Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2014, Vol. 33, No. 1, pp. 2–14

Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test Within the Diagnostic Classification Framework

Laine Bradshaw, Andrew Izsák, Jonathan Templin, and Erik Jacobson,
The University of Georgia

We report a multidimensional test that examines middle grades teachers' understanding of fraction arithmetic, especially multiplication and division. The test is based on four attributes identified through an analysis of the extensive mathematics education research literature on teachers' and students' reasoning in this content area. We administered the test to a national sample of 990 in-service middle grades teachers and analyzed the item responses using the log-linear cognitive diagnosis model. We report the diagnostic quality of the test at the item level, mastery classifications for teachers, and attribute relationships. Our results demonstrate that, when a test is grounded in research on cognition and is designed to be multidimensional from the onset, it is possible to use diagnostic classification models to detect distinct patterns of attribute mastery.

Keywords: cognitive assessment, diagnostic classification model, mathematics education, multiplicative reasoning, test construction

Multiplicative reasoning has been one primary focus of mathematics education research for several decades (e.g., Greer, 1992; Lamon, 2007) because it is critical for a wide range of topics and poses perennial difficulties for students and, in many cases, teachers. Multiplicative reasoning is fundamental to whole number multiplication and division, fractions, ratios and proportions, linear functions, and more (e.g., Vergnaud, 1983). Students' and teachers' difficulties with such topics have been widely documented (e.g., Ball, Lubienski, & Mewborn, 2001; Greer, 1992; Lamon, 2007; Ma, 1999). Example difficulties relate to conceptual meanings for the operations of multiplication and division, conceptual underpinnings for computations such as multidigit multiplication and division by fractions, and multiplicative relationships as models of situations presented through word problems or other formats. This body of research has demonstrated that multiplicative reasoning is complex and multifaceted.

Identifying knowledge that teachers need to effectively teach students has been a second primary focus of mathematics education research (e.g., Ball et al., 2001; Ball, Thames, & Phelps, 2008; Hill, Sleep, Lewis, & Ball, 2007). Establishing links between teachers' knowledge and students' achievement eluded researchers for decades, but a few recent studies (e.g., Baumert et al., 2010; Hill, Rowan, & Ball, 2005) have done so using measures inspired by the knowledge categories Shulman (1986) discussed as necessary for teaching—including subject matter knowledge, pedagogical knowledge, and pedagogical content knowledge. These measures have relied on unidimensional item response theory (IRT) models and have not attempted to model multifaceted complexity like that characteristic of multiplicative reasoning.

This study is inspired by the recent advances in conceptualizing and measuring teachers' knowledge mentioned above. In particular, we investigated the possibility of developing a test of teachers' multiplicative reasoning that would detect meaningful differences and that could be used to give constructive feedback to teachers about which components of this complex domain they do and do not understand. We narrowed our focus to reasoning about fraction arithmetic because the research literature suggested several distinct yet related dimensions in this subdomain of multiplicative reasoning. Given our purpose to provide pointed feedback with respect to multiple facets of this subdomain, we developed the test in anticipation of using diagnostic classification models (DCMs; e.g., Rupp, Templin, & Henson, 2010). In so doing, we addressed a main question for the field of psychometrics: Is it possible to construct tests that reliably measure multiple, distinct dimensions and that are practical to administer within realistic testing conditions?

Laine Bradshaw and Jonathan Templin, Department of Educational Psychology, The University of Georgia, 323 Aderhold Hall, Athens, GA 30602; laineb@uga.edu, jtemplin@uga.edu. Andrew Izsák and Erik Jacobson, Department of Mathematics and Science Education, The University of Georgia; izsak@uga.edu, erdajaco@indiana.edu. *The following members of the Diagnosing Teachers' Multiplicative Reasoning project helped draft items for the test we report: Bridget Druken, Laura Giberson, Joanne Lobato, Cynthia Lopez, Jessica McCreary, Caitlin O'Connor, Chandra Orrill, Amanda Paganin, Becky Stephens, and Lauren Susoeff. We thank the many teachers who took pilot versions of the test and helped us understand how to refine our items and the anonymous reviewers who provided helpful comments on an earlier draft.*

This research was supported by the National Science Foundation under Grant Nos. DRL-0903411 and SES-1030337. The opinions expressed are those of the authors and do not necessarily reflect the views of NSF.

DCMs are a family of psychometric models that hold promise for supporting the construction of practical, multidimensional tests. Most research on DCMs to date has focused on psychometric issues fundamental to model development, including unifying model specification with generalized linear parameterizations and testing estimation properties of the models in simulated settings. These theoretical advances are necessary precursors to practical uses of DCMs, but a successful demonstration that a test can be designed to measure a multidimensional construct with DCMs has yet to be reported.

This study takes this next step, reporting the process by which we developed a multidimensional test assessing fraction arithmetic and the results of modeling data from a large-scale administration using a DCM. An interdisciplinary team of psychometricians and mathematics education researchers developed the test as part of the National Science Foundation–funded Diagnosing Teachers' Multiplicative Reasoning (DTMR) project. The test examines knowledge that middle grades mathematics teachers have of fractions content and emphasizes using problem situations (e.g., word problems) and drawn models (e.g., number lines and rectangular areas). These emphases are consistent with recent U.S. standards for K-12 mathematics curricula (e.g., Common Core State Standards [CCSS] Initiative, 2010; National Council of Teachers of Mathematics [NCTM], 2000), and thus are consistent with knowledge that teachers need for their practice.

First, we provide background on fractions research and mathematics curricula. Second, we introduce DCMs and discuss previous analyses that have been conducted with this newer class of psychometric models. Third, we explain how we developed the test that was well aligned with the DCM framework. Fourth, we describe a general DCM, the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), that we used to model the test data. Finally, we provide results demonstrating that the test measured a multidimensional construct and that estimation was feasible with a practical test length. The concluding discussion addresses implications for psychometrics and mathematics education.

Background on Fractions Research and Mathematics Curricula

The mathematical content of the DTMR Fractions test is informed by a substantial research base on students' and teachers' reasoning about fractions and by current curriculum standards. There has been increasing recognition over the past several decades that encouraging students to memorize steps in numerical procedures as a means to learn mathematics leads to numerous problems: Because such instruction does not support conceptual understanding of arithmetic, students are prone to make errors because they forget steps or make inappropriate generalizations (e.g., National Research Council, 2001). As one example, students who have been taught to divide fractions by remembering to "invert and multiply" are often confused about whether they should invert the divisor or the dividend—for instance, a child may have to guess whether to solve $\frac{2}{3} \div \frac{3}{4}$ by computing $\frac{3}{2} \times \frac{3}{4}$ or $\frac{2}{3} \times \frac{4}{3}$.

In response, there has been a sustained effort to reform how children are taught mathematics in the United States. Both the NCTM (2000) standards and the CCSS (2010) consistently emphasize the importance of (a) having students solve problems in which numbers are embedded in problem situations as measures of quantities (e.g., lengths and areas) and (b) using drawn models (e.g., number lines and rectangular areas) as the basis for building meaning for arithmetic operations and general numerical methods for computation. Because there are many school teachers who were only taught procedural steps, such as "invert and multiply" for fraction division, there is urgent need to help preservice and in-service teachers develop understandings necessary for using problem situations, drawn models, and meaning for operations to teach students why computation methods like "invert and multiply" make sense. In fact, numerous studies have reported teachers' conceptual difficulties with different aspects of this mathematical domain (e.g., Ball, 1990; Borko et al., 1992; Izsák, 2008; Izsák, Jacobson, de Araujo, & Orrill, 2012; Ma, 1999; Sowder, Philipp, Armstrong, & Schappelle, 1998; Tirosh & Graeber, 1990). Before discussing the test-construction process, we first introduce DCMs to overview the framework that supported and guided our diagnostic test construction.

Background and Previous Uses of DCMs

DCMs, Rule Space Methodology (RSM; e.g., Tatsuoka, 1990) and the Attribute Hierarchy Method (Leighton, Gierl, & Hunka, 2004) aim to provide feedback with respect to multiple attributes, and models within these frameworks have all been referred to as *cognitive diagnosis* models (Leighton & Gierl, 2007). To distinguish latent class–based models from others that fall under the cognitive diagnosis umbrella, we use the term diagnostic classification model.

DCMs conceptualize latent constructs as sets of related categorical traits and diagnose mastery states with respect to those traits. Consistent with DCM literature, we will use the term *attribute* to refer to the categorical latent traits that tests are developed to measure. For educational tests, most DCMs assume dichotomous attributes where examinees are either a master or a nonmaster of each. The unique patterns of attribute mastery and nonmastery define the latent classes or groups by which DCMs categorize examinees. Because attributes, and therefore classes, are defined prior to analyses, DCMs are *confirmatory* latent class models where examinee classifications are determined by the item responses, the statistical properties of the items, and the population-level base rates of examinees that are masters of each attribute.

Due to the confirmatory nature of DCMs, designing a diagnostic test first requires delineating a set of attributes suggested by cognitive research to be critical for a given domain. Subsequently, each test item is constructed to measure one or more of the attributes. Because an item can measure more than one attribute, multidimensionality can exist within as well as between items. The item–attribute alignment is expressed in a *Q-matrix* where an entry of "1" indicates that an item measures an attribute and an entry of "0" indicates that an item does not. DCM classification accuracy hinges on the correct alignment of items with attributes, as Q-matrix misspecifications result in more frequent misclassification (Rupp & Templin, 2008).

Instead of establishing item–attribute alignment during the test construction process, most DCM research to this point has relied on retrofitting DCMs to existing test data. Model–data misfit is expected when multidimensional DCMs are fit to data from tests developed to assess one dimension. In particular, DCMs have been retrofitted to tests designed

for Item Response Theory (IRT)—including the TOEFL (von Davier, 2005), TIMMS test (Lee, Park, & Taylon, 2011), NAEP (Xu & von Davier, 2008), and large-scale state tests (e.g., Cheng, 2009)—and to tests designed for RSM, including the Examination for Certification for Proficiency in English (e.g., Templin & Bradshaw, in press) and Tatsuoka's (1990) fraction subtraction test.

Due to model–data misfit, retrofitting unidimensional data is not ideal for investigating the promise of DCMs for constructing multidimensional tests. The purpose of many retrofitted analyses has been to illustrate methodological advances in DCMs. Rarely has emphasis been placed on interpreting or scrutinizing results with respect to cognitive or learning theories in the tested domain or giving valid feedback to examinees (see Jang, 2009, for exception). For example, although the Tatsuoka fraction subtraction data have been used in over 10 methodological DCM publications in the last 8 years, DeCarlo (2011) demonstrated that a flaw in the Q-matrix leads to misclassification, an incompatibility that has yet to be resolved. When results of DCM analyses are thoroughly examined, the conclusion is often that the items do not fit the model well (e.g., Kunina-Habenicht, Rupp, & Wilhelm, 2009) or that a unidimensional model best represents the data (e.g., Lee, de la Torre, & Park, 2012; Templin & Bradshaw, in press).

Despite limited demonstrations of what can be achieved with DCMs, these models are attractive because they present a methodological solution for a common scenario in educational testing: Multidimensional feedback is desired, yet testing time is limited. In comparison to multidimensional IRT (MIRT) models, DCMs need far fewer items per dimension to yield reliable examinee estimates (Templin & Bradshaw, 2013). That MIRT models require more items than can be administered feasibly may be one explanation for why unidimensional tests remain predominant in education despite federal requirements (NCLB, 2002) and teachers' needs (Huff & Goodman, 2007) for more nuanced feedback.

DCMs can be viewed as alternatives to MIRT models that trade categorical for continuous latent variables in exchange for multidimensional feedback within common testing conditions. Although many researchers have questioned whether DCMs are practical for real-world testing programs because proof-of-concept studies have not demonstrated strong model–data fit, these studies have significant limitations because the test data were not designed to be multidimensional. This study investigates what might be possible in DCM applications when using an extensive research base in a targeted content area to build a test from the ground up to diagnose multiple attributes.

Construction of the DTMR Fractions Test

Attribute and Item Construction and Validation

The DTMR Fractions test was developed by a collaborative team that included mathematics education researchers with expertise in the correct and incorrect ways that students and teachers reason about fractions. We designed our test for teachers of Grades 5–7, the grades during which fraction arithmetic, especially multiplication and division, is typically taught. We assumed teachers know how to compute numerically with fractions and defined our target construct, instead, as content knowledge necessary for using problem situations and drawn models of quantities as the basis for developing general numeric methods for fraction arithmetic.

We began our test development by synthesizing the relevant literature to identify a set of core competencies that could serve as an initial set of attributes or dimensions upon which to base the test. The substantial body of research on teachers' and students' capacities to reason about fractions in terms of quantities provided a good opportunity for identifying such attributes. A main challenge was to identify a set of attributes that would span the critical competencies teachers need and then operationalize that set as distinct traits of which teachers exhibited clear mastery or nonmastery. Because the attributes we sought to diagnose are typically used in various combinations to solve problems, we expected them to be distinct, but related traits.

Once we identified the initial set of attributes, we completed three cycles of writing items, interviewing teachers to determine how they interpreted and answered the items, and revising the attributes and items in light of the interview data. Each cycle led to improved alignment among the attributes, items, and teachers' reasoning. Ensuring that we had written enough items to elicit different features of the target attributes helped establish that our test adequately covered the breadth of each attribute, which was important for establishing *construct representation* (Messick, 1989). Verifying item–attribute alignment through our iterative test construction process provided evidence that items were measuring attributes we intended them to measure, which was important for establishing the *content validity* (Boorsboom & Mellenberg, 2007) of our test. Because DCMs readily accommodate items measuring multiple attributes, we did not restrict items to measure one attribute. Rather, we allowed our understanding of the mathematical content to determine the number of attributes an item elicited.

During the first cycle of item development, project members drafted items intended to measure one or more of the initial attributes identified through the synthesis of the literature. Once we had drafted enough items to create short-test forms, we recruited 22 in-service teachers from several school districts in two states, one in the western and one in the eastern United States. The teachers answered the items and then participated in semistructured (Bernard, 1994, Chapter 10) videotaped interviews during which they explained how they interpreted and answered the items. Video was important because many of the items included drawings of numbers lines, rectangular areas, and other representations, and teachers often pointed to or marked on these when explaining their thinking. We used the teachers' responses during the interviews to determine accuracy of item–attribute alignment. That is, we looked to see if teachers' explanations indicated that they answered items correctly by using the intended attributes appropriately (true positives), answered items correctly without demonstrating the intended attributes (false positives), answered items incorrectly and did not demonstrate the intended attributes (true negatives), answered items incorrectly but did demonstrate the intended attributes (false negatives).

Data from the first cycle of item development revealed several shortcomings both with the attributes and with the initial items. In one case, we found that we could not use teachers' written responses to reliably discriminate between two related attributes because the written responses did not capture the nuanced differences in teachers' reasoning with respect to these attributes that we observed during

the interviews. In this case, we combined the two attributes into a single attribute. In other cases, teachers found ways to answer items correctly that circumvented the intended attributes—for instance, by setting up and solving algebra equations.

We responded to these shortcomings in two ways. First, we revised items that could be repaired and dropped items that could not be repaired. Second, we planned a new round of semistructured interviews during which we had teachers solve a series of constructed response tasks to help us better understand how to write items that would distinguish teachers who were masters and nonmasters of our attributes. We conducted these videotaped interviews with a new sample of 14 teachers and used the data to inform a further round of item writing. To investigate the validity of our newly developed items, we conducted a third and final round of videotaped interviews with another sample of 25 teachers. We made small refinements to some items as we progressed through these interviews. When the interviews were complete, we analyzed the data item-by-item for true and false positives and negatives as described above. Nearly all of the items performed well, and we used those well-performing items to construct the final DTMR Fractions form.

Final Form of DTMR Fractions Test

The final DTMR Fractions test was based on four attributes and included 22 stems with a total of 29 individual items, 20 multiple choice and 9 constructed response. The DTMR attributes emphasize components needed to reason about fraction arithmetic in terms of quantities, such as lengths and areas. These contrast with attributes used in previous applications of DCMs that have emphasized either procedural steps in numeric computations (e.g, Kunina-Habenicht et al., 2009; Tatsuoka, 1990) or entire branches of mathematics such as geometry and algebra (e.g., Lee et al., 2012). Mathematics education researchers perceive significant differences between procedural steps, like finding a common denominator or subtracting numerators, and the attributes described below. The former can be accomplished by memorizing and executing steps in a rote manner, while the latter focus on making sense of the conceptual underpinnings for fraction arithmetic.

The first attribute, Referent Units (RUs; Attribute 1; α_1) has to do with making different choices for the whole to which fractions refer. The ability to identify appropriate RUs is critical when numbers are embedded in problem situations. To illustrate, one interpretation of the division statement $\frac{2}{3} \div \frac{3}{4} = \frac{8}{9}$ is to ask how many $\frac{3}{4}$ths are in $\frac{2}{3}$. The answer, $\frac{8}{9}$, is $\frac{8}{9}$ths of $\frac{3}{4}$ths, not $\frac{8}{9}$ths of the whole.

The Partitioning and Iterating attribute (PI; Attribute 2; α_2) combines partitioning a quantity into equal sized pieces and concatenating unit fractions to create larger fractions. Proficiency in partitioning depends on knowledge of whole-number factors and multiples to partition in stages—for instance, one might use the fact that $3 \times 5 = 15$ to anticipate that subdividing a whole into thirds and then subdividing each third into five equal-sized pieces would create fifteenths. In the case of fraction arithmetic, one has to partition based on common multiples of two denominators in some situations and of two numerators in others. Proficiency in iterating is based on a particular meaning for fractions. At least in the United States, the part-whole definition for fractions is used widely in school curricula. According to this definition $\frac{A}{B}$ is interpreted to mean a subset of size A taken from a set of size B. As an example, one might illustrate the meaning of $\frac{3}{4}$ by saying "three of four cookies are chocolate chip." One problem with this definition is that it is hard to interpret improper fractions: How could you have "five of four cookies?" Iterating unit fractions supports an alternative to the part-whole definition in which $\frac{A}{B}$ means A copies of the unit fraction one B^{th}. Using this interpretation, one can interpret $\frac{3}{8}$ as 3 one-eighths and $\frac{9}{8}$ as 9 one-eighths. This is the definition of fraction adopted by the CCSS (2010, p. 24). Partitioning and iterating are a single attribute because they go hand in hand in many problem-solving situations. Initially, we treated partitioning and iterating as two separate attributes, but in interviews we observed that teachers rarely had partitioning facilities without iterating facilities, and vice versa. Preliminary DCM analyses confirmed that these attributes functioned as a single attribute, so we combined them.

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

Candice has 4/5 of a meter of cloth. She uses 1/8 of a meter for a project. How much cloth does she have left after the project?

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.

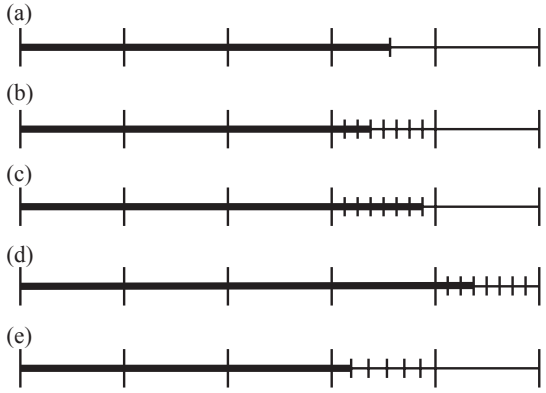

FIGURE 1. An item that measures referent unit and partitioning and iterating using common multiples of denominators.

The Appropriateness attribute (APP; Attribute 3; α_3) has to do with identifying an appropriate operation or mathematical expression for a given problem situation. A master of APP can identify word problems that call for multiplication, that call for division, and so on. For example, when teachers are asked to write a word problem that illustrates dividing $1\frac{3}{4}$ by $\frac{1}{2}$, many write problems that require multiplying by $\frac{1}{2}$ (e.g., Ball, 1990; Ma, 1999).

Finally, the Multiplicative Comparison attribute (MC; Attribute 4; α_4) has to do with forming comparisons by asking "How many times as great is one value than another?" or "What portion or fraction of one value is another?" Teachers and students form MC with whole numbers more easily than with fractions: Thinking that 12 is 3 times as much as 4 is easier than thinking that 10 is 5/2 times as much as 4.

Sample Item to Assess Reasoning with Fractions

Figure 1 shows an item similar to Item 18 on the DTMR Fractions test. The correct response is (b). A teacher who chose (a) or (c) would likely be unclear about the RU for 1/8: (a) shows 4/5 minus 1/8 of 4/5 (the mistake is taking the RU for 1/8 to be 4/5), and (c) shows 4/5 minus 1/8 of 1/5 (the

mistake is taking the RU for 1/8 to be 1/5). A teacher who reexpressed 1/8 as 5/40 would still have to choose from (b), (d), and (e) because they all show 5 parts removed from an interval of length 1/5. Choice (d) shows 5/5 minus 1/8 of 5/5 (the mistake is removing 1/8 of 5/5 from the whole segment, not from 4/5 of the whole segment). To discriminate between (b) and (e), a teacher could subdivide the length of the entire segment using the partition of the fourth interval as a guide. In case of (e), the result is 5 groups of 6 pieces that create 30ths, an incorrect partition. In case of (b), the result is 5 groups of 8 pieces that create 40ths, a correct partition. Thus, choice (b) is consistent with identifying the correct RU for 1/8 and partitioning intervals appropriately.

Methods

Psychometric Model: The LCDM

We designed the DTMR Fractions test from the perspective of a general DCM, the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009). The LCDM maps item responses onto latent attributes using a generalized linear model framework. As such, the LCDM is similar to an analysis of variance (ANOVA) model for binary data where attributes measured by an item represent fully crossed and reference-coded design factors. Consequently, the item parameters of the LCDM reflect the ANOVA-style main effects for each attribute and, for items measuring more than one attribute, interactions between attributes. Parameter values reflect the degree to which mastering additional attributes increases the probability of a correct response.

To demonstrate how the LCDM relates item response probabilities to attribute mastery status, consider an item like the one presented in Figure 1 that measures two attributes: Attribute 1 (RU, α_{e1}) and Attribute 2 (PI, α_{e2}). The LCDM provides the log-odds of a correct response as:

$$\ln\left(\frac{P\ (X_{ei}=1|\boldsymbol{\alpha}_e)}{P\ (X_{ei}=0|\boldsymbol{\alpha}_e)}\right)=\lambda_{i,0}+\lambda_{i,1(1)}(\alpha_{e1})+\lambda_{i,1(2)}(\alpha_{e2})+\lambda_{i,2(1*2)}(\alpha_{e1}\alpha_{e2}). \quad (1)$$

The parameter $\lambda_{i,0}$ is the intercept and represents the predicted log-odds of a correct response for examinees in the reference group—examinees who have not mastered RU (Attribute 1) or PI (Attribute 2). The parameter $\lambda_{i,1(1)}$ is the simple main effect for mastery of RU, representing the *increase* in the log-odds of a correct response for examinees who have mastered RU ($\alpha_{e1}=1$) but not PI ($\alpha_{e2}=0$). Similarly, the parameter $\lambda_{i,1(2)}$ is the simple main effect for mastery of PI representing the *increase* in the log-odds of a correct response for examinees who have mastered PI ($\alpha_{e2}=1$) but not RU ($\alpha_{e1}=0$). Finally, the parameter $\lambda_{i,2(1*2)}$ is the interaction effect for mastery of RU and PI that represents the *change* in log-odds for examinees who have mastered both attributes ($\alpha_{e1}=1$ and $\alpha_{e2}=1$).

To make these interpretations more concrete, consider if an item had estimated model parameters of $\lambda_{i,0}=-1.5$, $\lambda_{i,1(1)}=1$, $\lambda_{i,1(2)}=1.25$, and $\lambda_{i,2,(1*2)}=0.5$. The log-odds of a correct response for examinees who have mastered neither RU nor PI is –1.5, which corresponds to a probability of a correct response of .18 that can be calculated by the inverse log-odds function from Equation 1. For examinees who have mastered RU but not PI, the log-odds of answering the item correctly is $-1.5+1.0=-0.5$, corresponding to a probability of a correct response of .38. To quantify the strength of association between attributes and an item in DCMs, odds ratios can be used to indicate the effect sizes for item parameters. The conventional metrics for evaluating the effect size of an odds ratio apply (see Chinn, 2000, where $1.44<$ small <2.47, $2.47<$ medium <4.25, and large >4.25). For example, the effect size for the simple main effect of RU was 2.72, indicating a moderately sized effect. This odds-ratio is notated by $\theta_{i,\alpha_1|\alpha_2=0}$ and indicates that, conditional on examinees being nonmasters of PI (i.e., $\alpha_{e2}=0$), the odds of RU masters (i.e., $\alpha_{e1}=1$) answering the item correctly are 2.72 times the odds for RU nonmasters. The interaction term is interpreted similarly by conditioning on values of the attributes. For example, for masters of PI, the odds of correct response for examinees additionally mastering RU are 4.48 times the odds of a correct response for examinees who have only mastered PI. This strong odds ratio provides evidence that the interaction term is needed as both attributes are necessary to have a high probability of answering the item correctly.

Our test construction process focused on developing strong conjectures about item–attribute alignment but not about ways that the attributes would interact at the item level. We planned to inspect these interactions empirically using the LCDM. The LCDM can model attribute effects on each item response in a compensatory or noncompensatory manner, depending on the size and direction of the LCDM item parameters. Attributes are compensatory when an examinee can answer an item correctly using a strict subset of the measured attributes. In this case, the lack of mastery of one or more attributes is compensated for by mastery of other attributes. Attributes are noncompensatory when an examinee can only answer an item correctly using all attributes measured by the item. Other DCMs differ from the LCDM in that they impose the same compensatory or noncompensatory constraints across all items on the test; the LCDM relaxes these constraints and provides a flexible framework to empirically test item–attribute relationships at the item level.

Structural Portion of the LCDM

The previous section discussed the measurement portion of the LCDM, or how the LCDM parameterization relates attributes to items. The structural portion of the LCDM parameterizes how attributes are related to each other. We used a structural model that parameterized the *base rate* probabilities of mastery for each of the 16 unique attribute mastery patterns with a log-linear structure (see Chapter 8 of Rupp et al., 2010) and then, to help describe the bivariate relationship between attributes, derived tetrachoric correlations among pairs of attributes from these probabilities. The base rate probabilities of attribute profile mastery are mapped onto a series of linear model parameters including simple main effects for all attributes and all possible interactions of attributes. For these structural models, nonsignificant higher order interaction terms negligibly influence attribute relationships and do not need to be estimated (Rupp et al., 2010; Xu & von Davier, 2008).

Sample and Test Characteristics

We collected a national sample of 990 in-service middle grades mathematics teachers' responses to the DTMR Fractions test. Descriptive statistics for our sample were in rough accord with representative national samples of middle grades

teachers in other studies (e.g., Hill, 2007). Teachers reported an average of 12 years of teaching experience, were mostly White (75%, $n = 743$), and were mostly women (81.3%, $n = 805$). Nearly all (93%, $n = 922$) had experience teaching Grades 6–8 mathematics, and 885 (89%) were fully credentialed in mathematics.

We scored the teachers' responses to the DTMR Fractions test items as correct or incorrect. Each multiple choice item had one correct answer, and our team of mathematics education researchers scored the nine constructed response items. We removed Item 20 from the analyses due to difficulties in interpreting responses, leaving a total of 28 items on the test. Twenty items measured a single attribute and eight items measured two attributes. The test measured the four attributes, RU, PI, APP, and MC, with 15, 10, 5 and 5 items, respectively.

Results

Using a *conjecture-based* approach, we estimated the data using the LCDM and *Mplus* 6.11 (Muthén & Muthén, 1998–2013; Rupp et al., 2010). Our approach that we describe as *conjecture-based* began with specifying the LCDM in accord with our conjectures of item–attribute alignment (i.e., our Q-matrix). Then, we utilized the LCDM general parameterization which allows for all possible main effects and interactions among attributes in the structural and measurement components of the model. As in a general linear model framework, we empirically evaluated the significance of these parameters and removed those that were statistically nonsignificant.

Utilizing the LCDM as a tool for gathering empirical evidence to evaluate theory-based conjectures is similar to the approach used with confirmatory factor analysis (CFA): Current theory delineates the construct and hypothesizes relationships among the latent and observed variables, and modification indices provide empirical evidence for refining those hypotheses (e.g., Jöreskog, 1993). As in CFA, our procedure was not completely confirmatory because we were open to modifications based on empirical evidence that might refute our theory-based conjectures. In an ideal case, researchers combine substantive and statistical evidence in a cyclic fashion to refine their understandings of constructs and tests without over-relying on either evidence source. We view this study as an initial cycle to hone our test and theory.

We will describe first our process of model-fitting with the LCDM as it is a more general and less familiar approach than many DCMs appearing in the literature that are maximally constrained versions of the LCDM (i.e., deterministic inputs noisy and gate [DINA] model, e.g., Haertel, 1989; deterministic inputs noisy or gate [DINO] model, Templin & Henson, 2006). We then turn to results for our best-fitting model. These results focus on describing the diagnostic quality of the test items, mastery classifications for teachers, and attribute relationships.

Model Specifications

Structural model. We used a log-linear parameterization for the structural model to freely estimate the hypothesized correlations between attributes. The final structural model specification we selected was the 2-way structural model, which constrained all 3- and 4-way interactions to equal zero. Other specifications including higher order interaction terms did not converge. The 2-way structural model converged appropriately and all 2-way interaction terms and main effects were significantly greater than zero according to the Wald test ($p < .05$).

Item parameter significance. We can use item parameters in the LCDM to statistically test whether or not each item measures the intended attributes: If the main effects and relevant interactions for an attribute are both near 0, then empirically the item does not measure the attribute. Thus, beginning with a general modeling strategy analogous to that used in a fully crossed factorial ANOVA model, our first analysis included all possible interaction effects among attributes on each item. Initial results indicated that the interaction effect for RU (α_1) and MC (α_3) on Item 5 was not significant (Wald $Z = -.017$, $p = .987$). Removing this effect yielded a more parsimonious model that better represented the data, as shown by a nested model likelihood ratio test yielding $\chi\ (1) = 1.81$, $p = .179$. We continued to remove nonsignificant interaction effects and test for improved model specifications in this way, which resulted in removing 8 of the 10 interaction terms. We continued to estimate interactions for Items 14 and 17.

We examined main effects next. Main effects cannot be evaluated solely according to the Wald test statistic or likelihood ratio test because they are constrained to be greater than zero. This constraint results in a boundary violation and yields an overly conservative test (see Templin & Bradshaw, in press). Therefore, we trusted the findings of nonsignificant main effects and reported effect sizes for the remaining main effects. Results indicated that nine main effects should be removed. Five of these main effects were for the MC attribute. Of the four attributes, item–attribute alignment for this attribute was the hardest to infer when we were developing our items. To be conservative, we included the attribute as a potential main effect for all items that could be answered using MC with the expectation that statistical analysis might lead to refinements in the Q-matrix. Both main effects for Item 14 were nonsignificant, resulting in Item 14 being the one item on the test that behaved like a DINA model item with completely noncompensatory attributes. The main effect for Item 19 was removed, which led us to remove the item from the test because it did not offer information to classify teachers. The item was expected to be very difficult, even for masters of the APP attribute, because it presents a fraction division situation with which many teachers are unfamiliar. That 92.3% of teachers missed this item meant that many masters of the attribute still missed the item.

Final LCDM specification. After removing nonsignificant main effects, we estimated our final model which measured four attributes with 27 items, as shown in Figure 2. This figure depicts the LCDM for the DTMR Fractions test using a path diagram similar to those found in CFA. The unidirectional arrows show the Q-matrix structure, the bisected shapes indicate that the attributes and item responses are dichotomous, and the bidirectional arrows indicate tetrachoric correlations between attributes. The following sections provide results for this model.

Multidimensionality of the DTMR Fractions Test

Tetrachoric correlations among the attributes ranged from .626 to .781, as shown in Figure 2. These correlations are

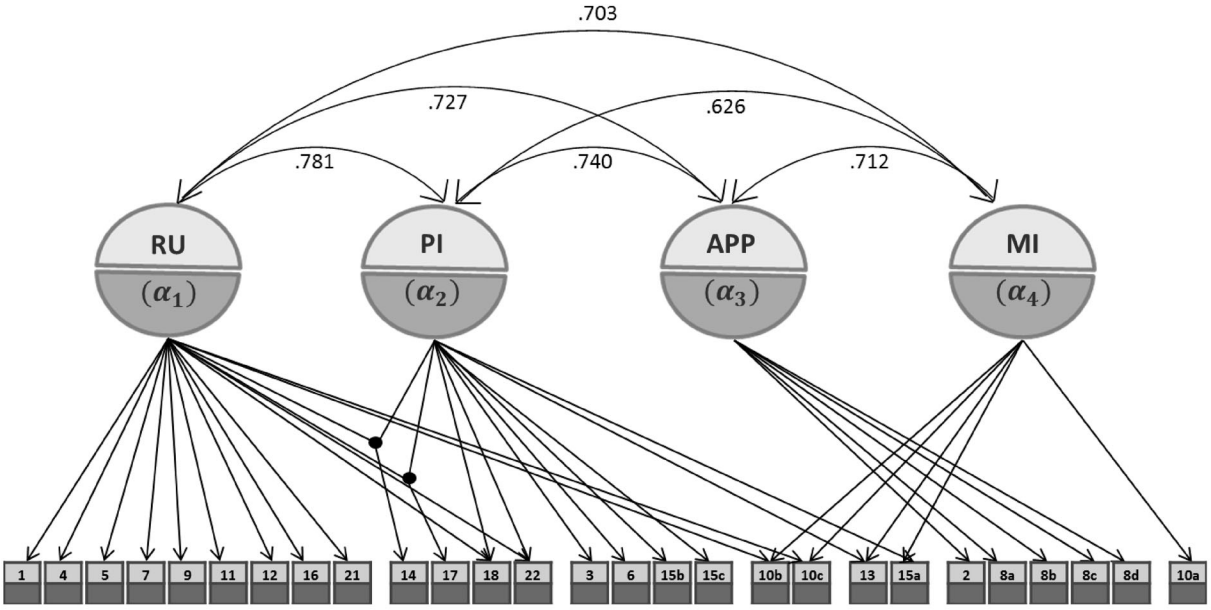

FIGURE 2. Path diagram for DTMR fractions test. Unidirectional arrows indicate the set of binary latent attributes (indicated by bisected circles) that influences each observed, dichotomous item response. The converging paths indicate significant attribute interactions for the item. The bidirectional arrows among the attributes indicate the correlations among these latent variables (values provided).

strongly positive, but not near one, providing evidence that the DTMR Fractions test is able to distinguish targeted components of the multidimensional construct. We expected the attributes to be distinct but related; however, correlations that were too large would have provided evidence that either (a) ontologically, the attributes are not distinct traits, or (b) operationally, the attributes were not separable as distinct traits by this test. This result would have replicated findings from previous studies that retrofitted DCMs. We are cautious about interpreting the range of correlations we report, because without more tests grounded in cognitive psychology or learning theory that measure fine-grained components of reasoning there is limited research against which to compare. Research on constructing multidimensional tests in education has mainly targeted dimensions that more closely resemble Thurstone's (1947) multiple factors approach, where dimensions are broadly defined as composite abilities, such as math or verbal ability (e.g., DiBello, Stout, & Roussos, 1995).

Item Parameter Estimates

The item parameter estimates and standard errors are provided in Table 1. On average, items had an intercept of –1.38, meaning roughly 20% of teachers who had not mastered any of the measured attributes answered the items correctly, presumably by guessing. Average main effect parameters ranged from 1.40 to 3.23 for individual attributes and the average of the two interactions was 1.41. The size of these effects are relative to the size of the intercept, where generally items with lower intercepts and higher main effects and interaction terms are more discriminating between masters and nonmasters of the attribute(s). Because the attributes' impact on item responses is lost at the aggregate level, item characteristic bar charts (ICBCs) for all 27 items are shown in Figures 3 and 4. Analogous to an item characteristic curve in IRT, an item characteristic bar chart plots the response probabilities on the vertical axis as a function of attribute mastery on the horizontal axis. Figure 3 provides the ICBCs for items measuring only one attribute (simple structure items). For example, the estimated probability of answering Item 1 correctly was .75 for masters of RU and .25 for nonmasters. Figure 4 provides ICBCs for items measuring more than one attribute (complex structure items). For these items, how attributes interacted at the item level can be examined. For example, consider Item 18, which was similar to the item provided in Figure 1. Figure 4 shows that the attributes were partially compensatory: The probability of a correct response to Item 18 increases from .27 to .52 to .53 to .78 when comparing examinees who mastered neither attribute, only RU, only PI, and both attributes, respectively.

Diagnostic Utility of Items

The diagnostic utility of an item can be evaluated with respect to how discriminating the item is for masters and nonmasters of the target attribute(s). To quantify the diagnostic utility, we calculated and evaluated the significance of odds ratios of correct response comparing masters and nonmasters. For items measuring one attribute, odds ratios ranged from 1.91 (Item 4) to 126.8 (Item 10a) and were all statistically significant, indicating strong effect sizes and corroborating evidence from our cognitive interviews. The conditional odds ratios for the eight items that measured two attributes ranged from 1.47($\hat{\theta}_{\alpha_4|\alpha_2=0}$ and $\hat{\theta}_{\alpha_4|\alpha_2=1}$ for Item 13) to 96.83($\hat{\theta}_{i,\alpha_4|\alpha_1=0}$ and $\hat{\theta}_{i,\alpha_4|\alpha_1=0}$ for Item 10c). All but two of these conditional odds ratios were statistically significant: The 95% confidence interval was (0.87, 15.60) for $\hat{\theta}_{i,\alpha_1|\alpha_4=0}$ on Item 10c and (0.95, 2.29) for $\hat{\theta}_{\alpha_4|\alpha_2=0}$ on Item 13. For Items 10c and 13, three of the four effect sizes indicated strong relationships between attributes and items, indicating these complex structure items also had strong diagnostic quality to distinguish between some, although not all, profiles.

Table 1. DTMR Item Parameter Estimates

i	$\lambda_{i,0}$	RU(α_1) $\lambda_{i,1(1)}$	PI(α_2) $\lambda_{i,1(2)}$	APP(α_3) $\lambda_{i,1(3)}$	MC(α_4) $\lambda_{i,1(4)}$	RU/PI $\lambda_{i,2(1,2)}$
1	− 1.12 (0.12)	2.24 (0.20)				
2	0.59 (0.13)			1.27 (0.22)		
3	− 2.07 (0.22)		1.70 (0.24)			
4	− 1.19 (0.11)	0.65 (0.19)				
5	− 1.67 (0.14)	1.52 (0.20)			*	
6	− 3.81 (0.47)		2.08 (0.50)			
7	− 0.73 (0.09)	1.20 (0.22)				
8a	− 0.62 (0.25)			4.25 (0.64)	*	
8b	− 0.09 (0.17)			2.16 (0.24)		
8c	0.28 (0.13)			0.87 (0.18)		
8d	− 1.03 (0.17)			1.81 (0.21)		
9	− 1.22 (0.10)	0.76 (0.19)				
10a	− 0.50 (0.18)	*			4.84 (0.55)	
10b	− 4.01 (0.74)	1.32 (0.28)			4.26 (0.73)	
10c	− 4.89 (0.87)	1.30 (0.26)			4.57 (0.87)	
11	− 0.88 (0.01)	1.25 (0.18)			*	
12	− 1.29 (0.11)	1.89 (0.21)				
13	− 0.74 (0.14)		0.45 (0.20)		0.39 (0.21)	
14	− 2.14 (0.14)					1.59 (0.21)
15a	− 2.48 (0.29)		2.72 (0.26)		1.05 (0.28)	
15b	− 0.56 (0.18)		2.94 (0.28)		*	
15c	− 0.44 (0.17)		3.04 (0.31)		*	
16	− 0.86 (0.01)	1.55 (0.23)				
17	− 2.08 (0.23)		1.22 (0.27)			1.27 (0.34)
18	− 0.99 (0.14)	1.13 (0.26)	1.10 (0.24)			
19				*		
21	− 1.50 (0.13)	1.69 (0.19)				
22	− 1.25 (0.16)	1.47 (0.28)	1.43 (0.25)			
Average	− 1.38 (0.21)	1.40 (0.22)	1.86 (0.29)	1.46 (0.21)	3.23 (0.55)	1.41 (0.24)
Med	− 1.12 (0.14)	1.55 (0.23)	1.30 (0.27)	1.54 (0.21)	1.52 (0.26)	1.41 (0.24)

Note. Standard errors for parameters are given in parenthesis. Item 20 was removed due to scoring. Asterisks (*) indicates the parameter was estimated in the initially hypothesized parameterization.

Attribute Classifications

Because the DTMR Fractions test measures four binary attributes, teachers were classified into 2^4 or 16 possible patterns of attribute mastery or latent classes. The LCDM estimates the probability that each teacher is a member of each latent class. These estimates were aggregated across examinees and are provided in Figure 5 with the horizontal grey bars. The most likely attribute profile (25.5% membership) was the last profile where all four attributes were mastered, and the second most likely profile (21.2%) was the first, where none of the four attributes were mastered. Seven of the last eight classes have very low class membership because RU (Attribute 1) is only mastered by 31.2% of the sample. The individual attribute mastery proportion for each attribute is shown with the vertical bars shaded black in Figure 5. These results show that the other three attributes are mastered by 55–63% of teachers, which is consistent with previous research indicating that a sizeable proportion of teachers struggle with the content tested by the DTMR Fractions test.

Test Feedback for Teachers

To demonstrate the types of feedback possible from DCMs, Figure 6 provides the marginal probabilities of attribute mastery for three teachers, referred to as Teachers A, B, C. All three teachers answered 11 of 27 items correctly, so they would be viewed as having equal amounts of ability with respect to fractions if a total score or a Rasch IRT model were used. In contrast, the DCM classifications provide very different diagnoses for the three teachers. Teacher A is a master of APP and MC; Teacher B is a master of PI; and Teacher C is a master of PI and MC. This feedback could be used to plan professional development that focused on teachers' individual weakness rather than on areas redundant with their strengths.

Attribute Reliability

Using mastery classifications for teacher feedback requires the classifications to be both valid and reliable. Item–attribute validity was discussed previously in the description of the test development process. To quantify reliability of the attribute classifications, we used the DCM measure of reliability from Templin and Bradshaw (2013). Classification reliabilities were .928, .916, .888, and .891 for RU, PI, APP and MC, respectively. As expected, the results show higher reliabilities for attributes measured by more items, as RU and PI were measured by 15 and 10 items. However, in an absolute sense reliability was very high even for APP and MC, which were measured by five items each.

Discussion

Our results demonstrate that by coupling strong understanding of the target construct with careful task development, we were able to create a multidimensional test and use DCMs to enable diagnostically useful and reliable interpretations of teachers' abilities with fraction arithmetic. In mathematics

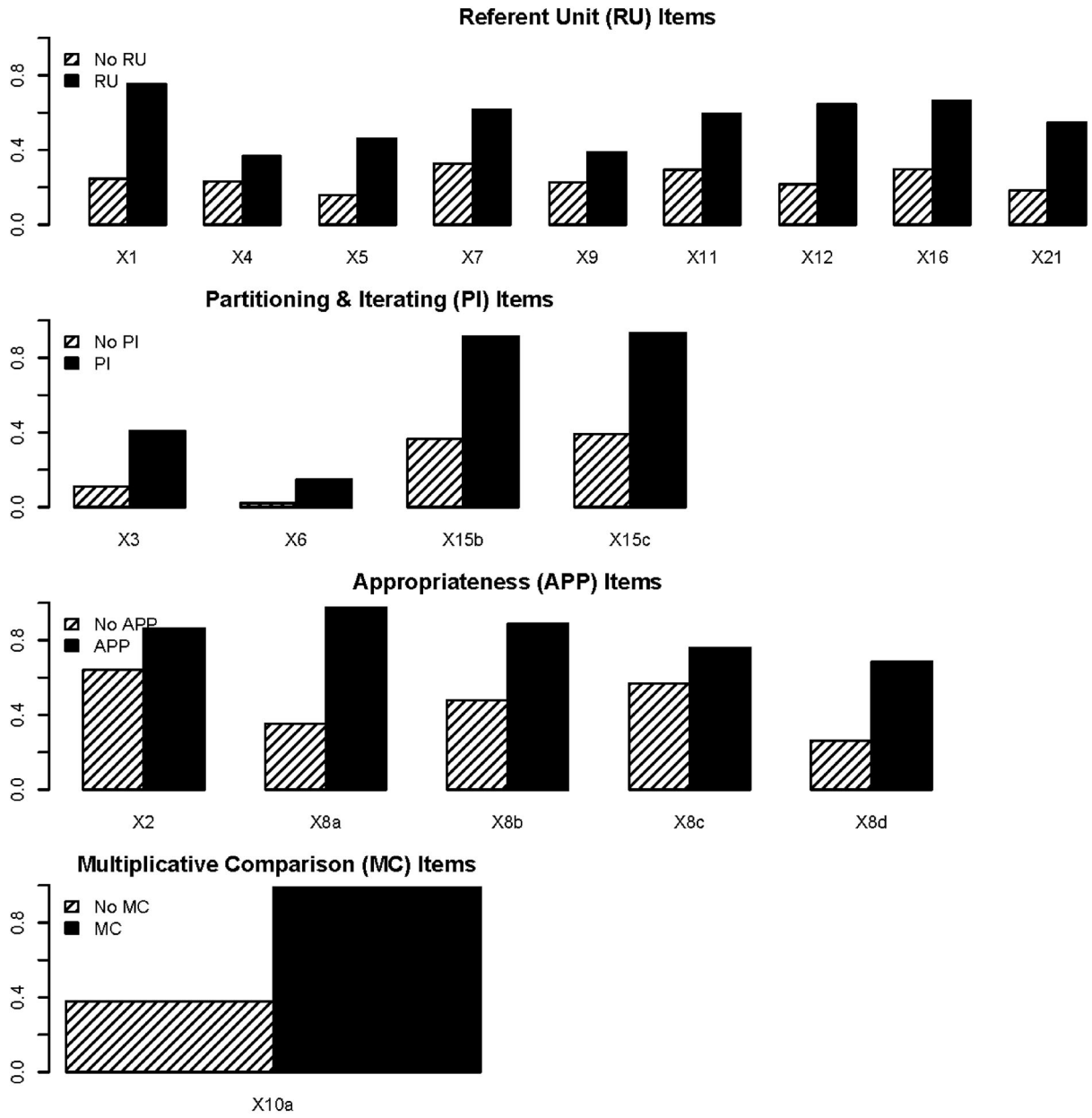

FIGURE 3. Item characteristic bar charts (ICBCs) for simple structure items. Nineteen items on the DTMR Fractions test measured one of four possible attributes (RU, PI, MC, and APP). For each of these items, the figure provides the ICBC which displays the probability of a correct response (vertical axis) by the discrete attribute mastery state (horizontal axis). For example, for Item 1 (X1), nonmasters of Referent Unit have a .25 probability of answering the item correctly and masters have a .75 probability of correct response.

education, this provides information about the components of multiplicative reasoning with which teachers must be facile if they are to be skillful with current standards-based mathematics curricula. Although considerable interest has grown in applying psychometric models to build tests of the mathematical knowledge teachers need for their practice, the majority of studies (e.g., Hill, 2007; Hill, Schilling, & Ball, 2004; Sadersholm, Ronau, Brown, & Collins, 2010; Shechtman, Roschelle, Haertel, & Knudsen, 2010) have developed tests for use with unidimensional IRT models, and a few have used mixture-IRT models (Izsák, Orrill, Cohen, & Brown, 2010; Izsák et al., 2012). Tests that locate teachers on a unidimensional scale of ability provide considerably less information about particular areas of strengths and weakness in a complex content area, like fractions, than, for example, the information we were able to report about teachers A, B, and C (Figure 6). Thus, developing a fractions test along the four dimensions of RU, PI, APP, and MC constitutes a significant advancement for mathematics education.

Results from the DCM analyses provide new insights into teachers' understandings of fraction arithmetic. First, given the extant literature, it is not surprising that teachers had difficulty with the RU items, but it is worth knowing that the RU attribute had far fewer masters in our sample than any other attribute. Second, there is literature on students' capacities for partitioning, but almost nothing is known about teachers' capacities. Our result that PI can pose significant challenges for teachers might surprise some mathematics education researchers. Third, the low rates of mastery for these attributes (31%–63%) demonstrate the need for quality

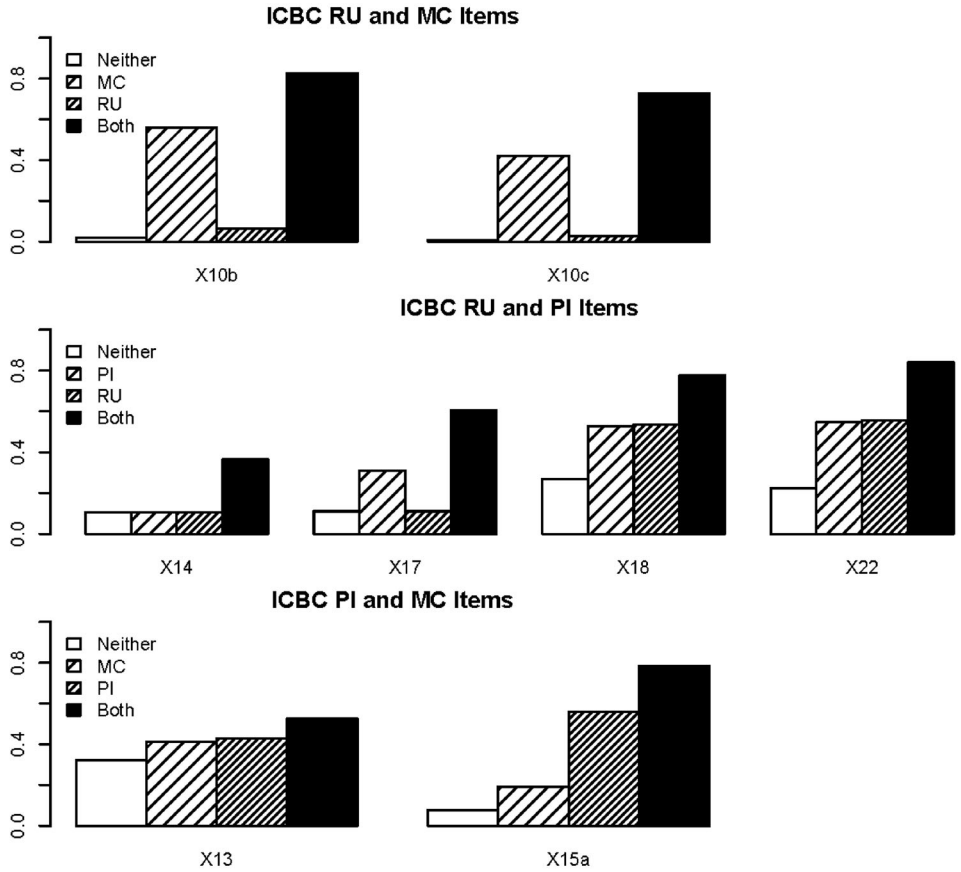

FIGURE 4. Item characteristic bar charts (ICBC) for complex structure items. Eight items on the DTMR Fractions test measured two of the four possible attributes [i.e., Referent Unit (RU); Multiplicative Comparisons (MC), Partitioning and Iterating (PI), and Appropriateness (APP)]. For each of these items, the figure provides the ICBC which displays the probability of a correct response (vertical axis) by the discrete attribute mastery states (horizontal axis). For example, for Item 10b (X10b), nonmasters of RU and MC have a near zero probability of a correct response, masters of RU only have a .06 probability, masters of MC only have a .56 probability, and masters of both RU and MC have a .83 probability of correct response.

teacher education to implement recent curriculum standards. These results corroborate findings from numerous smaller scale studies that report teachers' difficulties with reasoning about fraction arithmetic in terms of quantities.

Our results also have broader implications for developing multidimensional tests. First, the study contributes to the practice of multidimensional test construction for psychometrics by providing a model for DCM analyses that others in the measurement field could follow. Second, the ability to design tests that provide statistically sound diagnoses illustrates a critical feature of DCMs: These models hold promise for making multidimensional tests practical because traits can be measured reliably using relatively few items. We could not reasonably administer more than 30 items due to the time it took teachers to respond to our items, yet the theoretical properties of DCMs provided the ability to measure multidimensional traits with reliabilities above .85 with as few as five items for some dimensions. Studies, such as this one, complement existing simulation-based research on DCMs and demonstrate the feasibility of DCM methodology for operational use. As a contrasting example, consider end-of-grade tests in K-12 education. Depending upon the state, these high-stakes tests require between 60 and 90 items to provide an estimate of a general ability for students with a suitable reliability (e.g., Templin & Bradshaw, in press), so providing multidimensional scores in this framework is not feasible.

Third, a main lesson we learned is that identifying workable attributes can be a significant empirical problem in its own right. An initial set of attributes may be hypothesized *a priori* but will likely need more development to define a set of distinct yet related traits that can be used to separate examinees into distinct groups. We emphasize that mathematics education research has not had identifying attributes as an explicit goal, and that it took considerable effort to organize the extant literature around attributes that served as the basis for our test construction. The extensive research base on students' and teachers' reasoning about fractions provided a solid foundation from which to construct provisional attributes, but we arrived at our final set through cycles of item development and refinement. This result echoes the discovery and refinement process in personality psychology where theories have been proposed, refined, and altered since the advent of multidimensional exploratory and CFA in the 1930s. We expect identifying workable attributes to surface again as a critical challenge in future projects where researchers seek to develop multidimensional tests that are suitable for DCMs.

We close by identifying several directions for future research with respect to our test and theory refinements. First, whether there is a hierarchy among the four attributes in

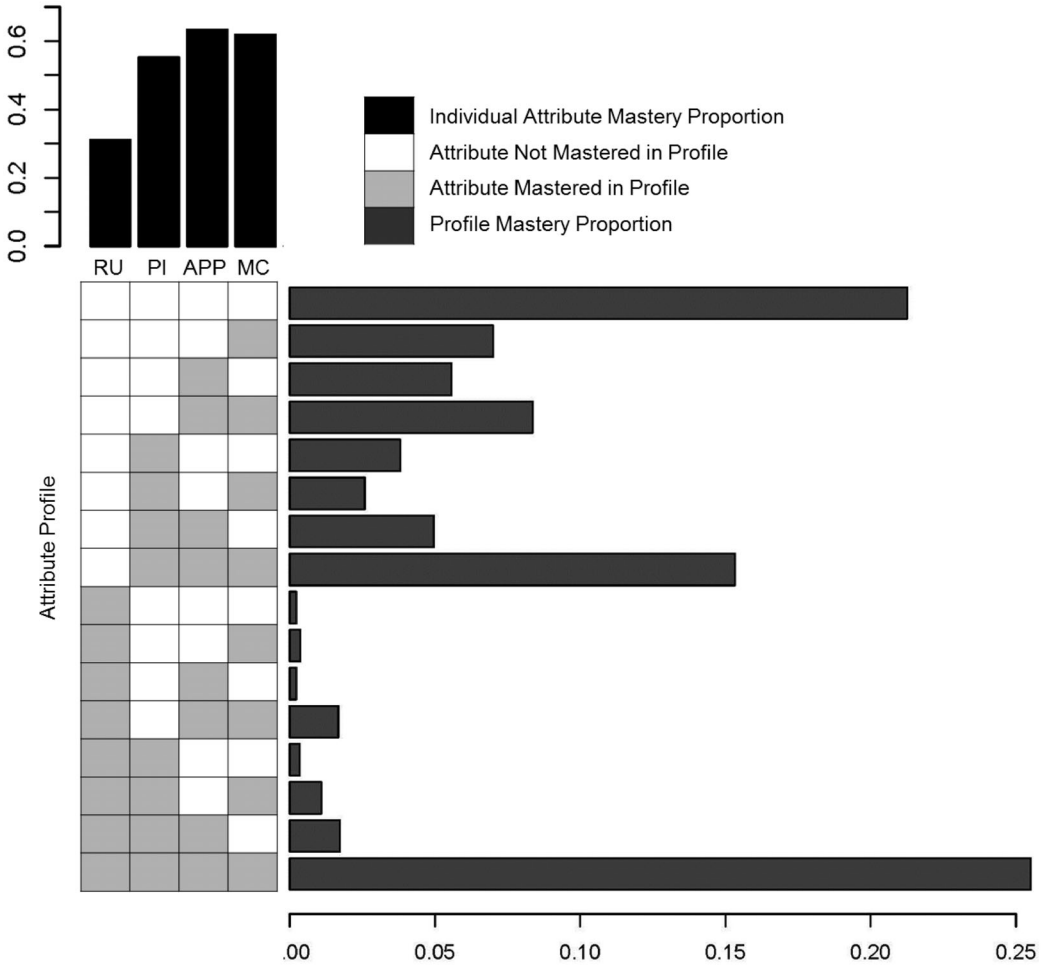

FIGURE 5. Individual Attribute and Attribute Profile Mastery Proportions. The horizontal bars in the lower portion of the figure represent the proportions of teachers who are classified with each attribute pattern of mastery, with the pattern indicated on the vertical axis by white and grey shading according to the four attributes RU, PI, APP, and MC. The vertical bars in the upper portion of the figure represent the proportion of teachers who have mastered each individual attribute.

FIGURE 6. Example teacher feedback. This figure presents individual attribute mastery probabilities for three teachers who each answered 11 of 27 items correctly.

the DTMR Fractions test is an open question. Although the distribution of attribute profiles in Figure 5 suggests that the RU attribute might be dependent on the remaining three attributes, preliminary tests we conducted did not provide evidence to support such a hierarchy. A more thorough examination of attribute hierarchies in this domain is one area for future research. Second, further empirical work is needed to provide external validity evidence to support the accuracy of the model-based teacher classifications. For example, we could conduct a study to determine the degree to which expert-based classifications judged upon interview data analyses match DCM-based classifications. Third, future research should examine how useful the DTMR Fractions test is in its intended application, teacher professional development. Feedback from the DTMR Fractions test provides probabilities that teachers are or are not masters of each attribute, and these categorical decisions could inform which aspects of fractions teachers should focus on during professional development. Furthermore, we would like to know the extent to which a test like the DTMR Fractions test is sensitive to the growth and change in teachers' knowledge as a result of professional development. These topics of future research are important steps in utilizing data-based efforts to contribute to the ultimate goal of improving teachers' capacities to facilitate student learning.

Our results demonstrate how a diagnostic test can be created and modeled with a DCM for a critical content area and that it is possible to diagnose mastery with respect to components critical for reasoning about fractions. At the same time, they raise a host of questions about relationships between knowledge, item responses, and learning for future research on applications of DCMs to consequential, practical problems in education.

References

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

Ball, D., Lubienski, S., & Mewborn, D. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, D.C.: American Educational Research Association.

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

Bernard, H. (1994). *Research methods in anthropology* (2nd ed.). Thousand Oaks, CA: Sage.

Boorsboom, D., & Mellenberg, G. D. (2007). Test validity in cognitive test. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic test for education: Theory and applications* (pp. 19–60). Cambridge, UK: Cambridge University Press.

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*, 194–222.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*, 3127–3131.

Common Core State Standards Initiative. (2010). *The common core state standards for mathematics*. Washington, DC: Author.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Lawrence Erlbaum.

Greer, B. (1992). Multiplication and division as models of situations. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 276–295). New York, NY: Macmillan.

Haertel, E. H. (1989). Using restricted latent clas models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, *74*, 191–210.

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

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

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

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic test. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic test for education: Theory and applications* (pp. 19–60). New York, NY: Cambridge University Press.

Izsák, A. (2008). Mathematical knowledge for teaching fraction multiplication. *Cognition and Instruction*, *26*, 95–143.

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

Izsák, A., Orrill, C. H., Cohen, A. S., & Brown, R. E. (2010). Using the mixture Rasch model to assess middle grades teachers' reasoning about rational numbers. *Elementary School Journal*, *110*, 279–300.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing*, *26*, 31–73.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Lang (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*, 64–70.

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

Lee, Y. S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Education Review*, *13*, 333–345.

Lee, Y. S., Park, Y. S., & Taylon, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and practices*. New York, NY: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

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

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, & B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107–110, 115 Stat/ 1449–1452 (2002).

Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, *68*, 78–98.

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

Saderholm, J., Ronau, R., Brown, E. T., & Collins, G. (2010). Validation of the diagnostic teacher assessment of mathematics and science (DTAMS) instrument. *School Science and Mathematics*, *110*, 180–192.

Shechtman, N., Roschelle, J., Haertel, G., & Knudsen, J. (2010). Investigating links form teacher knowledge, to classroom practice, to student learning in the instructional system of the middle-school mathematics classroom. *Cognition and Instruction*, *28*, 317–359.

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.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–486). Hillsdale, NJ: Lawrence Erlbaum.

Templin, J., & Bradshaw, L. (2013). The comparative reliability of diagnostic model examinee estimates. *Journal of Classification*, *10*(2), 251–275.

Templin, J., & Bradshaw, L. (in press). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

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

Vergnaud, G. (1983). Multiplicative structures. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 127–174). New York, NY: Academic Press.

von Davier, M. (2005). *A general diagnostic model applied to language testing data (RR-05–16)*. Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data (RR-08–27)*. Princeton, NJ: Educational Testing Service.