as represented by the $\theta_j$s. IRT models use one of several different forms for these functions. For the Rasch model (a one-parameter logistic model), the functions are of the following form:

$$\mathrm{P_i}(\theta) = \frac{1}{1+\exp(-(\theta - \mathrm{b_i}))} \quad (1)$$

The location parameters $b_i$ (also called the difficulty parameters) describe points on the ability scale. The subscript, *i*, indicates that each item has its own location. If a teacher has ability equal to the difficulty of a particular item, $b_i$, then $P_i(b_i) = 0.5$. Thus, an item location parameter of –1 (see Figure 1) means that, according to the model, a teacher with ability one standard deviation below the population mean ($\theta = -1$) has a 50-50 chance of answering the item correctly. A teacher with more ability ($\theta > -1$) would have a better than 50-50 chance of answering the item correctly, and so on. From this it follows that harder items have higher locations on the ability scale, and easier items have lower locations on the ability scale. (For the interested reader, Hambleton, Swaminathan, and Rogers, 1991, provide a good introduction to IRT.) The third example that we discuss in a subsequent section of the paper uses very different functions to model the probability that a teacher will answer an item correctly.

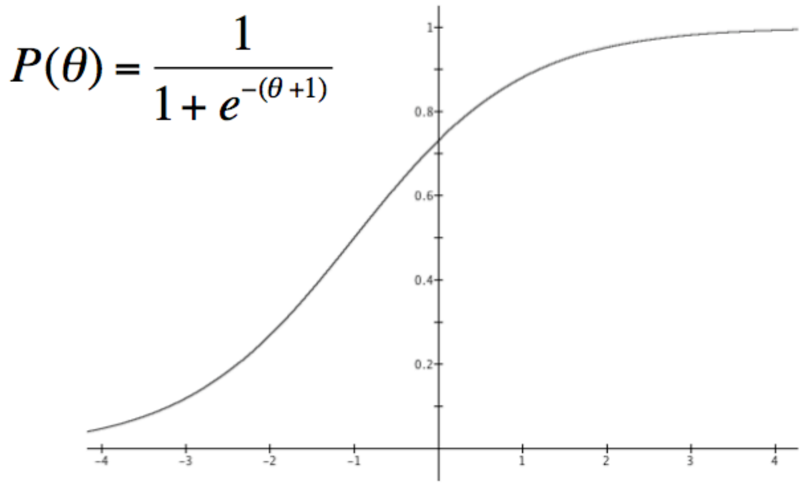


*Figure 1*. A Rasch model item characteristic curve with a location parameter of – 1.

**Example 2 (Using mixture IRT models to detect systematically different understandings of mathematics):** The second example comes from the on-going NSF-funded research project called *Does it Work?: Building Methods for Understanding Effects of Professional Development* (DiW). Orrill is principal investigator; Izsák and Cohen are co-principal investigators. One main difference between this project and the MKT/LMT work summarized above is that we examine a narrower slice of mathematical content in greater detail. At the center of the DiW project is a professional development course designed to help in-service middle grades teachers develop their capacities to reason about arithmetic with fractions, decimals, and proportions that are embedded in problem situations—for instance, as drawn length or area quantities. Numerous studies have reported that teachers struggle to justify numeric procedures in this domain (e.g., Ball, 1990; Borko et al., 1992; Ma, 1999; Sowder, Philipp, Armstrong, & Schappelle, 1998; Tirosh, 2000; Tirosh & Graeber, 1990).

An essential mathematical issue that surfaces when numbers are used to describe physical quantities is that of referent units. To illustrate, consider the following two problems:

1. Carrie has run 1/2 mile. If she walks another 1/3 mile, how far has she traveled?

2. John has 1/2 cup of flour for baking cookies. If each recipe requires 1/3 cup of flour, how many recipes can he make?

In the first problem, 1/2, 1/3, and the answer, 5/6, all refer to the same unit: one mile. In the second problem, 1/2 refers to the cups of flour John has, 1/3 refers to the cups of flour required per recipe, and the answer, 3/2, refers to the number of recipes that he can make. In contrast to Problem 1, each number in Problem 2 refers to a different unit. A main goal for the professional development at the center of the DiW project was to help teachers improve their capacity to reason with referent units consistently and appropriately across situations.

We emphasize that, in contrast to broad categories such as common content knowledge and specialized content knowledge of number and operations, we focused on a more fine-grained understanding when emphasizing referent units. Furthermore, we attended explicitly to the fact that knowledge use is context sensitive: A teacher might reason with referent units appropriately in one situation but not another. Thus, a valid measure requires providing teachers with opportunities to reason about referent units in several different situations and examining their performance across those situations. Our attention to context sensitivity is reflected in the domain map we used to develop the professional development course (see Table 2).

First, we divided the content into two large categories, fractions and decimals. We then subdivided these categories into two further categories based on referent units for numbers. The first category includes those tasks where the referent unit is the same for all fractions or decimals (similar to Problem 1). These include tasks that involve comparing sizes (or ordering), reasoning about parts of one fixed whole, and adding and subtracting (white cells). Tasks in the second category include those where different numbers refer to different referent units (similar to Problem 2). These include multiplication, division, and ratio and proportion tasks (gray cells).

TABLE 2: *Item domain map for* the *Does it Work? instrument*

| | | **Numeric** | **Verbal** | **Drawing** |
|---|---|---|---|---|
| Fractions and Percent | Compare/Part Whole | 2 | 7 | 1 |
| | Addition and subtraction | 1 | 4 | (2) |
| | Multiplication | 2 | 2 | 2 (5) |
| | Division | 1 | 4 | (3) |
| | Ratio and proportion | 1 | 1 (4) | (4) |
| Decimals | Compare/Place Value | 1 (1) | - | 5 |
| | Addition and subtraction | 1 | - | - |
| | Multiplication | 1 | - | 4 |
| | Division | 1 | 4 | - |

We then crossed the nine rows with three columns. For each row, we considered items that were about numeric methods, items in which numbers referred to quantities presented verbally (i.e., word problems), and items in which numbers referred to quantities presented visually (i.e., drawings of lengths, areas, or volumes). Originally we planned to use the LMT instrument reported by Hill (2007), but it was not adequately aligned with our course. With the generous permission of Hill, Ball, and colleagues we constructed our own assessment by combining items from the LMT instrument with items that we developed. Table 2 shows the distribution of the 64 items on the DiW instrument. The numbers indicate the number of items in each cell. Numbers without parentheses count items from the LMT instrument (45 in all), and numbers in parentheses count items developed for DiW (19 in all). In contrast to the LMT instrument (see Table 1), the DiW instrument takes a narrower slice of mathematical content and deliberately builds in multiple, diverse contexts in which teachers can reason about referent units.

Ms. Lovejoy gave her students the following problem to solve:

The distance from John's to Susan's house is 2/3 of a mile. If he started out from his house and has 1/4 of a mile still to go, how far has he already walked?

She had students use the number line so that they could draw the lengths. Which of the following diagrams shows the solution? Assume all intervals are subdivided equally.

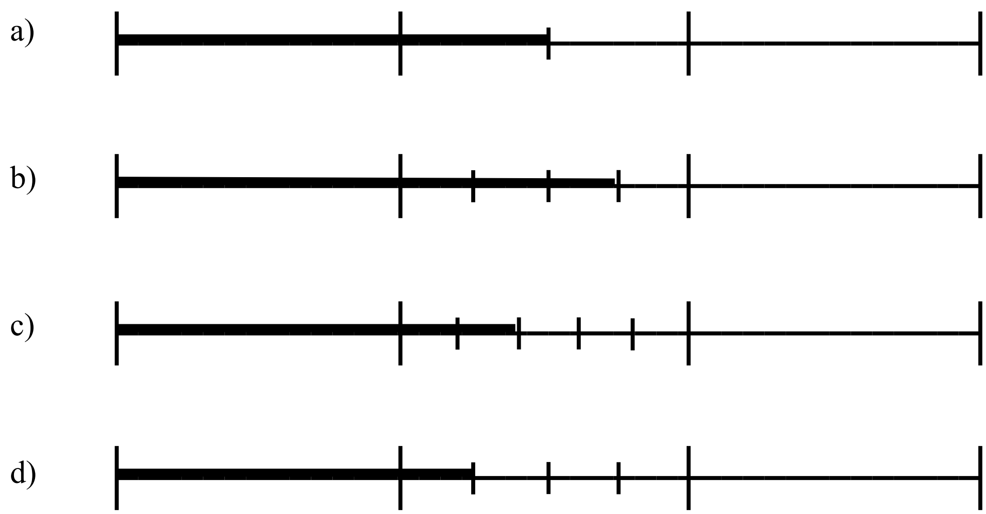


*Figure 2*. A DiW subtraction item.

Figure 2 shows a demonstration item similar to one we developed for the DiW test (actual test items are secure). In Table 2, this item would be in the addition and subtraction row and in the drawing column. The correct response is (d). A teacher who chose (a) or (b) would likely be unclear about the referent unit for 1/4: (a) shows 2/3 minus 1/4 of 2/3, and (b) shows 2/3 minus 1/4 of 1/3. A teacher who re-expressed 1/4 as 3/12 would still have to choose between (c) and (d) because both choices show 3 parts removed from an interval of length 1/3. To discriminate between these choices, a teacher could subdivide the unit interval using the partition of the second interval as a guide. In case of (c), the result is 3 groups of 5 pieces that create $15^{\text{ths}}$.

In case of (d), the result is 3 groups of 4 pieces that create $12^{ths}$. Thus, choice (d) is consistent with identifying the correct referent unit for 1/4 and subdividing intervals appropriately. Notice that selecting the correct choice requires attending not only to referent units but also to multiplicatively nested units. We will return to this point when discussing the third approach to coordinating descriptions of mathematical knowledge with psychometric models.

We coordinated our description of mathematical knowledge that emphasized context sensitive attention to referent units with a psychometric model that combines classification with the Rasch model. In particular, the mixture Rasch model allows one to ask whether, when estimating ability and location parameters as discussed above, the best model fit occurs when all teachers are treated as a single group, two groups, three groups, etc. The model detects groups based on homogeneities in response patterns. Different patterns of correct and incorrect responses are thought to correspond to different underlying cognitive strategies (Bolt, Cohen, & Wollack, 2001). As the name suggests, for each item the mixture Rasch model fits a function of the form shown in Equation 1 to each group. Classification occurs when the model assigns to each teacher a probability that he or she is a member of a particular group. Scaling occurs when the model locates a teacher's ability on a scale centered at 0 and measured in standard deviations.

We administered the DiW test to a sample of 201 teachers spread over 13 school districts in 4 states, used the mixture Rasch model to analyze our data, and found two groups (Izsák, Orrill, Cohen, Brown, 2008). Simply knowing that the best model fit occurred when teachers were classified as members of one of two groups did not tell us by itself what underlying differences in reasoning might account for those groups. Therefore, we conducted further analyses of raw test response data for the complete sample and of interview data from a subset of 16 teachers. We found that one group contained a much higher concentration of teachers who reasoned about referent units appropriately across a range of situations. (Items that made differences between the groups particularly visible were located in several different white and gray cells in Table 2.) Thus, group membership provided good indication of a teacher's capacity to reason about referent units. This is a first step toward squeezing out more fine-grained information about teachers' reasoning with numbers as quantities than is possible when using traditional IRT models to measure overall "amounts" of knowledge within broad categories such as common content knowledge and specialized content knowledge.

**Example 3 (Using Diagnostic Classification Models to assess components of reasoning):** The third example comes from a new NSF-funded project called *Diagnosing Teachers' Multiplicative Reasoning* (DTMR). Izsák is principal investigator, and the other authors of this report are co-principal investigators. The mathematical content at the center of this project is very similar to that of the DiW project. In particular, we are concentrating on reasoning about fractions, decimals, and ratios embedded in problem situations, oftentimes as drawn length or area quantities. This time we are coordinating a more elaborated description of fine-grained rational number knowledge with Diagnostic Classification Models (DCMs).

DCMs are a family of recently developed models that are being actively researched by psychometricians. Some DCMs combine scaling with classifying, others classify only. What distinguishes DCMs from the psychometric models discussed above is that DCMs provide a "profile" of "attributes" that a person has "mastered." A main point is that the profiles generated by DCMs provide substantially more information than other psychometric models about teachers' fine-grained attributes. The following paragraphs explain how we interpret the terms "profile," "attribute," and "mastery" in our work.

When using DCMs to construct multiple-choice tests, content experts first specify "attributes" which are components of reasoning in a given domain. Test question are then constructed around different subsets of attributes so that any given response, whether correct or incorrect, provides information about those attributes to which the teacher may be attending. Responses across all items on a test provide information about whether a teacher is or is not a "master" of each attribute. We interpret the statement that a teacher is a master of a particular attribute to mean that the teacher consistently uses that component of reasoning appropriately across situations. This is consistent with a description of knowledge that emphasizes context sensitive use of fine-grained understandings. Notice that because each attribute serves as a dichotomous categorical variable (a person either is or is not a "master" of that attribute), a test built around $K$ attributes defines $2^K$ groups to which a teacher might possibly belong. Each group corresponds to a different "profile" that describes those attributes that teacher has and has not "mastered." Thus, DCMs classify teachers in ways that provide information about multiple strengths and weaknesses in understanding.

In our application of DCMs, we are using as attributes several components of reasoning that have been identified as important in the research base on students' and teachers' thinking with fractions, decimals, and ratios. Three of these attributes are:

**Norming**: Establishing standard units for measurement from alternate choices.

**Referent Units**: Attending to the units to which numbers refer.

**Nested Units**: Constructing and interpreting multi-level unit structures.

We are in the process of developing a pool of items similar in spirit to the item shown in Figure 2 where each answer choice provides information about these and other attributes to which the teacher is attending. The final pool of test items will be summarized by what is called a Q-matrix. Columns in the matrix correspond to attributes, rows correspond to items. The Q-matrix is completed by using a 1 to indicate that a given item requires a particular attribute and a 0 to indicate that a given item does not require that attribute. Figure 3 shows a hypothetical Q-matrix. Item 1 could be the fraction subtraction item shown in Figure 2 because, as discussed above, discriminating between choices (c) and (d) requires reasoning about referent units and nested units. Item 2 and Item 3 illustrate the feature that each item can load onto different combinations of attributes.

| | Norming | Referent Units | Nested Units |
|---|---|---|---|
| Item 1 | 0 | 1 | 1 |
| Item 2 | 1 | 1 | 0 |
| Item 3 | 1 | 0 | 1 |
| … | | | |
| Item i | $q_{i1}$ | $q_{i2}$ | $q_{i3}$ |
| … | | | |

*Figure 3*. A Q-matrix with three attributes.

As mentioned above, diagnostic classification models contain a family of related models. Once again, we limit our discussion to one model. As in the case of item response theory, tests are administered to teachers and responses are scored dichotomously. The data are then used to estimate group membership for each teacher (that is probabilities are assigned to each teacher that describe how likely that teacher is a member of a particular group). In our example with three attributes, there are eight possible groups that correspond to having mastered each possible combination of norming, referent units, and nested units. Thus, classification occurs according to the attributes built into the test.

One important feature of the DCM approach is that the probability of a particular teacher getting a specific item correct is no longer modeled by a function of a continuous independent variable, $\theta$. Instead, the probability is conditional on the set of attributes that the teacher posses. Equation 2 shows how one diagnostic classification model (called the reduced re-parameterized unified model, or reduced RUM; DiBello, Stout, & Roussos, 2007) does this:

$$P\left(X_{ij} = 1 \mid \alpha_j\right) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{(1-\alpha_{jk})q_{ik}} \qquad (2)$$

Equation 2 says that the probability that person *j* gets item *i* correct depends on $\alpha_j$, where $\alpha_j$ is a k-tuple of 0's and 1's describing which attributes the person has mastered or, in other words, to which group he or she belongs. The probability is expressed as a product of $\pi_i^*$ (the probability that a person with mastery of all the required attributes gets the item correct) and "penalties" $r_{ik}$. Notice that exponents of these penalties are either 0 or 1, and they are one when the item requires an attribute ($q_{ik} = 1$) but the person does not have that attribute ($\alpha_{jk} = 1$). Figure 4 shows a graph of a hypothetical function that could model the probability of getting the fraction subtraction item correct (see Figure 2). Recall that this is Item 1 in the Q-matrix (see Figure 3). We generated this function using the following values: $\pi_1^*$ = 0.8, $r_{12}$ = 0.5 (penalty for not "mastering" referent units), $r_{13}$ = 0.6 (penalty for not "mastering" nested units). The function is hypothetical in the sense that we have not estimated any of the parameters in Equation 2 by applying the reduced RUM model to an actual dataset.

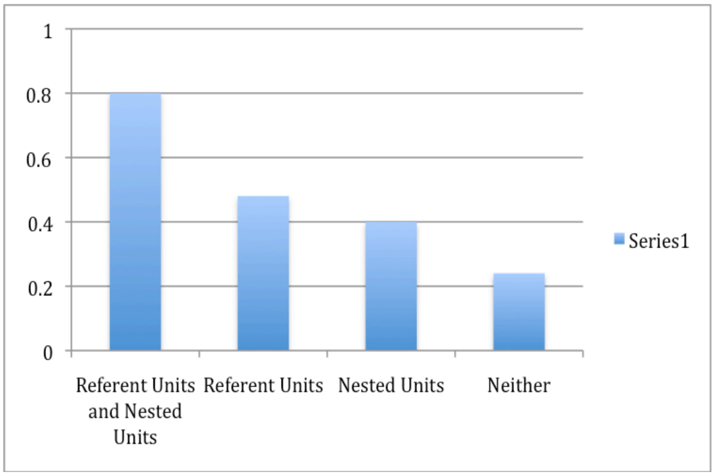


*Figure 4*. A hypothetical reduced RUM model item characteristic curve.

If we are able to develop a pool of multiple-choice items that effectively separate teachers into different groups that correspond to distinct profiles, we will be able to squeeze significantly more information about strengths and weaknesses in teachers' capacities to use things like norming, referent units, and nested units than we were able to achieve with the mixture Rasch model. Such information would be very useful for tailoring and focusing goals for professional development.

**Implications for Research in Undergraduate Mathematics Education**: The examples presented here demonstrate that there are a variety of options for combining descriptions of mathematical knowledge with psychometric models. Different options are likely to be useful for different kinds of research questions. Measuring overall "amounts" of common content knowledge and specialized content knowledge using item response theory models is useful for research that informs educational policy and the design of teacher education programs. This same approach, however, is likely to be insensitive to growth and change that might occur in teachers' knowledge during a professional development course. For this purpose, the second and third approaches hold more promise. Whether this promise can be realized hinges on the ability of researchers to harness more recently developed psychometric models, such as the diagnostic classification models.

Although the examples we have discussed are all based on multiple-choice tests for teachers, neither the item format nor the population of examinees are required. The psychometric models only require a reliable scoring system and could be used with multiple-choice or constructed-response items. What is required to extend the classification approaches presented here to research in undergraduate mathematics education is a research base that allows researchers to identify key components of reasoning in a particular content area. Finally, although we are working with very fine-grained components of reasoning such as norming, referent units, and nested units, it is possible that the classification models could be used in conjunction with attributes at a somewhat coarser grain-size. The essential criteria is that, whatever the grain size, an item pool can be developed that separates examinees into different groups each with a distinct profile.

**References**

Ball, D. L. (1990). Prospective elementary and secondary teachers' understanding of division. *Journal for Research in Mathematics Education, 21*(2), 132-144.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 81-104). Westport, CT: Ablex.

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.

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.

Borko, H., Eisenhart, M., Brown, C., Underhill, R., Jones, D., & Agard, P. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23*(3), 194-222.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Volume 26, Psychometrics* (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*, Newbury Park, CA : Sage Publications.

Hill, H. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind Act. *Educational Evaluation and Policy Analysis, 29*(2), 95–114.

Hill, H., Ball, D.L., & Schilling, S. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education,* 39(4), 372-400.

Hill, H., Schilling, S., & Ball, D. L. (2004). Developing measures of teacher's mathematics knowledge for teaching. *The Elementary School Journal, 105*(1), 11-30.

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

Izsák, A., Orrill , C. H., Cohen, A. S., Brown, R. E. (2008). *Using the Mixture Rasch Model to Assess Middle Grades Teachers' Reasoning About Rational Numbers*. Manuscript submitted.

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

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

Sowder, J., Philipp, R., Armstrong, B., & Schappelle, B. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction: A research monograph*. Albany: State University of New York Press.

Tirosh, D. (2000). Enhancing prospective teachers' knowledge of children's conceptions: The case of division of fractions. *Journal for Research in Mathematics Education, 31*(1), 5–25.

Tirosh, D., & Graeber, A. (1990). Evoking cognitive conflict to explore preservice teachers' thinking about division. *Journal for Research in Mathematics Education, 21*(2), 98–108.