Research Report

Validating Classifications From Learning Progressions: Framework and Implementation

ETS RR–19-18

Yigal Attali
Meirav Arieli-Attali

December 2019

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Validating Classifications From Learning Progressions: Framework and Implementation

Yigal Attali[1] & Meirav Arieli-Attali[2]

1 Educational Testing Service, Princeton, NJ
2 ACT Next, Iowa City, IA, and Fordham University, The Bronx, NY

Learning progressions (LPs) have seen a growing interest in recent years due to their potential benefits in the development of formative assessments for classroom use. Using an LP as the backbone of an assessment can yield diagnostic classifications of students that can guide instruction and remediation. In operationalizing an LP, assessment items are classified as measuring specific LP levels and, through the application of a measurement model, students are classified as masters of specific LP levels. To support the use of LPs in instructional planning and formative assessment, the reliability and validity of both item and student classifications should stand up to scrutiny. Reliability of classifications refers to their consistency. Validity of classifications refers to alignment of these classifications with test data. A framework for testing these classifications is proposed and implemented in a validation study of a rational number LP for elementary school mathematics. As part of this study, 400 items were classified in terms of LP level of understanding, a cognitive diagnostic model of student mastery level within the LP was fitted to the data, and analyses were conducted to assess the reliability and validity of these classifications.

Keywords Learning progression; cognitive diagnostic modeling; formative assessment; mathematics

doi:10.1002/ets2.12253

Learning progressions (LPs) articulate a trajectory of learning and understanding in a domain in the form of levels or steps in student understanding toward proficiency (Clements & Sarama, 2004; Heritage, 2008; Post, Cramer, Harel, Kieren, & Lesh, 1998; Simon & Tzur, 2004; Weaver & Junker, 2004). LPs have been developed in many areas, including English language arts (Song, Deane, Graf, & van Rijn, 2013), mathematics (Arieli-Attali, Wylie, & Bauer, 2012; Carr & Alexeev, 2011; Clements & Sarama, 2004), and science (Alonzo & Gotwals, 2012; Duschl, Maeng, & Sezen, 2011).

By articulating a trajectory of learning and understanding in a domain, LPs can provide the big picture of what is to be learned, support instructional planning, and act as a guide for formative assessment (Heritage, 2008). An LP enables the capture of connections among the various skills and concepts of a curriculum and can take into account a variety of factors affecting student progression, including maturation (Graf & Arieli-Attali, 2015). In particular, an LP can be useful in the creation of diagnostic tools for formative assessment. By mapping LPs and developing an assessment around them, the assessment can provide teachers with information regarding the location of their students on the progression and from that derive information needed to move the students forward. For example, diagnosing a student as being at Level 2 of a five-level LP implies the student is ready to learn the material of Level 3 but may not be ready for the material of Level 4, even if the curriculum dictates Level 4 material at that point in time. Using this diagnostic information in the classroom may facilitate how teachers guide group activities and address misconceptions. However, extracting useful diagnostic information from an LP assessment requires that students' classifications to LP levels be valid and reliable.

The purpose of this paper is to describe a framework for validating student classifications into LP levels and describe an implementation of this framework in the context of an LP for rational numbers in mathematics. In the remainder of this section, we argue that an LP, as a theory about learning, needs to be operationalized through assessment tasks and measurement models. This operationalization provides a way to test the predictions of the theory, but such tests are rare in the literature. We then propose a framework for validation of operationalized LPs, present the LP that is the focus of our implementation of this framework, and conclude with specific research questions that were addressed in this validation effort.

Corresponding author: Y. Attali, E-mail: yattali@ets.org

Learning Progressions as Testable Theories

At their essence, LPs are theories of the paths students follow in learning a subject. The usefulness of LPs as a practical tool in teaching and learning is dependent on validating the hypotheses that an LP posits about how students progress toward their learning targets (Stevens, Shin, & Krajcik, 2009).

In order to support the empirical evaluation of LPs, Corcoran, Mosher, and Rogat (2009) argued that every LP should include operational definitions of what students' understanding and skills would look like at each of the stages of progress defined by the progression. In other words, the hypotheses that LPs posit should be testable (Stevens et al., 2009).

These requirements emphasize the link between LPs and the assessments that can measure student understanding of the key concepts or practices and can track student developmental progress over time (Corcoran et al., 2009). In particular, this tracking requires that assessment tasks (or items) have been designed to determine if students achieved a particular LP level. If a particular item requires knowledge and skills that are necessary at a particular level of the LP (but not at a lower level of the LP), then correctly answering this item can provide evidence that a student has achieved this particular level in his or her development.

Measurement Models for Learning Progressions

With assessment tasks thus designed, researchers can try to incorporate the LP information into their models of assessment data. Typically, the purpose of modeling assessment data is descriptive. Using either classical test theory or item response theory (IRT), the information from a test is used to determine where a student is along a given ability scale. However, several psychometric approaches have been developed that incorporate additional information about the construct measured by specifying what processes, strategies, and knowledge are involved in responding to items. This approach is diagnostic in the sense of explaining item responses through cognitive mechanisms.

Item Response Theory Framework

A possible expansion of the traditional descriptive approach to test the diagnostic power of an LP is through the use of explanatory item response models (De Boeck & Wilson, 2004), which are based on the principle that item responses can be modeled as a function of different types of predictors. Specifically, LP levels can be viewed as a characteristic of items that may predict or explain the variability in item responses. Confirming that items associated with higher levels of the LP are also more difficult supports the basic claims of the LP.

A limitation of this approach is that it is focused on the item side of the data, whereas LPs ultimately make claims about students and their mastery of specific levels in the LP. In other words, diagnostic measurement models in the context of an LP need to classify students dependably into one of the LP levels.

Under the IRT framework, this kind of classification of students has been achieved through variations of so-called Wright maps (Wilson & Draney, 2000) that link the continuous latent ability estimate to discrete levels in the learning progression (van Rijn, Graf, & Deane, 2014; Wilmot, Schoenfeld, Wilson, Champney, & Zahner, 2011). Specifically, for each level of the LP, cut scores on the continuous ability scale are created by locating (through the item response functions) the ability levels that correspond to certain probabilities of success on tasks associated with this level. For example, one can set a minimal probability of 65% and set a cut score for this level as the average ability levels that correspond to this probability across all items associated with this level. The validity of the LP is supported inasmuch as the order of cut scores that arises from this process corresponds to the order hypothesized by the LP (lower levels have lower cut scores) and the cut scores associated with a particular level across test forms or item groups are similar.

Cognitive Diagnostic Models

Cognitive diagnostic models, or CDMs (Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; Tatsuoka, 2009), present an alternative measurement approach for LPs. CDMs are latent class models (Haagenars & McCutcheon, 2002) that classify test takers into groups according to the similarity of their responses to test items. The latent classes are defined by attributes or skills that are associated a priori with test items. Each attribute has two possible values, *mastery* or *nonmastery*, and each item is defined in terms of the combination of attributes that are required to perform successfully on the item. The model then probabilistically classifies test takers in terms of their latent pattern of mastery over all attributes from their observable pattern of responses.

As compared with common psychometric models (e.g., IRT), CDMs are based on categorical instead of continuous latent variables. As a result, CDMs directly estimate the skill profile of test takers, as opposed to the two-step IRT approach described above: estimate continuous scores and then discretize them based on (somewhat subjectively defined) cut scores.

The *attribute hierarchy method* (AHM; Leighton, Gierl, & Hunka, 2004) is a special type of CDM that defines attributes that have an ordered, hierarchical relationship, in contrast to most CDMs, which assume that attributes are independent or nonhierarchical. This feature makes the AHM an appealing candidate for the modeling of LPs, which also makes explicit the hierarchical distinctions in student understanding as it becomes more sophisticated (Briggs & Alonzo, 2012).

A Framework for Testing Learning Progression Classifications

To date, most efforts around LPs have focused on their theoretical development. Although several authors have illustrated different types of validation analyses and report preliminary results (Briggs & Alonzo, 2012; Gotwals & Songer, 2013; Steedle & Shavelson, 2009; van Rijn et al., 2014; West et al., 2012; Wilmot et al., 2011), there have been no attempts to empirically test the validity of classifications from an LP in a comprehensive way.

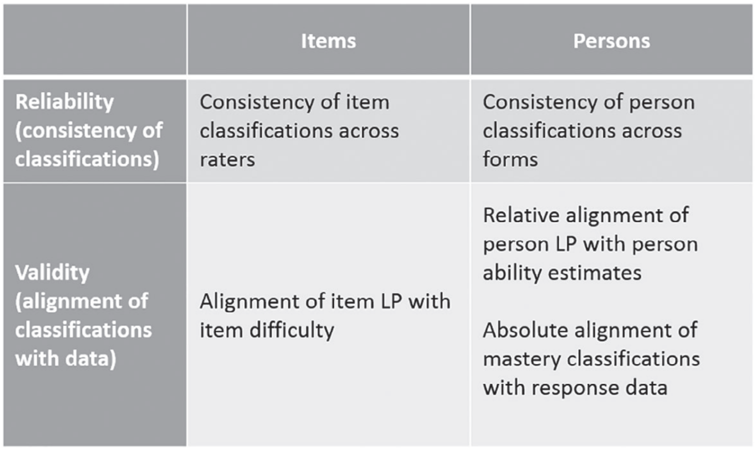
To facilitate and inform our own efforts to support the use of LPs in instructional planning and formative assessment, a framework for testing classifications from an LP was developed. This framework focuses on validation activities that should be performed following an operationalization of the LP — an assessment development, administration, and analysis effort that follows the development of the LP itself. In particular, the context for this framework was a data collection effort where assessment items were developed and classified as measuring specific LP levels, students were administered these items, and through the application of a measurement model, were classified as masters of specific LP levels. In other words, the LP theory has been operationalized and is being used to classify students into LP levels for instructional or remedial purposes.

Supporting the use of test scores is the purpose of test validation. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) defines validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses” (p. 11). These standards, following modern validity theory nomenclature (Kane, 2006; Messick, 1995), refer to different types of validity evidence (rather than different types of validities). Two types of validity evidence are particularly relevant to the development of an LP as a theory of learning and development of test items as measures of LP levels: evidence based on response processes and evidence based on content (Leighton, 2004; National Research Council, 2001). Evidence based on response processes is at the heart of an LP as a theory of learning, supported by theoretical and empirical analyses of cognitive reasoning processes by students (e.g., Clements & Sarama, 2004; Confrey & Maloney, 2010; Gotwals & Songer, 2013; Weaver & Junker, 2004). Evidence based on test content is at the heart of evaluating the alignment between the LP and its levels and between items developed to operationalize it and can include logical analyses or expert judgment evaluations of the relevance of item content to the LP (e.g., Arieli-Attali & Cayton-Hodges, 2014).

These two types of validity evidence have been the focus of previous LP development efforts but do not address the evidence in support of classifications of students to LP levels. These classifications are the primary interpretation or use of a particular LP measurement, and we therefore need to gather evidence to evaluate the soundness of the proposed interpretation for its intended use (AERA, APA, & NCME, 2014, p. 19).

The purpose of the proposed framework is to organize evidence for the validity of LP classifications. The framework (see Figure 1) organizes evidence in two ways. First, both item classifications and person classifications should be evaluated. Second, both the reliability and validity of classifications should be examined. Both the item analyses (first column) and reliability analyses (first row) should be viewed as a necessary requirement for the soundness of person classification: Item classifications are necessary to develop person classifications, and reliability of classifications is necessary for validity of classifications. In other words, although the ultimate goal is to validate person classifications (see Figure 1, lower right cell), the potential to gather such evidence would be limited without supporting evidence from the other cells of the framework.

Reliability of classifications refers to their consistency. The term is used here, in a more general sense than *reliability coefficients*, to refer to the consistency of scores across replications of a testing procedure (AERA, APA, & NCME, 2014, p. 33). Although reliability and validity are often discussed separately, in fact reliability is closely related to validity because



	Items	Persons
Reliability (consistency of classifications)	Consistency of item classifications across raters	Consistency of person classifications across forms
Validity (alignment of classifications with data)	Alignment of item LP with item difficulty	Relative alignment of person LP with person ability estimates Absolute alignment of mastery classifications with response data

Figure 1 Framework for validation of learning progression classifications.

it ultimately bears on the generalizability and dependability of scores (Kane, 2006). Classification of items into LP levels is typically assessed through measures of interrater reliability of item classifications. Classification of persons into LP levels can be assessed through measures of consistency in person classifications across different sets (or forms) of items. An appropriate statistical measure of consistency for both types of analysis is Cohen's weighted kappa (Cohen, 1968) because it takes into account chance agreement levels and weighs the severity of disagreements.

Validity of classifications refers to predictions made by these classifications concerning the ordering of items and students. Specifically, these predictions should be supported by examining their alignment with empirical data. From the item side, item LP levels are expected to align with item difficulty—items with lower LP levels should be easier than items with higher LP levels. From the person side, person LP levels are expected to align with person test performance—persons with lower LP levels should have lower performance than persons with higher LP levels. Test performance may refer to the LP assessment or to other relevant external measures of performance. These types of analyses support the relative alignment of person classifications (see Figure 1, lower right cell) because they entail a claim about the relation between LP levels and test data. Thus, they can suitably be performed through correlational or analysis of variance methods.

However, person classifications also entail an absolute claim about mastery of specific LP levels (see Figure 1, lower right cell). A person at a specific LP level is expected to master the knowledge and skills necessary to answer assessment tasks that measure this level or any lower level of the LP. Test data therefore should reflect this expectation by showing that items classified at each LP level are answered correctly by a large majority of persons classified at this level or higher. Conversely, data should show that items classified at each LP level are answered correctly at a substantially lower rate by persons classified below this level. Because definitions of "large majority" are necessarily subjective, qualitative or graphical analyses should support this type of alignment. However, odds of at least 2 (67% correct) to 4 (80% correct) are reasonable thresholds for defining this term.

Previous research attempts to empirically test the validity of classifications from an LP considered some of the aspects of the proposed framework, usually focusing on the item side. The reliability of item classification was analyzed through interrater reliability analyses (Wilmot et al., 2011). The validity of item classifications was analyzed through the application of Wright maps, where the position of items on the continuous latent trait scale is plotted to examine (in a qualitative way) whether items from different LP levels are clustered as expected from the LP (e.g., Gotwals & Songer, 2013; Wilmot et al., 2011). The reliability of person classifications was addressed by van Rijn et al. (2014, Table 4) through an analysis of the agreement of person LP classifications among different assessment form pairs. The validity of person classifications was partially addressed by Steedle and Shavelson (2009, Table 4) in an analysis of the probability of showing an LP level on an item conditioned on students' LP level classification. It is clear from this review that previous research addressed different aspects of validation, but no comprehensive validation effort of LP classifications has been implemented.

Note that the suggested framework focuses on validation of the classifications from an LP and is therefore a validation of a particular operationalization of the LP in the context of a specific assessment. Therefore, it is possible that a specific operationalization (assessment system) would yield limited support for the theory, whereas another would yield more support.

The framework presented here is particularly relevant for a context in which test data are collected during a single point in time from a group of students who are expected to vary in their LP levels. Other contexts may require a somewhat different validation plan. One such context is a contrasting groups study, where student groups who are known to differ in their LP levels (e.g., because of differences in instruction) are compared. In this context, a direct comparison of the LP levels of the two groups would be called for, in addition to validating LP classifications within each of the groups. Another context is a longitudinal study of a group of students whose LP levels are measured before and after they learn the knowledge and skills associated with the LP. In this context too, a direct comparison of the LP levels of the students before and after instruction would be called for in addition to validating LP classifications within each of the groups.

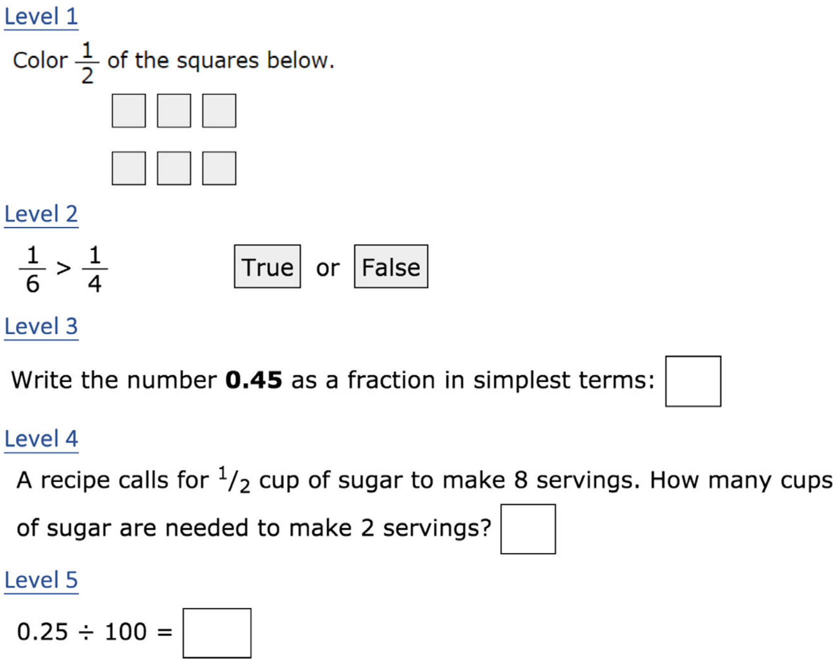
A Learning Progression for Rational Numbers

This framework was applied to an LP of rational numbers. Rational numbers have been given special attention in mathematics education research (e.g., the Rational Number Project at the University of Minnesota; see Post et al., 1998) due to their centrality in the learning of mathematics in elementary grades and their predictive value for future mathematics learning, specifically high school algebra learning (Siegler et al., 2012).

The rational number system poses new challenges for the young learner who is familiar only with the whole number system. In addition to the challenge of acquiring a new symbol system, there are difficulties in understanding that there can be different representations of the same quantity (equivalent fractions, as well as the equivalency between fractions and their corresponding decimals; Moss, 2005). The operation of translation between representations or even ordering rational numbers (fractions and decimals) does not come easily for students of all ages (Markovits & Sowder, 1991, 1994; Moss, 2005; Sowder, 1995). Most profoundly, it is difficult for students to grasp that there are different meanings (subconstructs) of the same fraction (Behr, Harel, Post, & Lesh, 1993; Behr, Lesh, Post, & Silver, 1983; Behr & Post, 1992; Kieren, 1995). For example, if a student previously knew that the number 2 stands for a group of two objects, now the symbol 2/5 is (a) part/whole (i.e., two pieces of pizza out of five pieces); (b) division (i.e., two items divided between five people); (c) ratio (i.e., 2 to 5 ratio); (d) a measure (i.e., 0.4; fixed quantity, number-line representation); and (e) multiplicative operator (i.e., operator that reduces [2/5] the size of another quantity). It is argued that introducing the rational number system to students requires a shift in understanding from an absolute or concrete way of perceiving the number system to a more relative perception of numbers (e.g., relative to a “whole”), along with acquiring new concepts such as the density of numbers (between each two fractions there is another fraction)—a precursor for continuous versus discrete concepts in later mathematics, and the multiplicative relationship between numbers (rather than additive, as in the whole number system; Moss, 2005). These factors and others make it difficult to master the rational number system, yet these are exactly the factors that explain why it is so important for middle-school students to master rational numbers before progressing to algebra and higher mathematics.

The rational number learning progression that is the focus of this study was developed by Arieli-Attali and Cayton-Hodges (2014) on the basis of research on how students think about the main ideas of rational numbers, what strategies they use—and how these ideas and strategies change over time. The LP spans the curriculum that is normally taught in Grades 2–5 but does not depend on a particular curriculum. An important characteristic of the structure of this type of learning progression is that it does not contain reference to age or grade level expectations but rather describes a sequence of increasing expertise depending on cognitive development (Heritage, 2008). This particular LP follows the conception and structure of learning progressions that are more prevalent in science education (e.g., Corcoran et al., 2009)—that is to say, a progression that covers conceptual understanding through a small number of wide grain-size levels. However, this structure of LPs can also be found in mathematics assessment (Arieli-Attali et al., 2012; Graf & Arieli-Attali, 2015; Graf & van Rijn, 2016; Kalchman & Koedinger, 2005; Kalchman, Moss, & Case, 2001; Wilmot et al., 2011). The benefits and usefulness of such a wide grain-size progression are particularly evident for diagnostic and remedial purposes as students enter middle school and should be ready for algebra.

A notable aspect of this rational number LP is that it combines fractions and decimals under one progression. Research in mathematics education tends to examine rational numbers separately for fraction understanding and decimal learning, viewing the decimal numbers primarily as a notation system. Hence, several groups of researchers have developed elaborated LPs for fraction understanding (e.g., Clements & Sarama, 2004; Confrey & Maloney, 2010; Kalchman et al., 2001), whereas in the decimal literature, researchers primarily discuss common misconceptions and errors due to the notation and representation system, with some suggestions for possible progressions or acquisition steps (e.g., Roche,



Level 1

Color $\frac{1}{2}$ of the squares below.

□ □ □

□ □ □

Level 2

$\frac{1}{6} > \frac{1}{4}$ True or False

Level 3

Write the number **0.45** as a fraction in simplest terms: □

Level 4

A recipe calls for $^1/_2$ cup of sugar to make 8 servings. How many cups of sugar are needed to make 2 servings? □

Level 5

0.25 ÷ 100 = □

Figure 2 Example items from each learning progression level.

2010; Sackur-Grisvard & Leonard, 1985). However, numerous studies have shown that ideas about fractions and decimals go hand in hand, and separating them for the purpose of teaching may not be ideal (e.g., Confrey, 1994; Moss & Case, 1999; Resnick et al., 1989; Thompson & Saldanha, 2003).

Intertwining fraction and decimal concepts has a developmental value. From a developmental point of view, once students grasp a particular concept within the fraction system or the decimal system, they are ready to learn (or implicitly already perceive) that concept in the other number/notation system. For instance, once a student perceives the concept of a fraction as a "measure" (i.e., a number on a number line), one of the manifestations of this understanding is that the student can "translate" from that fraction to its corresponding decimal (e.g., 2/5 is another name for 0.4 and he or she will be able to find 2/5 on a number line not only divided by 5 but also divided by 10). However, because learning depends on both maturation and instruction, it is possible for a student to show evidence of one level with fractions and a lower level with decimals because that student has not yet received decimal instruction (Arieli-Attali & Cayton-Hodges, 2014). The notion, however, is that the cognitive underpinnings necessary to achieve standing on a level may be present, even if instruction has not yet made possible the ability to perform at that level with both fractions and decimals.

The LP is presented in detail elsewhere (Arieli-Attali & Cayton-Hodges, 2014), but a summary of the LP appears in the appendix, and example items from each level are presented in Figure 2. The progression starts with the part-whole conceptual understanding of both fractions (using area representation, such as pizza pies) and decimals (using the money system, for example), develops into the concept of a fraction/decimal as a number in a continuum (e.g., on a number-line representation) and the ability to translate between them (grasp the equivalency between different representations of the same value), and later becomes the ability to operate fluently with fractions and decimals. The LP includes six levels. Level 0, which is the most intuitive level where no fractional understanding yet exists apart from a basic conception of halving and doubling, often characterizes students in the beginning elementary grades. Five more levels follow that describe conceptual development through Grades 2–5.

Empirical Study

The purpose of the empirical study reported here was to apply the framework presented above for the validation of classifications made according to an LP for rational numbers. As a specific application of this framework, several design decisions were made, three of which are noted here. First, we decided to use a CDM (specifically, the AHM) as the measurement model for classification of students into LP levels. As a latent class model, the CDM is particularly suitable as a measurement model for the categorical LP levels, but alternative models could have been used, including Bayesian networks (West et al., 2012) and IRT approaches (van Rijn et al., 2014). Second, we used state assessment scores as the external measure against which LP classifications were compared. These scores were preferred over alternatives (such as class grades) for their comprehensiveness as measures of mathematics achievement and high reliability. Third, we administered the assessment to middle school students, who already completed instruction for both fractions and decimals curricula. Due to this decision, we can also assume that lack of decimal instruction was not an obstacle in this study.

This study was based on results from a pilot study of a formative assessment system that was designed to measure mastery and fluency with concepts of the number system and operations with numbers. We classified 400 items in terms of LP level of understanding, a cognitive diagnostic model of student mastery level within the LP was fitted to the data, and analyses were conducted to assess the reliability and validity of these classifications.

The research questions were informed by the validation framework and accordingly were concerned with both the reliability and validity of LP classifications and also with both item-related issues and student-related issues. Item-related questions were as follows:

1. What is the reliability of item classifications to LP levels, as evidenced by the interrater agreement in their classifications?
2. What is the validity of item LP classifications, as evidenced by the contribution of item LP levels in explaining item response variance?

Student-related questions were as follows:

1. What is the validity of student classifications, as evidenced by the alignment between the expected responses of students according to their model-based LP levels and the actual responses of students?
2. What is the validity of student classifications, as evidenced by the relation between student LP classifications and test scores as well as external measures of mathematical ability?
3. What is the reliability of student LP classifications, as evidenced by the relation between student classifications based on different test forms?

Method

Participants

For this study, a New Jersey middle school was recruited. The school was paid 10 dollars per participating student. All Grade 6–8 students (in almost equal numbers per grade) in the school participated, 693 in all. Girls and boys were 50% each, and 48% were Asian, 31% were White, 12% were Hispanic, and 8% were Black. The average mathematics ability of the students, as indicated by their state assessment scores, was relatively high: 258, 249, and 234 for Grades 6–8, respectively. The corresponding standardized differences for the average New Jersey student were .51, .54, and .43 (calculated with data from the New Jersey Department of Education, 2014, p. 67). In other words, the average school scores were about half a standard deviation above the state mean.

Instruments

The test items that were used in this study are part of Quick Math (Attali & Arieli-Attali, 2015a, 2015b), an adaptive practice and assessment system of elementary and middle school mathematics. This system is itself part of the *CBAL®* research initiative to develop assessments that maximize the potential for positive effects on teaching and learning (Bennett, 2011). Quick Math is used to assess and strengthen procedural and representational fluency with concepts of the number system and operations with numbers.

Questions can be generated on the fly from item models (Bejar, 1993), which are problem schemas having parameters that can be instantiated with specific values. For example, the model "X + Y =? ", where X and Y can be whole numbers in the range 1–10, has two parameters that can be instantiated to display an actual exercise.

From a practical standpoint, the use of item models greatly expands the potential number of exercises while at the same time allowing repetition, the cornerstone of practice. From a theoretical standpoint, item models provide an opportunity for a construct-driven approach to item development because they can be tied to a mapping of the construct through an analysis of the cognitive mechanisms related to item solution and item features that call on these mechanisms (Embretson, 1983). In this context, these item models constitute foundational tasks that form the basis for understanding number properties and operations. In particular, many of the item models in Quick Math (Attali & Arieli-Attali, 2015a, 2015b) were focused on rational numbers and have been informed by the work on the rational number LP (Arieli-Attali & Cayton-Hodges, 2014). Other item models focused on whole number concepts and understanding.

The set of items used in this study comprised 50 item models with eight instances from each model, with a total of 400 items. Instances were generated to cover a representative range of model parameters.

Design

The complete set of 400 items was divided into four nonoverlapping sets (or forms), with each set composed of two instances from all 50 item models. Each set was further divided into two subsets with 25 item models (and two instances) in each. Each participant was randomly assigned to one of the sets and was further randomly assigned to take one of the subsets in the first test session and the other in the second test session.

Procedures

Teachers were asked to schedule two test sessions for each class at their convenience. Each test session lasted one class period (the median time to complete each session, including instructions and posttest questions, was 20 minutes) and was conducted in a computer lab. The two sessions were completed within 1 day of each other by most students (70%) and within 5 days by almost all students (except a few make-up sessions).

Analyses

First, we report descriptive statistics for response accuracy and response time, including Cronbach's alpha reliability coefficients and validity coefficients with respect to students' state assessment scores. Next, we report interrater agreement results for the classification of the items to the LP levels.

Third, we explore the potential of LP levels to predict response accuracy and response time and explain the variation across item models and instances within models. To do that analysis, the results of unconditional linear mixed models were compared with those that include the LP levels covariate. The mixed models included crossed random effects for person and item models, as well as instance-within-model. For response accuracy, a generalized linear model with the Bernoulli distribution and logistic link function was used. This model is similar to a one-parameter IRT model, except that items (models and nested instances) were treated here as random effects rather than fixed effects (see De Boeck, 2008). Treating items as randomly selected from a population naturally fits the concept of item models and instances drawn from it (as is used in Quick Math, Attali & Arieli-Attali, 2015a, 2015b) but also presents several measurement advantages (De Boeck, 2008).

The linear component of the model is a linear combination of predictors:

$$\eta_{pi} + \theta_p + \beta_i$$

where $\theta_p \sim N\left(0, \sigma_\theta^2\right)$ is the random person ability effect and β_i is the item easiness effect:

$$\beta_i = \beta_L X_{iL} + \varepsilon_i$$

where β_L is the weight of the LP level (0–5), $X_{iL}$is the value of item I on the LP level property, and $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$ is the random residual item variance. This model is referred to as the *linear logistic test model* with error (De Boeck, 2008).

To estimate the contribution of the LP levels to the prediction of item variance, the percent reduction in estimated item (both model and instance) variance from the unconditional to the conditional model was calculated.

Finally, we estimated an AHM from the response accuracy data for each form using the R package CDM (Robitzsch, Kiefer, George, & Uenlue, 2016). Formally, a deterministic-input, noisy-and-gate model (see Rupp et al., 2010, p. 116) with a reduced or restricted skill space was used. The LP levels (0–5) were defined as attributes, or skills, which resulted in six attributes for the Q-matrix. To account for the hierarchical nature of the LP, a reduced skill space was defined (Leighton et al., 2004) that corresponded to a linear attribute hierarchy (Gierl, Leighton, & Hunka, 2007), where the attributes were sequentially ordered in a single chain. In such a chain, knowing that a student has mastered a particular attribute (say, the one corresponding to Level 4 in the LP) implies that the student has mastered all preceding attributes in the chain (Levels 0 through 3). Formally, the reduced skill space is defined through a 6x6 lower triangular matrix (that is, all entries below and on the main diagonal are 1). In this matrix, if $i > j$, then the entry ij is 1, signifying that if a student masters Level i in the LP, then he or she should also master all levels lower than i. As a consequence, instead of the full space of 2^6 patterns of skills, only seven skill patterns are possible: Students can be classified as masters of Level 5 (which entails they master all lower levels too), or Level 4, or Level 3, or Level 2, or Level 1, or Level 0, or nonmasters of Level 0 (for examples of setting up this model in the CDM package, see Robitzsch et al., 2016).

The models were evaluated in terms of their absolute model fit by computing a range of indices proposed in the literature. Two indices are based on summaries of the discrepancies between observed and expected (according to the model) pairwise item response correlations: The statistic MADcor denotes the average absolute deviation between observed pairwise item correlations and model-predicted correlations (DiBello, Roussos, & Stout, 2007), and the SRMSR (standardized root mean square root of squared residuals; Maydeu-Olivares, 2013) is also based on comparing these correlations. It is generally recommended that values of .05 or lower indicate excellent fit (Maydeu-Olivares, 2013).

A related index based on the difference of Fisher transformed correlations was proposed by Chen, de la Torre, and Zhang (2013). For each item pair, the transformed correlation and its standard error are derived to compute a z-score (denoted X2) to test whether the residuals differ significantly from zero, applying the Holm procedure to adjust for multiple comparisons. The authors proposed the use of the maximum of these X2 measures as a global test of absolute model fit (indicating whether or not at least one item pair does not fit the model). However, due to the large number of items in our data (and a very large number of item pairs), we report on and analyze instead the number of significant item pairs.

The validity of student classifications to LP levels based on the models was analyzed by (a) calculating convergent validity relationships with test scores and state assessment scores, (b) conducting a graphical analysis of the alignment between the expected responses of students according to their model-based LP levels and the actual responses of students, and (c) calculating the reliability of student LP classifications from two submodels, each based on half of the items.

Results

Descriptive Results

Table 1 presents psychometric properties of response accuracy and response time, or slowness (natural log of time in seconds to answer). The average percentage correct score was 68%, and average mean log response time was 2.58 (corresponding to 13 seconds). The internal consistency of response accuracy was very high, with Cronbach's alpha coefficients around .97 for the 100-item scores and slightly lower for the 50-item-model scores (each score the sum of two instances from each item model). The internal consistency of response time was similar, with Cronbach's alpha coefficients of .96 for the 100-item scores and slightly lower for the 50-item-model scores.

The validity coefficients of the accuracy scores (number of correct responses across the two sessions) with respect to the students' state assessment scores (obtained a year before the study) were high: .84, significantly higher than the correlation between students' mathematics grades for the entire year and state assessment scores and .68 ($p < .01$ for the test of correlated correlations). The validity coefficients of the slowness scores (average of log seconds) with respect to

Table 1 Psychometric Properties of Response Accuracy and Response Time

		Percentage correct				Average response time (log sec.)			
Form	*N*	*M*	*SD*	α items	α models	*M*	*SD*	α items	α models
1	172	0.680	0.203	0.967	0.960	2.601	0.349	0.962	0.957
2	178	0.668	0.200	0.966	0.958	2.561	0.359	0.963	0.957
3	173	0.678	0.199	0.966	0.958	2.586	0.357	0.963	0.957
4	170	0.710	0.186	0.962	0.955	2.574	0.350	0.962	0.958
Total	693	0.684	0.197			2.580	0.353		

Note. All forms share the same 50 models with two different instances for each test.

Table 2 Generalized Linear Mixed Model Results

	Response accuracy				Response time			
Parameter	Model 1		Model 2		Model 1		Model 2	
Random variance estimates								
Person	2.274	38%	2.275	59%	0.118	13%	0.118	16%
Model	3.190	54%	1.202	31%	0.259	28%	0.103	14%
Instance	0.465	8%	0.358	9%	0.062	7%	0.046	6%
Residual	—		—		0.476	52%	0.476	64%
Total	5.929		3.834		0.915		0.742	
Fixed-effect estimates								
Intercept	1.328	(0.279)	3.643	(0.271)	2.566	(0.079)	1.760	(0.083)
LP			−0.781	(0.068)			0.271	(0.022)

Note. Percentages for random effects are out of total variance for model. Standard errors for fixed effects in parentheses. The R package *lme4* does not compute standard errors for random effects. No residual term for accuracy (binomial distribution). LP = learning progression.

the students' state assessment scores were lower, −.38, but as expected, indicating that faster response is associated with higher levels of mathematical proficiency.

Classification of Items to Learning Progression Levels

Two content experts familiar with the LP independently categorized each of the 400 items in terms of the lowest level of understanding in the LP (Levels 0 – 5) necessary to answer the item correctly or whether the LP did not apply to the item (which in this context meant in practice that the item addressed whole number concepts). The agreement between the raters was high, with quadratic-weighted kappa of .90 (*SE* = .02). Around one fifth of the items were classified by one or both raters as not applicable to the LP. A similar agreement level was found for the items that both raters agreed were applicable to the LP, with quadratic-weighted kappa of .86 (*SE* = .03). The raters proceeded by discussing disagreements and adjudicating them. For the final decisions, the percentage of items falling under the categories not applicable and Levels 0 through 5 were 21%, 11%, 3%, 6%, 19%, 28%, and 12%, respectively. In the following analyses, we focus on the 316 items that were classified as applicable to the LP.

Learning Progression Levels as Predictors of Variability in Item Difficulty and Response Time

In this section, we explore the potential of LP levels to predict response accuracy and response time and explain the variation across item models and instances within models. To do that exploration, the results of unconditional linear mixed models are compared with those that include the LP levels as a covariate. The mixed models include crossed random effects for person and item models, as well as instance-within-model. For response accuracy, a generalized linear model with the Bernoulli distribution and logistic link function was used. This model is similar to a one-parameter IRT model, except that items (models and nested instances) were treated here as random effects rather than fixed effects.

Table 2 shows variance components and fixed-effect estimates for the unconditional models and the LP models. With respect to response accuracy, inclusion of the LP reduced item model variance by 62% (from 3.190 to 1.202) and instance variance by 23%. Overall item variance (model and instance) was reduced by 57%. The results for response time were very similar: Inclusion of the LP reduced item model variance by 60% (from 0.259 to 0.103) and instance variance by 27%. Overall response time item variance (model and instance) was reduced by 54%. In summary, for both response accuracy and response time, LP levels explained more than 50% of the item variance.

Cognitive Diagnostic Model of Student Mastery Level

Table 3 presents absolute model fit for the CDMs developed for each form. The table shows that MADcor values (.08 – .13) and SRMSR values (.10 to .13) fell short of the .05 threshold for excellent fit but were not far from it. The number of significant residual item pair correlations was small (0.4% of item pairs), and an inspection of the item pairs showed that

Table 3 Model Fit Indices for Cognitive Diagnostic Models

Form	Items	*MADcor*	*SRMSR*	Sig. X2
1	82	0.081	0.104	15
2	76	0.089	0.112	8
3	82	0.108	0.135	24
4	76	0.088	0.108	4

Note. Sig. X2 = number of item pairs with significant X2 (α = .05). Total number of item pairs is 3,321 for forms 1 and 3, and 2,850 for forms 2 and 4.

almost half of these cases involved item instances that belonged to the same item model. This result suggests that part of the lack of fit was due to the use of multiple instances for each model.

To determine the LP level of students, posterior probabilities for the level classification of each student given their responses were obtained. Most students were unambiguously classified into only one level—75% were assigned to one of the levels with at least 89% probability (that is, with at most 11% probability for all other levels combined), and 90% of students were assigned to one of the levels with at least 66% probability. One way to assign students to LP levels is to select the level with the highest probability from the model (maximum a posteriori). An alternative that better accounts for possible uncertainty in probabilities is to compute the weighted average of LP levels with posterior probabilities as weights and round the result to the nearest whole number (expected a posteriori). In our data, both methods resulted in exactly the same LP assignments.

The most frequent student LP level was Level 5, with 51% of students classified into it. The percentage of students classified into Levels 4 to 0 was 10%, 11%, 4%, 10%, and 11%, respectively, and 3% of students were classified as lower than Level 0.

Validity and Reliability of Student Learning Progression Classifications

One way to test the classifications made by an LP is to show the convergent validity of student LP classifications with both internal (test scores) and external measures. Figure 3 presents the relation between LP level classification and percent

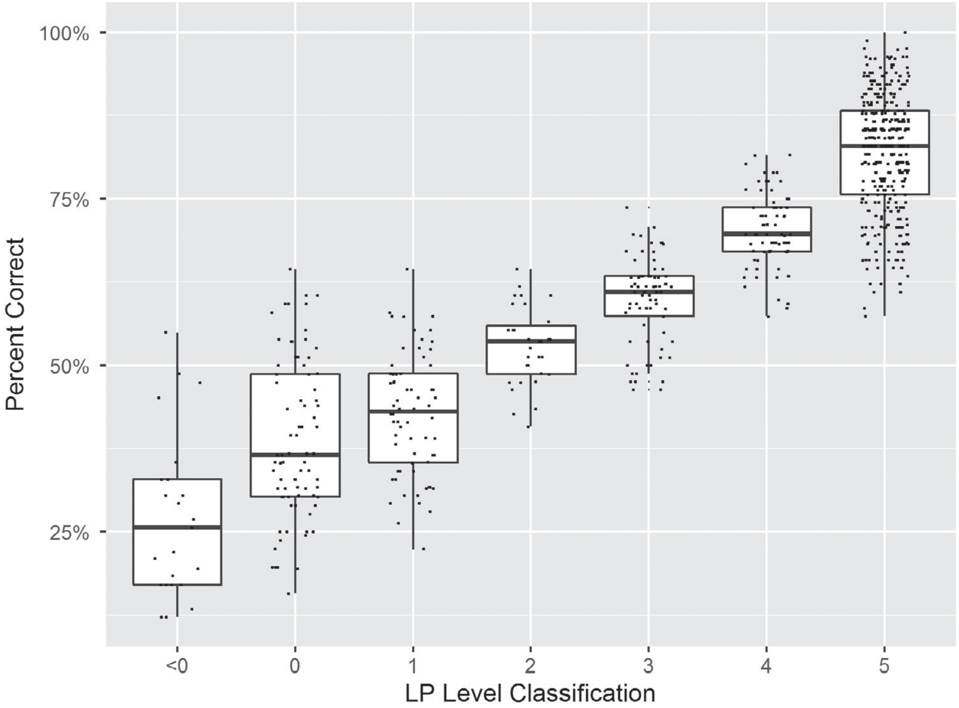


Figure 3 Box plots of percent correct by learning progression level classification.

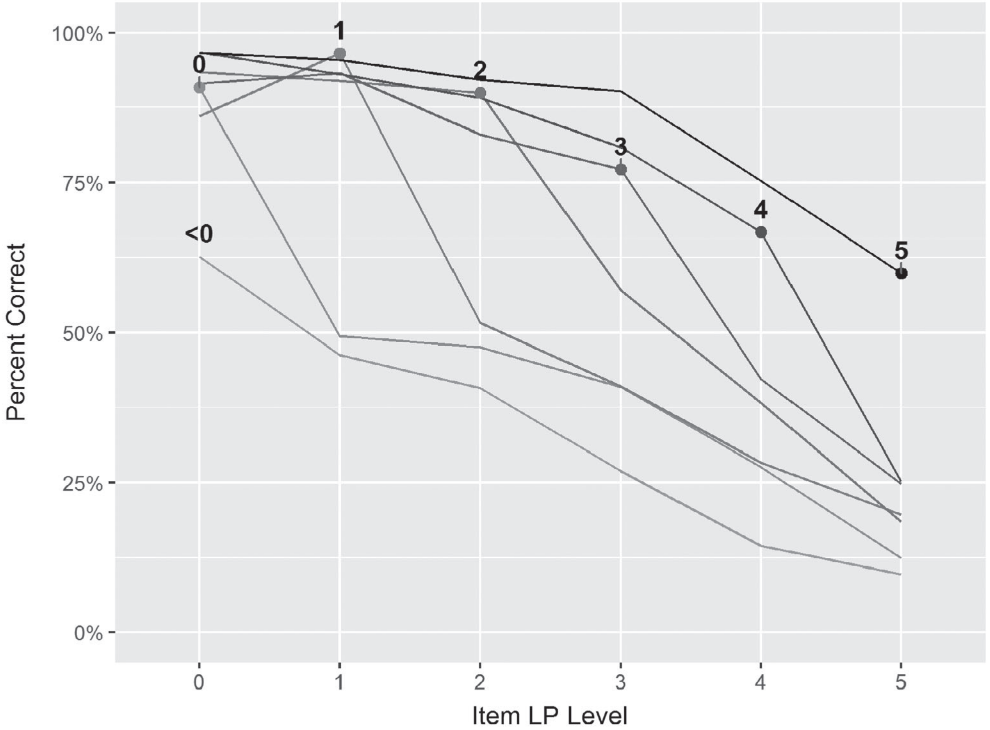


Figure 4 Percent correct across item learning progression levels and student learning progression levels (labeled lines).

correct responses on the test. As expected, the correlation between the two variables was high, .89. Similarly, the correlation between the LP student classifications and the students' state assessment scores was .74.

The validity of student classifications can also be supported by showing an alignment between the expected responses of students according to their model-based LP levels and the actual responses of students. For example, students classified at Level 4 are expected to answer correctly a high proportion of items at Level 4 or lower and a lower proportion of items at Level 5. Figure 4 presents the rates of correct answers for each combination of student LP level and item LP level. The figure shows that, in general, success of students in answering items from different LP levels aligned well with the students' LP level. Rates of correct responses were high in cases where the students' LP level equaled or exceeded the items' LP level, and these rates were significantly lower when students' LP level was lower than the items' LP level.

For example, for items with LP Level 0 (leftmost column of data), rates of correct responses for students with LP Levels of 0–5 were all higher than 85%, but it was 62% for students who did not reach Level 0 (<0). For items in the next level (Level 1), rates of correct responses were around 90% for all student levels greater than 0 and around 50% for student Levels 0 and less than 0. The same pattern can be seen for higher item LP levels and is summarized in Figure 5, which shows a gap of around 40% in the rates of correct answers for students below, versus at or above, the item LP level.

Although this gap supports the validity of the LP, Figures 2 and 3 also show a downward trend of the lines for higher item LP levels, resulting in at or above rates for the highest item LP, Level 5, that is not absolutely high (60%). In other words, some students at Level 5 were not answering correctly a large majority of the items they were supposed to master.

The reliability of student LP classification was estimated by splitting the items students answered into two groups, with one instance of every model in one group and the second instance in the other (excluding as before the items judged as not applicable to the LP). CDMs were then estimated separately for each group of items, and two LP levels were calculated for each student from the posterior level probabilities of the models. The quadratic-weighted kappa for the two classifications was .78 (and the correlation between the classifications was .80), indicating a high level of consistency in classifications.

Discussion

As theories of learning, LPs should be supported by multiple types of evidence for their validity. Evidence based on response processes and evidence based on content are particularly relevant to the early stages of the development

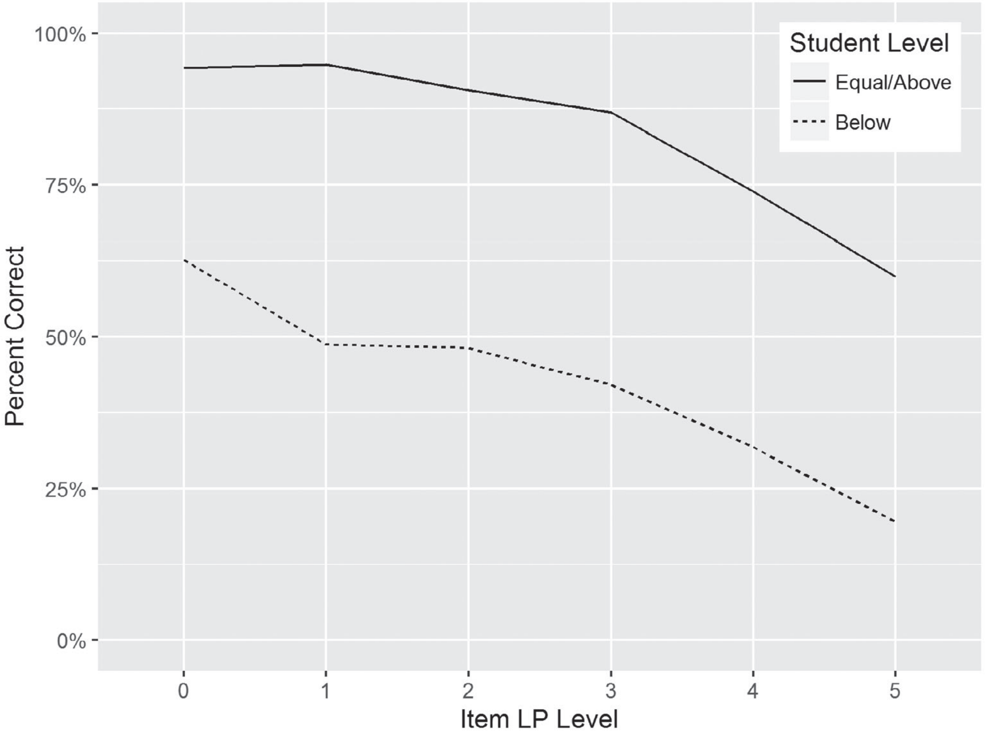


Figure 5 Percent correct for student learning progression levels below or equal/above item learning progression level.

of an LP. However, because a primary use of the LP is to classify students into LP levels through a particular measurement for diagnostic and instructional purposes, we therefore need to gather additional validation evidence to evaluate the soundness of these classifications.

In this paper, we propose a framework for validating classifications from an LP and reports results from an implementation of the framework to validate classifications made according to a rational number LP for elementary school mathematics, administered to middle school students. The framework calls for validating both item and person classifications, and both the consistency (reliability) of classifications as well as the alignment (validity) of classifications with test data. The empirical study included evidence on all four aspects of the framework with a relatively large sample of students and items. Our main findings indicated that the CDM fitted the response data reasonably well, and good alignment was found between the expected responses of students according to their model-based LP levels and the actual responses of students. In addition, we found a strong relation between student LP classifications and both test scores and external measures of mathematical ability (state assessment scores). Finally, we found high levels of reliability of student LP classifications.

Although results generally supported the validity of classifications, the following discussion highlights limitations and issues that could be addressed in future research. As several authors have noted (Briggs & Alonzo, 2012; Steedle & Shavelson, 2009; van Rijn et al., 2014; West et al., 2012; Wilmot et al., 2011), the process of validating an LP is iterative, and intermediate findings should be used to revise the LP. Our findings suggest one such revision. Specifically, the lower levels of success for students at Level 5 on items at Level 5 (60% correct, compared to 80% or more for lower levels), suggests that Level 5 is too widely defined. In fact, in the process of rating items to levels, the raters noted that some items seem to "exceed Level 5" (not aligned with other Level 5 items), but because this level is the highest level defined by the progression, they assigned those items to Level 5. In the current LP, Level 5 includes understanding of the multiple facets of fractions, with specialization and generalization across contexts as well as decontextualization from complex word problems to equation and expressions. This decontextualization has been found to be most difficult for students (Kilpatrick, Swafford, & Findell, 2001). For example, an item such as "If 1 Euro is worth 1.5 U.S. dollars, then 1 U.S. dollar is worth ___ Euro." requires that the student decontextualize the question (i.e., extract the appropriate mathematical model before solving the actual math problem). Indeed, items from this model were found to be among the most difficult items in the item bank. These results (and theory supporting them) suggest that the LP may benefit from dividing Level 5 into two levels.

An important aspect of the validation study is that it was based on a middle school sample. Because the rational number LP targets curriculum that is taught in earlier grades, students were expected to master the knowledge and skills covered by the LP (i.e., the fractions and decimal concepts). Therefore, the fact that 50% of the students were classified in the highest level of the LP should not necessarily be interpreted as high. The results thus highlight the importance of remedial instruction on the topic of rational numbers during middle school. For example, if a student was diagnosed at Level 3, the LP specifies what students at this level know and can do and, specifically, what they still need to acquire (Arieli-Attali & Cayton-Hodges, 2014). Students' classification can be used directly by teachers in determining the remedial work necessary for groups of students. Interestingly, in our study, students from higher grades did not necessarily perform at higher levels, and the (cross-sectional) grade effect on test scores was not significant (average percent correct across Grades 6–8 was 66%, 68%, and 66%, respectively). This outcome may have been the result of the decreased emphasis in the curriculum on rational numbers (and increased emphasis on algebra) during the middle grades.

However, a replication of this study with a sample of students from lower grades is called for to further support the rational number LP, because it was originally designed for curriculum learned in lower grades. Another possibility is to collect longitudinal data that tracks student development across a long enough time period to observe the transition for individuals across progression levels. This type of data collection would provide a direct link between the LP and individual learning (as opposed to a cross-sectional data collection). From a psychometric point of view, a replication with a lower performance sample of students raises other issues. In principle, the CDM results should not be affected by a lower performance sample. In other words, similar to an IRT model, the model parameters should be invariant to the performance distribution of the student sample. In addition, lower performance would entail more evenly distributed LP levels and therefore could increase the measurement accuracy of students categorized with lower LP levels. However, a lower performance sample could also indicate a larger proportion of students who are in transition through the LP levels. Response patterns of such students could present a different kind of challenge to the validity of the LP. A related issue is the relatively small number of items at Levels 2 and 3 compared to other levels (3% and 6%, respectively). Although this lack of items did not seem to affect the classification results of students at these levels (see Figures 3 and 4), it is not clear how this would have played out with a lower ability student sample, and therefore should be addressed in future research.

The results of the particular implementation of the LP using Quick Math (Attali & Arieli-Attali, 2015a, 2015b) demonstrated that the LP classifications can be used by teachers (and students) to assess students' level of understanding with rational numbers and readiness for algebra and target instruction or special interventions to help students overcome difficulties or gaps accrued with this foundational topic in mathematics education, specifically at the crossroad before entering middle school. An advantage of using Quick Math for this purpose is the speed with which such an assessment can be carried out—the median response time for an item was 13 seconds (see Table 1). As a result, even with the fixed-item forms administered in this study, test reliability was high for a median total testing time of just over 20 minutes. An interesting topic for future research is the implementation of adaptive testing procedures whose goal is to minimize testing time for LP classification rather than for measurement error, as in regular adaptive testing. This altered goal requires modifications to the item selection criteria and may be implemented through an IRT framework (Reckase, 1983) or through a dynamic CDM that incorporates new responses as they are accumulated.

In conclusion, this study showed the potential for using LPs to reliably classify students according to their development in mastering an important foundational topic in mathematics education, opening up the possibility of using these classifications in instructional decisions.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Alonzo, A., & Gotwals, A. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam, The Netherlands: Sense Publishers. https://doi.org/10.1007/978-94-6091-824-7

Arieli-Attali, M., & Cayton-Hodges, G. A. (2014). *Expanding the CBAL™ mathematics assessments to elementary grades: The development of a competency model and a rational number learning progression* (Research Report No. RR-14-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12008

Arieli-Attali, M., Wylie, E. C., & Bauer, M. I. (2012, April). *The use of three learning progressions in supporting formative assessment in middle school mathematics*. Paper presented at the 2012 AERA Annual Meeting, Vancouver, Canada.

Attali, Y., & Arieli-Attali, M. (2015a). Gamification in assessment: Do points affect test performance? *Computers and Education, 83*, 57 – 63. https://doi.org/10.1016/j.compedu.2014.12.012

Attali, Y., & Arieli-Attali, M. (2015b, April). *An adaptive mathematics assessment with on-the-fly item generation*. Paper presented at the AERA 2015 Annual Meeting, Chicago, IL.

Behr, M., Harel, G., Post, T., & Lesh, R. (1993). Rational numbers: Toward a semantic analysis — Emphasis on the operator construct. In T. P. Carpenter, E. Fennema, & T. A. Romberg, (Eds.), *Rational numbers: An integration of research* (pp. 13 – 47). Mahwah, NJ: Erlbaum.

Behr, M., Lesh, R., Post, T., & Silver, E. (1983). Rational number concepts. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 91 – 125). New York, NY: Academic Press.

Behr, M., & Post, T. (1992). Teaching rational number and decimal concepts. In T. Post (Ed.), *Teaching mathematics in grades K – 8: Research-based methods* (2nd ed., pp. 201 – 248). Boston, MA: Allyn and Bacon.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323 – 359). Hillsdale, NJ: Erlbaum.

Bennett, R. E. (2011). *CBAL: Results from piloting innovative K – 12 assessments* (Research Report No. RR-11-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02259.x

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 293 – 316). Rotterdam, Netherlands: Sense Publishers. https://doi.org/10.1007/978-94-6091-824-7_13

Carr, M., & Alexeev, N. (2011). Fluency, accuracy, and gender predict developmental trajectories of arithmetic strategies. *Journal of Educational Psychology, 103*, 617 – 631. https://doi.org/10.1037/a0023864

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnostic modeling. *Journal of Educational Measurement, 50*, 123 – 140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning, 6*(2), 81 – 89. https://doi.org/10.1207/s15327833mtl0602_1

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213 – 220. https://doi.org/10.1037/h0026256

Confrey, J. (1994). Splitting, similarity, and the rate of change: A new approach to multiplication and exponential functions. In G. Harel & J. Confrey (Eds.), *The development of multiplicative reasoning in the learning of mathematics* (pp. 291 – 330). Albany, NY: State University of New York Press.

Confrey, J., & Maloney, M. (2010). The construction, refinement, and early validation of the equipartitioning learning trajectory. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th annual conference of the learning sciences* (Vol. *1*, pp. 968 – 975). Chicago, IL: International Society of the Learning Sciences.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report # RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533 – 559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer-Verlag.

DiBello, L., Roussos, L., & Stout, W. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics*. Amsterdam, The Netherlands: Elsevier.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education, 47*, 123 – 182. https://doi.org/10.1080/03057267.2011.604476

Embretson, S. E. (1983). Construct validity: Construct representation vs. nomothetic span. *Psychological Bulletin, 93*, 179 – 197. https://doi.org/10.1037/0033-2909.93.1.179

Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242 – 274). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.009

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching, 50*, 597 – 626. https://doi.org/10.1002/tea.21083

Graf, E. A., & Arieli-Attali, M. (2015). Designing and developing assessments of complex thinking in mathematics for the middle grades. *Theory Into Practice, 54,* 195 – 202. https://doi.org/10.1080/00405841.2015.1044365

Graf, E. A., & van Rijn, P. W. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165 – 189). New York, NY: Taylor & Francis.

Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

Kalchman, M., & Koedinger, K. (2005). Teaching and learning functions. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics and science in the classroom* (pp. 351–396). Washington, DC: National Academies Press.

Kalchman, M., Moss, J., & Case, R. (2001). Psychological models for the development of mathematical understanding: Rational numbers and functions. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: 25 years of progress* (pp. 1–38). Mahwah, NJ: Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kieren, T. E. (1995). Creating spaces for learning fractions. In J. T. Sowder & B. P. Schappelle (Eds.), *Providing a foundation for teaching mathematics in the middle grades* (pp. 31–66). Albany: State University of New York Press.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

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. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Markovits, Z., & Sowder, J. T. (1991). Students' understanding of the relationship between fractions and decimals. *Focus on Learning Problems in Mathematics, 13*, 3–11.

Markovits, Z., & Sowder, J. T. (1994). Developing number sense: An intervention study in grade 7. *Journal for Research in Mathematics Education, 25*, 4–29. https://doi.org/10.2307/749290

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 13*, 71–101. https://doi.org/10.1080/15366367.2013.831680

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Moss, J. (2005). *Pipes, tubs, and beakers: New approaches to teaching the rational-number system*. In National Research Council (Ed.), *How students learn: History, math, and science in the classroom* (pp. 309–349). Washington, DC: The National Academies Press.

Moss, J., & Case, R. (1999). Developing children's understanding of rational numbers: A new model and an experimental curriculum. *Journal for Research in Mathematics Education, 30*(2), 122–147. https://doi.org/10.2307/749607

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

New Jersey Department of Education. (2014). *NJASK 2013 technical report: Grades 3–8*. Retrieved from http://state.nj.us/education/assessment/ms/5-8/

Post, T. R., Cramer, K., Harel, G., Kieren, T., & Lesh, R. (1998). Research on rational number, ratio and proportionality. In S. Berenson (Ed.), *Proceedings of the twentieth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, PME-NA 98* (Vol. *1*, pp. 89–93). Raleigh: North Carolina State University.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50023-8

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education, 20*(1), 8–27. https://doi.org/10.2307/749095

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). *CDM: Cognitive diagnosis modeling*. R Package Version 4.7-0. https://CRAN.R-project.org/package=CDM

Roche, A. (2010). Decimats: Helping students to make sense of decimal place value. *Australian Primary Mathematics Classroom, 15*, 4–12.

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

Sackur-Grisvard, C., & Leonard, F. (1985). Intermediate cognitive organizations in the process of learning a mathematical concept: The order of positive decimal numbers. *Cognition and Instruction, 2*, 157–174. https://doi.org/10.1207/s1532690xci0202_3

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*, 691–697. https://doi.org/10.1177/0956797612440101

Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning, 6*, 91 – 104. https://doi.org/10.1207/s15327833mtl0602_2

Song, Y., Deane, P., Graf, E. A., & van Rijn, P. W. (2013). *Using argumentation learning progressions to support teaching and assessments of English language arts* (R&D Connections No. 22). Princeton, NJ: Educational Testing Service.

Sowder, J. T. (1995). Instructing for rational number sense. In J. Sowder & B. P. Schappelle (Eds.), *Providing a foundation for teaching mathematics in the middle grades* (pp. 15 – 29). New York, NY: SUNY Press.

Steedle, J., & Shavelson, R. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*, 669 – 715. https://doi.org/10.1002/tea.20308

Stevens, S. Y., Shin, N., & Krajcik, J. S. (2009, June). *Towards a model for the development of an empirically tested learning progression*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge.

Thompson, P. W., & Saldanha, L. A. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick, G. Martin, & D. Schifter (Eds.), *Research companion to the principles and standards for school mathematics* (pp. 95 – 114). Reston, VA: National Council of Teachers of Mathematics.

van Rijn, P. W., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts. *Psicología Educativa, 20*, 109 – 115. https://doi.org/10.1016/j.pse.2014.11.004

Weaver, R., & Junker, B.W. (2004). *Model specification for cognitive assessment of proportional reasoning* (Department of Statistics Technical Report No. 777). Pittsburgh, PA: Carnegie Mellon University.

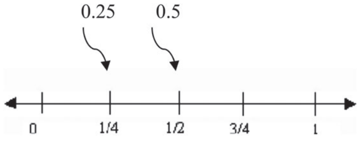
West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K. E., . . . Behrens, J. T. (2012). A Bayesian network approach to modeling learning progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 257 – 292). Rotterdam, Netherlands: Sense Publishers. https://doi.org/10.1007/978-94-6091-824-7_12

Wilmot, D., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning, 1*, 259 – 291. https://doi.org/10.1080/10986065.2011.608344

Wilson, M., & Draney, K. (2000, May). *Standard mapping: A technique for setting standards and maintaining them over time*. Paper presented at the International Conference on Measurement and Multivariate Analysis, Banff, Canada.

Appendix

Rational Number Learning Progression (LP) Summary

Level	Conceptual understanding	Fraction example	Decimal example
Prior knowledge Half and halving	The concept of half; halving or splitting into two equal parts	Breaking a cookie to share between two people	
Level 1 Early part/whole understanding	The beginning of part/whole and part/group understanding; repeated halving and equipartitioning into number of parts 2^n	1/2, 1/4, 1/8 In context	0.5 as equal to half, 0.25 is a quarter; in context (e.g., money), $2.50 means 2 and a half dollars
Level 2 Fraction as unit	The establishment of part/whole concept of a fraction; equipartitioning with all numbers Unit fraction concept and common fractions smaller than one whole (proper functions)	Unit fractions 1/2; 1/3; 1/4; 1/5; 1/6; 1/7; etc. Common fractions 2/5 conceived as 2 times 1/5	Basic unit decimals 0.01 = penny; 0.05 = nickel; 0.10 = dime; 0.25 = quarter; nonbenchmark decimals, such as $2.35
Level 3 A fraction as a single number	The shift to the concept of fraction as a single number; a number-line representation Early integration of fraction and decimal notation (benchmarks); fraction as a measure		

Appendix Continued

Level	Conceptual understanding	Fraction example	Decimal example
Level 4 Representational fluency	Smooth translating between different notations of rational numbers Established multiplication of fractions with fractions, i.e., partitioning of a partitioning (partitioning a third into fourths results in 1/12 size parts); fraction as an operator Convert smoothly between fractions and decimals	Transform 2/3 into 2/5 multiplicatively	Convert 5/8 into decimal form, in any algorithm
Level 5 General model of rational number	General model of fraction; multiplicative structure; contextual fluency; fraction as quotient Explain fraction/decimal multiplication in context Explain fraction division in context (quotative division concept)	What is a situation in which you would divide 3/4 by 5/8?	If three pounds of pasta salad cost $4.50, how much pasta salad can you buy for $6.00?

Suggested citation:

Attali, Y., & Arieli-Attali, M. (2018). *Validating classifications from learning progressions: Framework and implementation* (Research Report No. RR-19-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12253

Action Editor: Heather Buzick

Reviewers: Randy Bennett and Peter van Rijn