Visualizing and Reporting Content-Referenced Growth on a Learning Progression

April 28, 2025

Derek C. Briggs, University of Colorado Boulder

Kyla McClure, University of Colorado Boulder

Sanford Student, University of Delaware

Sarah Wellberg, University of Virginia

Nathan Minchen, Curriculum Associates

Olivia Cox, University of Colorado Boulder

Erik Whitfield, University of Colorado Boulder

Nicolás Buchbinder, University of Colorado Boulder

Laurie Davis, Curriculum Associates

Pre-print version of the publication

Briggs, D. C., McClure, K., Student, S., Wellberg, S., Minchen, N., Cox, O., Whitfield. E., Buchbinder, N., & Davis, L. (2025). Visualizing and Reporting Content-Referenced Growth on a Learning Progression. *Educational Assessment*, 1–23. https://doi.org/10.1080/10627197.2025.2503288

Abstract

This paper presents and illustrates a framework for visualizing large-scale assessment results in a dynamic score reporting interface to support teachers in making content-referenced interpretations of student growth. The reporting interface maps student performance to locations along a research-based learning progression to facilitate interpretations of the quantitative differences along a vertical score scale. We illustrate content-referenced growth reporting in the context of a learning progression for how students understand fractions. Two key aspects of the illustration include evidence showing that discrete levels of the learning progression have a moderate to strong association with the difficulty of assessment items that were coded to the levels, and results from a small-scale pilot test of the reporting interface with practicing classroom teachers. We speculate about important aspects of implementation that would strengthen the validity of CRG score reporting interpretations and uses.

Keywords: Growth, Learning Progression, Classroom Assessment, Large-scale Assessment, Score Reporting, Item Mapping, Formative Assessment

One of the most underappreciated challenges of educational measurement comes in the presentation of the results from a test instrument, an activity often characterized as "score reporting." Unlike the activities related to test design, administration, and modeling, which are internally directed with content and measurement specialists envisioned as a target audience, score reports are externally directed with a different audience in mind: teachers, parents, and students. And when scores are aggregated at the classroom or school level, the target audience is often the general public. It would therefore behoove managers of large-scale assessment programs to conceptualize and design their score reports *before* finalizing the nuts and bolts of test blueprints, item response modeling, and test form equating (Zenisky & Hambleton, 2012).

When an educational assessment in a given subject domain is administered to students on multiple occasions over time, it is common practice to expect the assessments to provide information about student growth. Numerical estimates of growth depend upon the availability of a common scale to determine whether students are successively answering more difficult questions correctly at different points in time. In the United States context of large-scale "interim assessments" (Perie, Marion & Gong, 2009) for which students are tested on multiple occasions within a school year (e.g. Curriculum Associates, 2018; NWEA, 2019, among others) vendors commonly use two sorts of growth metrics–one that is purely normative (e.g. Betebenner, 2009), and another that is criterion-referenced (e.g. so-called "growth to standard" as described by Castellano & Ho, 2012). These metrics are used to give teachers growth targets for their students. These targets are usually expressed as a percentile or in terms of some number of scale score units that each student would need to meet to exceed normative expectations, or to be on track to meet a future proficiency standard.

There is certainly some value in presenting to a teacher or principal, in the form of a class or school-level score report, information about the number of students who have met normative and criterion-referenced growth targets. Similarly, there is value in conveying to a parent or caregiver how their child's growth compares to these targets. However, these numeric targets do not give teachers or parents substantive, qualitative insights about what a student who has grown, say, 20 scale score points is likely to have learned. Most existing score reports from interim assessment vendors are designed to contextualize, at a high level, what it is that students are likely to know and be able to do given their most recent scale score[1]. The focus is not placed on how this has changed across occasions and how this change can be interpreted. If taken in isolation, a focus on growth categorizations with respect to numeric targets in terms of percentiles, or on a score scale with ambiguous meaning, might convey the message that the primary use of the assessment is as a tool for monitoring and evaluating: a student has either met or not met a given target.

A premise behind this article is that when an educational assessment has been designed to support teachers in using test results for formative purposes[2], then (1) score reporting should emphasize growth as much (or more) than it does status, and (2) this should be done in a way that encourages teachers to connect their interpretations of student growth to the content of the assessment. We argue that a way to accomplish this is to take a "content-referenced" (as opposed to a criterion-referenced) approach to the interpretation of growth, and for this reason we have

[1] For examples, see the growth reports portfolio available at the nwea website (https://www.nwea.org/resource-center/brochure/46865/MAP-Growth-Reports-Portfolio_NWEA_brochure.pdf/) or search for score reports at the websites of interim assessment vendors such Curriculum Associates and Renaissance.

[2] For examples, see https://www.curriculumassociates.com/programs/i-ready-assessment/diagnostic, https://www.renaissance.com/products/star-assessments/, https://www.nwea.org/, https://portal.smarterbalanced.org/library/en/interim-assessments-overview.pdf

taken to characterizing the approach as *content-referenced growth*. We introduce a framework for a content-referenced growth (CRG) approach to score reporting, and illustrate how this framework can be applied to create an interactive digital reporting interface. The illustrative prototype we present here was designed to give teachers the opportunity to visualize and interpret student growth interactively, using a score scale whose interpretation is facilitated by qualitatively distinct reference locations and exemplar items. The CRG approach is motivated by the goal of providing teachers with instructionally meaningful interpretations about numeric growth magnitudes. As a concrete example, which we will elaborate in what follows, instead of indicating that a student has grown 60 points in math over the last three years of school, a CRG reporting interface could help a teacher to convey that a student has gone from solving problems that involve a part-whole conceptualization of fractions, to solving problems that require the student to locate fractions on a number line.

There are four core elements underlying our framework for the CRG approach to score reporting that we consider novel and an important contribution to the research literature when taken in combination. The first is the use of *item mapping* to support interpretations about student growth with respect to reference distances along a score scale. We argue that teachers are most likely able to make sense of reference distances along a scale when those distances are instructionally meaningful. This is accomplished by identifying classes of items that can be ordered according to their difficulty, such that a teacher can connect the average difference in difficulty between the item classes to the time needed to provide instruction. A second element of our framework is ascribing qualitative meaning to growth magnitudes through use of a *learning progression*: a theory about how a concept in a given subject domain becomes more sophisticated over time with adequate curricular and instructional support (e.g. Duschl,

Schweingruber, & Shouse, 2007). It is this theory that helps to explain why we would expect to observe individual differences in test performance both within and across occasions, and it provides a basis for the order in which certain classes of items are expected to become easier to solve given the knowledge and skills that students are developing over time. The third and fourth elements of the CRG framework involve *test design* and *modeling*: tests that have been designed with overlapping item content across test occasions (e.g., grades) and the use of the Rasch Model[3] to calibrate the tests on a common vertical scale. The presence of overlapping item content makes it possible to distinguish differences in the difficulty of the items from differences in the proficiency of students at a particular point in time; the use of the Rasch Model makes it possible to define reference intervals along the scale that do not depend upon the specific items that have been used to define each end of the interval, provided that the model shows adequate fit to the data.

The purpose and focus of this paper is to present what amounts to an "existence proof" for a CRG approach to score reporting. The elements of the approach are informed and build upon frameworks and theories of action such as the assessment triangle (NRC, 2001), construct mapping (e.g., the BEAR Assessment System as described in Wilson, 2023), and the Cognitively Based Assessment of, for, and as Learning project (CBAL; Bennett, 2010). A key distinction is our focus on score reporting of student growth in the context of a large-scale standardized assessment (e.g., Briggs & Peck, 2015). The illustration of the CRG approach we present here focuses on a specific learning progression for the understanding of fractions, and as such we also present empirical evidence with regard to the validity of the progression and how teachers made

[3] This is meant generally to apply to any model with the family of "Rasch models" and would include the simple Rasch Model for dichotomously scored items as well as well-known extensions for polytomously scored items (Andrich, 1978; Masters, 1982; Rasch, 1960).

sense of the CRG reporting protoype. But our primary emphasis is on the feasibility of the approach in general, as opposed to its validity in this particular context.

In the following section, we provide a brief background on our theory of action for the role a dynamic score reporting interface has to play in improving teacher and student outcomes. An implicit context throughout is that of an assessment aligned with curriculum and instruction that is administered on multiple occasions either within a year (grade), across years (grades), or both. We also provide a background on graphic visualizations of growth and the core elements of the CRG framework: item mapping, learning progressions, test design, and vertical scale calibration. In the subsequent sections we turn to a concrete example of how we have implemented and iterated the CRG approach in the context of a learning progression for the understanding of fractions. We summarize the development of the learning progression, and the use of empirical student responses to a large-scale assessment to validate the ordering of learning progression levels. We also present results from a pilot test of the CRG approach via observations and interviews conducted with seven teachers as they interacted with a reporting prototype. We consider key aspects in the implementation of the approach that would strengthen the validity of interpretations and uses. We conclude with a discussion of the potential affordances and limitations of the CRG approach and consider directions for ongoing research on dynamic score reporting.

Background

Theory of Action

Figure 1 outlines a high-level theory of action for CRG reporting. The top of the diagram starts with the end goal, which is to facilitate improved learning outcomes for students. This is accomplished as teachers adjust their instruction based on the inferences about growth and learning they make from using a CRG reporting interface. Goertz et al., (2009) suggest that teachers typically use assessment data to make one of two types of adjustments to their instruction: procedural or conceptual adjustments. Examples of procedural adjustments include use of assessment data to create student groups, to identify what content to reteach, or to select students for intervention. Conceptual adjustments are characterized by using assessment results to understand how students are approaching problems and/or to identify different ways of explaining content. The authors argue that conceptual adjustments to instruction are most likely to help students learn, but that many teachers only use data to make procedural adjustments to their practice. We contend that by focusing teacher attention on the developmental path that students take as they learn a big picture topic, we will better support teachers in making conceptual adjustments to their practice that support student learning.

Insert Figure 1 about here

There are two secondary goals of the CRG reporting interface: 1) to contribute to teachers' professional learning, and 2) to improve attitudes about the usefulness of assessment. When teachers learn to connect growth in points on a quantitative scale to qualitative changes on

an associated learning progression, they will also be engaging in a process of professional learning that deepens their content area expertise. Similarly, as teachers learn—with support—to interact with the reporting interface, we hope they will see the utility of the tool for supporting their instruction and thus develop a belief that a large-scale assessment can be used to improve teaching and learning (for evidence that many teachers do not hold this belief, see Brown, 2004 and Barnes, Fives & Dacey, 2017). Finally, the above actions are made possible by the foundational elements of the CRG framework outlined in our opening section: a domain-specific learning progression, targeted test design, a scale established using item response theory, and items mapped from the scale to levels of the learning progression. The development of CRG score reporting with all four of these elements in place is what makes this approach distinct from the way growth on large-scale assessments is typically reported.

Research on Graphic Visualizations of Growth

A score report is a form of information visualization that uses visual metaphors, such as color, shape, or spatial arrangement, to convey meaning about abstract data, like test scores, that lack inherent interpretability. Effective visualizations require careful attention to "semantic mappings" Hegarty (2019)—the alignment between the display's design and the mental models being invoked by the target audience. In the present context, because the concept of growth is inextricably linked to changes in physical height, a report on student growth in some cognitive domain will be expected to map changes in proficiency (i.e., height), to changes in time. The more that any visualization of a student's cognitive growth departs from a physical growth metaphor, the more designers need to be cognizant of the disruption this poses to a person's mental model for the interpretation of growth. At the same time, if an intended semantic mapping

is maintained when the underlying metaphor has a tenuous basis in reality (e.g., when the units on the y-axis should not be interpreted "as if" they were analogous to the centimeters on a stadiometer), the resulting interpretations of growth might promote misconceptions.

Somewhat surprisingly given its importance, there appears to be very little published empirical research on how representations of student growth are interpreted by the stakeholders of score reports, let alone whether one visualization is more effective in some sense than another[4]. The primary exception comes in a small-scale study reported by Zenisky, Keller & Park (2019). In this study Zenisky et al. presented a convenience sample of adult respondents from Amazon Mechanical Turk with a variety of random pairings of alternative visual representations of student growth percentiles: a one-dimensional tabular array, two-line plots, and a bar plot. They then analyzed how the subjects endorsed statements that "identified" or "interpreted" student growth, and, in the case of the line plots and bar plots, statements that asked them to compare a student's growth over time. Unfortunately, the results presented in this study are difficult to interpret because the actual visualizations subjects viewed, and the complete set of statements they were given to endorse, are not provided. Nor is it clear what the researchers viewed as correct vs. incorrect interpretations and comparisons for each visual. Nonetheless, the authors' conclusion that the accuracy of growth interpretations and comparisons was sensitive to

[4] According to Zenisky, Keller & Park (2019) "To our knowledge there have been no published studies of reporting that have focused explicitly on growth reporting displays, though considerable efforts have been made by state education agencies to develop interpretive guides (across text, presentations, and video formats) to explain reports. Little is known about best practices for reporting growth, nor what elements of growth displays lend themselves to correct interpretations or misinterpretations by intended users." A literature search we conducted using the combinations of the keywords "score growth reporting visualization testing assessment" turned up no relevant studies between 2019 and 2024. As an anonymous reviewer of this manuscript pointed out, many large-scale assessment vendors do put thought and effort into score report designs and pilot test them with stakeholders, but the reports that describe these may not be publicly available or easy to locate by searching the internet.

the choice of visualization lends credence to the potential importance of this line of research. We note also that this small-scale study used a static report and visualization; how stakeholders respond when interacting with visualizations in a dynamic digital environment represents an entirely new frontier.

Item Mapping

The idea of using qualitative distinctions among items in an educational testing context as a basis for making sense of locations along a measuring scale has a long history (see e.g. Galton, 1883; Binet & Simon, 1916; Thurstone, 1925). The term "item mapping" captures the modern implementation of this idea as a tool for making sense of scale scores and achievement levels, exemplified by item mapping efforts for the National Assessment of Educational Progress[5] (NAEP; see Allen & Beaton, 1986; Zwick et al., 2001) and by standard setting procedures which use item mapping to establish the locations of achievement levels (e.g., Lewis et al., 2011).

As implemented in NAEP, item maps have tended to focus on providing qualitative interpretations for scale score levels through the locations of exemplar items, and we use them to this end in the CRG reporting application as well. However, in the CRG use context, the more important purpose of item mapping is to provide a qualitative interpretation of the numerical *distances* between scale score locations with respect to the qualitative differences of exemplar items at these locations.

[5] https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_itemmapping.aspx

Learning Progressions

Learning progressions, also sometimes called *progress maps* (Masters & Forster, 1996) or *learning trajectories* (Clements & Sarama, 2009; Lobato & Walters, 2017) have been defined as "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (Duschl, Schweingruber, & Shouse, 2007, p. 8-2). These successively more sophisticated ways of thinking are typically characterized by discrete levels or "waypoints" (Wilson, 2023) that are intended to represent significant intermediate steps in understanding commonly encountered as a student proceeds from "novice" to "expert" with respect to a conceptually rich learning target. One of the more interesting and challenging aspects of learning progression research and development is to put forward conjectures regarding the pathway students are likely to follow where it is not terribly useful to think of students dichotomously as those who "get it" and those who "don't." (Otero & Nathan, 2008; Peck et al, 2021).

A learning progression provides a principled basis for empirically testing developmental hypotheses through a combination of item design and item difficulty modeling (e.g. NRC, 2001; Briggs & Peck, 2015). More specifically, in the context of CRG, it is the hypothesized levels of a learning progression, and the items that are being used to distinguish between them, that provide a theoretical rationale for the activity of item mapping. When a collection of test items has been administered to students it will always be the case that the items vary in difficulty. Mapping the scale according to item difficulty is a purely descriptive process. But when items have been designed according to a learning progression hypothesis, they are expected to have a predictable ordering. It is only when this is borne out to an adequate extent that one proceeds to map item exemplars to a score scale as way to facilitate a content-referenced interpretation of growth.

Later in the concluding discussion section of this paper we return to the issue of how predictable this ordering needs to be before it can be considered "adequate."

Test Design

A defining feature of most learning progressions is that they are oriented around a "big picture" concept for which it may take multiple years of support—not just one—to develop some predefined goal of mastery. From an assessment perspective, this requires a departure from an approach in which the tests that a student takes over the course of adjacent grades are conceptualized as being comprised of entirely unique content. Instead, just as curriculum should be designed to be spiraled, tests must be purposefully designed to have overlapping content across temporal occasions. In an educational measurement context, this requirement would help to fulfill the conditions needed for a common item nonequivalent group design. In this design, students, typically in adjacent grades, are administered a subset of common items that can be thought of as an anchor test, and this makes it possible to disentangle differences between the abilities of students from the difficulty of the test items they have answered. When inspired by a learning progression, the design challenge is to create items that are able to discriminate between student ability at different levels of the progression, and this makes item writing a theory-based activity.

Calibrating a Vertical Scale with the Rasch Model

When students at unique temporal occasions (e.g., different months, different grade-levels) have been administered common items, it becomes possible to place the occasion-specific tests onto a common vertical scale. Although a variety of psychometric models facilitate vertical

linking (e.g., Tong & Kolen, 2007), in the CRG approach illustrated here, we use the Rasch Model. A known property of the Rasch Model, when it can be shown to adequately fit the data, is invariance of comparisons among persons to the choice of items used to define a reference interval. Specifically, under the Rasch Model, the choice of response probability threshold (e.g.. the 50%, or 75%) used to map item locations has no impact on the distances between items used to demarcate a reference distance along scale. In contrast, under more flexible IRT parameterizations such as the 2PL or 3PL, because item characteristic curves can cross, both the ordering of items on a common scale and the distances between their locations can change as a function of the criterion used to locate the items on the scale. In the discussion section we consider some implications of taking a CRG approach when a vertical scale and item mapping is not established using a Rasch Model.

Data and Methods

The development and pilot testing of the CRG reporting prototype we illustrate here makes use of data from the *i-Ready Diagnostic*, developed and maintained by the organization Curriculum Associates. The *i-Ready Diagnostic* is an assessment comprised of grade-specific standardized tests in reading and mathematics, administered during the fall, the winter and the spring of each academic school year. Students take each test on a digital interface, and tests are designed to be adaptive such that each new test item to which a student is exposed depends upon whether they answered prior items correctly. The mathematics test for students in grades K-12 consists of up to 66 items, and the content of these items is organized into four strands: Algebra, Geometry, Measurement, and Number and Operations. The data that we leverage for the CRG

reporting prototype consists of item and person parameter estimates. The item parameter estimates are part of the *i-Ready Diagnostic* operational item bank, previously calibrated by fitting the Rasch Model to dichotomously scored test item responses. Students who take the test adaptively can then be given a score on the same scale, by estimating their latent ability parameter with item parameters fixed to their operational bank values.

A distinguishing feature of Curriculum Associates' approach to working with school districts is that it seeks to bundle the *i-Ready Diagnostic* assessment with curricular resources that teachers can use as part of their efforts to facilitate student learning. To this end, scores on the fall and winter assessment occasions are used to determine students' "placement levels," and these placement levels are in turn associated with targeted curricular materials. Following the spring assessment, there is considerable attention on the extent to which students, individually and in the aggregate, have met normative and criterion-referenced targets for growth from fall to spring. As described in the opening section of this paper, this places the use case for student growth reporting into a fundamentally summative context—which students hit their growth targets, and which did not? Which teachers had larger proportions of students hitting their growth targets? In contrast, the CRG approach represents an opportunity to incorporate growth reporting in a more formative context (see Figure 1).

Several key elements of the CRG framework already exist as part of the *i-Ready Diagnostic*: within and across grade tests designed with overlapping content, a vertical scale calibrated using the Rasch Model, and the ability to map items at different locations on the scale to different levels of a learning progression. The items on the *i-Ready Diagnostic* have been aligned to college and career ready standards (e.g., Common Core State Standards, TEKS, FL BEST, etc.). Because many clusters of standards in math and reading are amenable to an ordering

that suggests a learning progression hypothesis, it follows that it is possible to retrofit subsets of *i-Ready* items across grades to the levels of one or more learning progressions, and then evaluate the defensibility of this retrofit through item difficulty modeling and sensitivity analyses.

Our process for developing unique CRG Reporting Prototypes in mathematics and reading included the following key steps:

(1) Specification of a research-based theory of sociocognitive development with respect to a "big picture" concept in the subject domain (i.e., the learning progression).

(2) A detailed articulation of the distinct levels of the learning progression in terms of what a student is expected to understand, what a student might not yet understand, and what a teacher could do to facilitate understanding via instructional activities.

(3) Consultation with an advisory panel comprised of subject matter experts in the content of the learning progression.

(4) The alignment of assessment items to each level of the learning progression based on the level of understanding a student likely needs to answer the items correctly.

(5) An evaluation of this alignment by regressing item difficulty onto learning progression levels (as well as other item-level covariates).

(6) An interactive graphical visualization (at the student and classroom levels) designed to emphasize the connection between growth in terms of vertical scale values on one side and the levels of a given learning progression on the other.

(7) Pilot tests of the CRG approach through think-aloud interviews with practicing teachers as they interacted with a CRG reporting prototype.

(8) Multiple prototype design iterations and internal critiques based on results from steps 1-7.

In the next section we turn to a presentation of one of the four prototypes we developed, and in the following section we present the results from our pilot test with teachers[6].

Content Referenced Growth for an Understanding Fractions Learning Progression

A Learning Progression for Understanding Fractions

Given the foundational nature of fractions understanding (Common Core State Standards Initiative, 2010; Empson et al., 2011; Bailey et al., 2012; Booth & Newton, 2012; Siegler et al., 2012; Torbeyns et al., 2015), we developed a learning progression (LP) for fractions that is meant to help students, teachers, and parents make sense of growth in this domain. Our theoretical LP based draws upon Kieren's (1976, 1980) five conceptualizations of fractions, and also includes key ideas that have been used to define levels in other pre-existing learning progressions for fractions (CCSSI, 2010; Arieli-Attali & Cayton-Hodges, 2014; Wright, 2014; Wilkins & Norton, 2018; and Yulia et al., 2019). The LP that forms the basis for our CRG prototype has four levels. (A detailed exposition of the levels can be found in Appendix Table 1.) Movement from the first to the second level of the LP requires a student to go from understanding fractions as primarily as **part-whole visual relationship** to understanding them as

[6] The results from developing an understanding fractions learning progression for this prototype, aligning levels to items, and validating the alignment through item difficulty modeling (i.e., steps 1-5 above) are the subject of a detailed report (Wellberg, Briggs & Student, 2022). We also have written a report on the results of our pilot test of the CRG prototype (Briggs, Cox, Student & Whitfield, 2023). Given space constraints, we only present a high-level summary of these results, with a focus on the two steps that involve novel empirical data analyses (item difficulty modeling and the pilot test). An important caveat is that although what we have created is a prototype for a dynamic reporting experience, we are necessarily describing it here in a static format. To remedy this, we include (blinded) links to the reporting prototype described in the next two sections (https://contentreferencedgrowth.github.io/prototype-react/) and also to a revised version based on the results of our pilot test and feedback from our advisory panel (https://contentreferencedgrowth.shinyapps.io/prototype-public/).

quotients that equipartition a whole into unit fractions whenever a person is asked to create "fair shares." Next, at level three of the LP students are able to understand fractions as **measurements** that represent a magnitude along a number line, which allows them to order fractional values and to understand the equivalent fractions that allow for the addition of fractions with unlike denominators. Finally, at level four, students are able to interpret fractions as **operators** that take a value and produce a new value that is proportional to the original through the activities of multiplication and division[7].

Empirical Validity Evidence Supporting the Learning Progression

The instructional sequence plays a critical role in any learning progression (Confrey, Maloney, & Corley, 2014). In mathematics, ideas tend to build upon themselves, and fraction concepts are no exception. While this does not necessarily mean that students must have completely mastered all previous concepts and procedures before moving on to more complex topics, it does imply that later-learned topics are likely to be more sophisticated and more complex than are those that are learned earlier (Confrey, Maloney, & Corley, 2014). Items that reflect these later concepts are, therefore, likely to be more difficult. In order to confirm this empirically, we began by inspecting the relationship between the grade-level ordering of the fractions-related lessons in Curriculum Associates' i-Ready Classroom Mathematics, which is highly aligned with the CCSS-M, and the estimated difficulties of the *i-Ready Diagnostic* items

[7] In our original development of this learning progression, we proposed a level located between the "measurement" and "operator" levels described here in which students could understand fractions in terms of ratios and rates. However, we found that this level did not align with the order in which the concept of ratio and rate are introduced in standards-based mathematics curricula, and following advice from content experts on our advisory panel, we decided that the concept of ratio and rate was better suited for a distinct learning progression on proportional reasoning that would extend into middle school.

that assess the content covered in those lessons. We then examine the association between the levels of our LP and the difficulty of *i-Ready Diagnostic* items that we were able to code as belonging to each level

Insert Table 1 about here

Content experts at Curriculum Associates group and characterize the items in the *i-Ready Diagnostic* assessment system according to the specific content knowledge and skills that students are presumed to require to accurately complete the problem. We refer to these item groupings as "assessed skills." Table 1 illustrates how a subset of six assessed skills related to fractions were associated with six curricular lessons and 54 items. In total we identified 107 assessed skills on the *i-Ready Diagnostic* that contained references to understanding fractions, and these were associated with 406 unique *i-Ready Diagnostic* items. Two members of our research team independently coded each assessed skill based on which of the four fraction conceptualizations (the four levels of our LP) would be most important to understand in order to correctly answer an item in that group. Initial agreement across raters was very high, with matches on 101 of the 107 (94%) assessed skills, and we discussed the six mismatched skills until we agreed upon a code for each. We also had access to the actual bank of items used in the *i-Ready Diagnostic* and confirmed that the items were consistent with the written description of the associated skill the item was intended to require in order for it to be answered correctly.

If our theoretical LP for fractions holds, then we would expect to see that students in lower grades were primarily exposed to less complex conceptualizations of fractions, and students in upper grades exposed to increasingly more complex conceptualization. To test this

hypothesis, we examined the lessons in the i-Ready Classroom Mathematics curriculum and recorded the lesson in which students would have first encountered content specific to each assessed skill. Using all *i-Ready Diagnostic* fractions assessed skills as the unit of analysis, when the 32 i-Ready Classroom Mathematics lessons specific to fractions are coded in the chronological order in which they appear in grades 3 to 6, we find that the rank correlation between lesson order and LP level of an assessed skill found in that lesson is .68—a moderate to strong association. As students move up in grade levels they are indeed more likely to receive instruction about fractions that would involve the more sophisticated conceptualization of fractions found in the higher levels of the LP.

Insert Figure 2 and Table 2 about here

Next, using *i-Ready Diagnostic* items as our units of analysis, we examined the association between the difficulty of the items, and the level of the LP to which we coded the items. Figure 2 presents a graphical visualization of our results. The values along the vertical axis are in *i-Ready Diagnostic* scale score units; items with lower/higher values are those that were easier/harder for students to answer correctly. As the LP level increases (from left to right on the horizontal axis), so does the average item difficulty. The rank correlation is .60. Results from an ANOVA followed by multiple pairwise comparisons indicated that differences in mean item difficulty by LP level can be considered statistically significant. There is, however, clearly some degree of overlap in these item difficulty distributions. For example, there are items we coded as involving a part-whole conceptualization of fractions (LP level 1) that are empirically harder for students to solve correctly than some items we coded as involving a quotient (level 2),

measurement (level 3) or even an operator (level 4) conceptualization of fractions. And vice-versa, there are items we coded as involving an operator conceptualization (level 4) that are easier than items we coded at levels 1, 2 and 3. As we discuss in more detail later in the "Strengthening Validity" section of this paper, this is a reminder that what we can infer about a student's understanding relative to a learing progression on the basis of their response to any single test item is equivocal.

To further establish that there are meaningful distinctions between the LP levels, we also regressed item difficulty on the item's corresponding LP Levels, but controlled for the grade level to which the item had been intended. The results from this exercise can be found in Table 2. On average, even after controlling for grade level, items coded to higher levels of the fractions LP tended to be significantly harder to solve correctly.

An Overview of the CRG Reporting Prototype

Figure 3 and Figure 4 are screenshots of version 1.0 of a CRG Reporting Prototype for the Understanding Fractions Learning Progression[8]. In both figures the left vertical axis is demarcated in increments of 25 units on the *i-Ready Diagnostic* scale, and the right axis is demarcated by the four levels of the fractions LP. In the class view (Figure 3), each location along the horizontal axis represents a unique student in a hypothetical teacher's class, each green dot represents a student's *i-Ready Diagnostic* scale score, and the students are ordered by the magnitudes of their *i-Ready* scale scores. In the student view (Figure 4), each location along the

[8] To interact with the dynamic version of the original prototype, see https://contentreferencedgrowth.github.io/prototype-react/

horizontal axis represents a unique test occasion (grade and season), and when paired with the *i-Ready* scale score for that occasion this maps out an empirical growth trajectory from the earliest to most recent test occasion for a student depicted as a dot in the class view. Clicking on a dot in the class view brings up the student view.

Insert Figure 3 and Figure 4 about here

In both class and student-level views a central goal of the reporting interface is to help teachers interpret quantitative differences in scale score with respect to qualitative differences in LP levels. One way this is done is to include horizontal lines that visually connect a specific threshold between LP levels to an *i-Ready Diagnostic* scale score. Another visualization that the user can select to make this connection (not shown here) provides color coding over the area of the plot corresponding to each level such that a horizontal region changes in gradient as one goes from the lowest to highest *i-Ready* scale score value. For version 1.0 of the CRG prototype we located the entry threshold for an LP level as the point at the *i-Ready* scale at the median of the item difficulty distribution of items written to that level (recall Figure 2). This means that a student with a scale score at this same location would be predicted to answer at least half of these items correctly.

Insert Figure 5 about here

Notice that with the exception of the "Part-Whole" level, the gray boxes providing labels for levels 2-4 of the LP depart from the formal names that derive from the math education literature

that we used in Figure 3. This was done purposefully to give teachers a better gist of the kind of fraction activity at the heart of each level conceptualization. To this end we substituted “Fair shares” for “Quotient”; “Number line” for “Measurement,” and “Multiply and Divide” for “Operator.” To give a better sense for what student thinking at each level entails, when a teacher clicks on any one of the LP labels they are given the view similar to the one shown in Figure 5 for the “Number Line” level. In this view, the teacher is given both a bulleted summary of what a student would and would not be expected to understand about fractions at this level, as well as an exemplar item they would be expected to be able to solve correctly. A detailed view of the full LP (shown in Table A-1 in supplementary materials) can also be downloaded as a 2-page pdf by clicking the question mark icon in the class view (see Figure 3).

The CRG reporting interface is dynamic in the sense that a teacher can easily toggle between student and class-level views, choose to turn on or off optional features (e.g., ordering students by score, adding a color gradient to demarcate LP levels), and can zoom in and out of level-specific LP interpretations that include exemplar items. Our contention is that the more that a teacher interacts with the CRG reporting interface and comes to understand the qualitative distinctions between LP levels, the more likely they are to interpret the scale score increase from the fall of grade 2 to the spring of grade 4 not so much as growth of about 60 scale score units (based on the left side of Figure 4), but as a student going from solving problems that involve a part-whole conceptualization of fractions to solving problems that involve a number line (i.e., measurement) conceptualization of fractions (based on the right side of Figure 4). In other words, growth is being given a content-referenced representation.

Pilot Test with Teachers

Methods

We recruited seven teachers to participate in 45-60 minute "think-aloud" sessions via Zoom in which they interacted with the CRG reporting prototype. Participating teachers were screened to ensure that they had prior experience with the *i-Ready Diagnostic*, and were purposefully sampled to have experience in grades 3, 4 or 5 (since the focal grade used in the prototype was 4), and to vary in geographic location (three were from California, one from Iowa, one from Massachusetts, one from Missouri; and one from New York). Four of the teachers were serving as instructional coaches in their school districts.

Three central questions motivated our observations of teachers during these sessions.

1. To what extent can a content-referenced growth report prototype facilitate meaningful interpretations about differences in student scale scores on the *i-Ready Diagnostic*?
2. To what extent are teachers able to understand and interpret the fractions learning progression by using the prototype?
3. How do teachers envision using information provided by the prototype?

In collaboration with User Experience researchers we developed an observation and interview protocol with open-ended questions and minimal pre-teaching or professional development about the prototype. As we later discuss, if the CRG reporting interface were to be implemented at scale, it would surely require some professional development for it to be used efficiently and effectively for its intended purposes. However, at this stage our aim was to find out whether the dynamic reporting interface we had built was intuitively sensible even in the absence of professional development activities related to the learning progression underlying the prototype.

In our Zoom sessions, after sharing a screen with the reporting prototype (e.g., see Figures 4 and 5 or https://contentreferencedgrowth.github.io/prototype-react/), the participating teacher was given mouse control, and then asked to explore the prototype while narrating their thoughts out loud. During this initial period of about 10 minutes, we only interrupted (as necessary) to remind the teacher to think aloud. Following this we asked 5-8 open-ended questions about how the teacher was interpreting and making sense of the information and options available to them in the class and student views, their general impressions about the utility of the interface, how it could be improved, and whether they had any other questions or comments.

Each interview was conducted with at least two members of our research team present, one member serving as facilitator, the other as note-taker. In addition, we were given permission by all participating teachers to record each interview. Following the interviews, we went through a process of rewatching the interviews and audio transcripts to engage in a process of deductive coding for themes related to answers teachers gave to our planned questions, and a process of inductive coding for themes that emerged more spontaneously. In what follows we report just the results from our deductive coding.

Evidence that the reporting prototype facilitated meaningful interpretations about differences in student scale scores

In all seven interviews teachers interacted with the prototype and responded to our questions in ways that indicated they were able to use the prototype to make connections between *i-Ready* scale scores and LP levels to support inferences about student growth. Statements about growth and progress most frequently occurred while teachers were in the

"student view" of the prototype (see Figure 4) and in the "class view" when teachers clicked a checkbox to show prior scores (see Figure 3). For example, when asked to describe what he was seeing and what it might represent when exploring the class view in the prototype, one teacher began to connect scale scores and LP levels by first reading the names of each level and then saying, "I'm guessing that those correspond to the scale score, and the types of problems that they'll be getting, that they either would get or they would need to understand." Then, turning his attention to the green dot representing one student, he went on, "so if I'm looking at LY… It's just above 450, 455. They understand fair shares, and their next progression would be to go towards number line understanding." Similarly, upon navigating to the student-level view of the prototype, another teacher moved her mouse pointer from the left side of the vertical axis, which displayed the *i-Ready* scale score, to the right side, which contained the LP levels, and said, "let's see, they're up to 475, so they're at the level where they can do things related to the number line."

Some teachers defaulted to prior interpretations of *i-Ready* growth from reports they had previously viewed that included scale score targets for a "typical" year of growth. For instance, when looking at a view that included prior scores in the class view, a teacher remarked, "this shows growth. This is what we're really interested in. They have a fall, a winter, and a spring." She then clicked on each of the dots representing the testing occasion and said, "you can click on it, it's 411 to 435. So, they made more than a year's growth in a year. Because typical growth is anywhere from 12 to 15 points, which is what we've seen." In this case, although the teacher used the prototype to make inferences about student growth, she drew primarily upon on her pre-existing conceptualizations of typical growth from other *i-Ready Diagnostic* score reports.

In contrast, a different teacher spoke about growth in terms of how students had progressed in levels of understanding on the LP, explaining that for the student shown in the student view of the reporting prototype, "they have a decent understanding of how it [fractions] can be represented on a number line. They are seeing the relationships between the fractions and decimals… they're not really ready for the multiplication and division." In both of these two cases the reporting prototype appeared to facilitate meaningful interpretations about differences in student scale scores, but the second case was more in line with the intended content-referenced interpretation of growth that we had intended to elicit in our design.

Evidence that teachers were able to use the reporting prototype to understand and interpret the learning progression

One structured part of our interview was to ask teachers questions that would prompt them to click through each of the levels of the LP (the gray boxes shown in Figure 3), and then later to download the detailed explication of the LP as a pdf. We would then ask if the differences between the levels of the LP made sense, and, more specifically whether the written descriptions and exemplar items were helpful in their sense-making process. All the teachers were able to follow the hierarchical distinctions in student understanding being made across levels of the LP. However, not all teachers agreed that we had the appropriate grain size for each level, or about the order of the two middle levels. For example, three teachers felt that level 2 of the LP (fair share or quotient conceptualization) contained too many distinct concepts, and that it represented "a big jump" to go from level 1 to 2. Two teachers felt that our level 2 conceptualization would be harder to teach than our level 3 (number line or measurement) conceptualization. Some support for this argument comes from the fact that the CCSS-M standards do in fact emphasize

an understanding of "fractions as numbers" as something that begins in grade 3 (see in particular 3.NF.1, which is part of our level 3) while an understanding of equivalent fractions is not fully consolidated until grade 5 (see in particular, 5.NF.3 which is part of our level 2).

Insert Figure 6 about here

More generally, this points to the fact that the levels of the fractions LP were not designed to be identical to the grade level ordering of standards found in the CCSS-M. To make this concrete, Figure 6 compares the two exemplar items we included in the prototype to distinguish levels 2 (Fair Shares) and 3 (Number Line). When teachers compared these two items, they argued that the Fair Shares exemplar item would be more cognitively demanding than the Number Line exemplar, both because of the understanding of fractions needed to answer correctly, and because the Number Line item included a visual representation, while the Fair Shares item did not. Indeed, on closer inspection the Number Line item can be shown to align with CCSS-M standard 3.NF.2 (Represent fractions on a number line diagram) while the Fair Shares item aligns with CCSS-M standard 5.NF.3 ("Interpret a fraction as division of the numerator by the denominator ($a/b = a \div b$). Solve word problems involving division of whole numbers leading to answers in the form of fractions").

How teachers could envision using the CRG reporting prototype

We asked all teachers a specific question about how they could imagine themselves using the information presented in the prototype in support of instructional planning or practice. Three different use cases emerged: (1) to place students into groups for instructional activities targeted

to the understanding of fractions, (2) to facilitate conversations with parents about student progress, and (3) to support professional learning of mathematics content among teachers.

Six teachers made comments about the potential usefulness of the reporting prototype for placing students into small groups. Although five out of the six teachers saw the most value in using the prototype to place students into more homogenous small groups, one teacher saw greater value in using it to create heterogenous groups so that students that differed the most significantly in their understanding of fractions could learn from each other.

Three teachers thought the reporting prototype could provide a useful way to communicate with both parents and students about their progress. One teacher focused on the possible use of removing student initials from the class view and then sharing out results with the class. Another mentioned that the prototype could be useful for parent-teacher conference meetings, making it "extremely easy to provide parents with information about what scores mean."

Five teachers saw potential in the CRG reporting prototype as a tool for teacher learning and development. Several participants suggested that newer teachers or teachers who were less familiar with math standards could use these elements of the prototype, especially the detailed LP pdf document (which we replicate in Appendix Table A-1), to deepen their content knowledge. One teacher, for example, was optimistic that other teachers would be able to use the document to "drill down into the standards." Taken together, teacher interview responses suggested to us that the information shared in the prototype (LP level descriptions, exemplar items, and the more detailed LP pdf) held some promise for helping teacher learn more about topics in math instruction and math standards. This reinforces some core elements of the CRG theory of action (recall Figure 1), particularly the role of the prototype in supporting professional

learning and changing teacher attitudes about not just assessment, but of their own identities as math educators.

Strengthening the Validity of CRG Interpretations

On the whole, we took the results of our pilot test of the CRG reporting prototype as evidence that we were on the right track, and that the approach is feasible has the potential to serve as the linchpin envisioned in our theory of action. However, there are a number of things that could be done to further strengthen the validity of CRG score reporting interpretations and uses.

Integration with Ongoing Professional Development and Standards-based Assessment

In our pilot test, teachers were asked to interact and interpret a reporting prototype premised on a learning progression for understanding fractions with a bare minimum of background context. As practicing teachers with multiple years of experience, these teachers all had varying degrees of expertise in teaching fractions to elementary school students. However, as Peck et al. (2021) argue, without good professional development opportunities, a learning progression framing of student assessment can come into conflict with a "count up points" framing. In the latter, learning is viewed as the acquisition of discrete pieces of knowledge that is either mastered or not mastered, and assessment is viewed as a matter of finding out how many items have been answered correctly and then counting up the points. Growth becomes a matter of looking for changes in the points earned across occasions, preferably using the same test twice. In contrast, a learning progression approach to assessment focuses attention on both quantitative

and qualitative changes in student reasoning over time. This is not to say that a learning progression approach is indifferent to student mastery of content standards; however, the focus is cumulative across multiple grades, and there is a far greater emphasis on how students are changing in the way that they reason about a big picture concept rather than whether they are "at grade level" at a particular point in time.

With respect to the example used in this illustration, Number & Operations-Fractions represent one of 11 specific domains in which the CCSS-M standards are organized across grades. The fractions domain includes 30 specific content standards across grades 3-5, and in principle each of these content standards could be associated with one or more of the standards for mathematical practices. In contrast, our understanding fractions LP is expressed with respect to just four ordered levels. The premise of the LP approach is that it is more productive and more efficient for teachers to be paying attention to how well their students invoke part-whole, quotient (relabeled "fair-shares" on the prototype), measurement (relabeled "number line" on the prototype), and operator (relabeled "multiply & divide" on the prototype) conceptualizations to solve math problems involving fractions than it is to focus more narrowly on a student's standard by standard mastery. In other words, the categories of the LP are meant to form a sensible instructional schema. But this would surely require initial and ongoing professional development and collaboration to support. And since content standards are likely to remain an important "coin of the realm" it would be important to create a version of the CRG reporting prototype with a "zoomed in" view in which teachers would be able to locate specific content standards as one of the finer-grained elements that comprise a given LP level.

Choosing Exemplar Items

Exemplar items used to characterize the distance between two scale score locations need to be chosen with great care. For example, in version 1.0 of the fractions CRG reportimg prototype, although the items we included as exemplars of levels 2 and 3 had a good theoretical alignment with each level, specific features of the level 2 item could have made it more difficult to answer correctly than the level 3 item. A better design principle when choosing a single exemplar item to represent each level would have been to choose items according to a standard criterion (e.g., an items that represents the most cognitively complex application of the fractions conceptualization at that level). An alternative principle would be to provide 3-4 items at each level that better characterize the variability in item difficulty. A problem with using just one exemplar per item is that it runs the risk of teachers treating a single item as the target for instruction, as opposed to treating it as just one of many ways that evidence of student reasoning could be elicited. On the other hand, if teachers are presented too many items, this can become cognitively overwhelming. The approach we have since settled upon as a healthy compromise is to pick one exemplar item per LP level according to a standard criterion, but to then allow for a "zoomed in" view of each level that includes an additional three items that span most of the range of scale scores encompassed by the LP level. (To see how this change was enacted, visit https://contentreferencedgrowth.shinyapps.io/prototype-public/)

Item Design

The levels of an LP are meant to capture hypotheses about qualitatively distinct ways that students think and reason when presented with a task. If a task has been written to discriminate between adjacent levels of an LP, the best way to find out if a student is using level X vs. level X

+ 1 thinking is to ask them to solve the problem and to show their steps and/or explain their reasoning. This is one reason that the exemplar items we used for our version 1.0 fractions reporting prototype were all open-ended. In contrast, the items on the *i-Ready Diagnostic* are almost all selected-response. While selected-response items may not be the optimal format for eliciting student thinking, other requirements of the CRG framework, such as test design and scoring, continue to make the use of selected-response items necessary. In the future, other selected-response options that provide more qualitative information could be used in a CRG reporting approach. For example, an ordered-multiple choice format could be considered in which an item's selected-response options are mapped to different levels of a LP, which would then facilitate partial credit scoring. Alternatively, with advances in AI it may become possible to elicit more concrete evidence of student thinking as, for example, students interact with a chatbot that can provide or withhold scaffolding as necessary in follow-up prompts after posing an initial task for the student to solve. A key point here is that an LP design approach may require assessment tasks that look quite different than traditional selected-response items. Hence, although it might be necessary for practical reasons to begin a CRG reporting approach by retrofitting pre-existing items to an LP, ongoing efforts to improve and validate the approach will likely benefit from new innovations in item design and development.

Triangulation

In the CRG reporting approach illustrated here, a scale score on the *i-Ready Diagnostic* can in fact be mapped to one of four levels of an understanding fractions LP. But the underlying relationship of interest—the ability to use an increasingly sophisticated conceptualization of fractions to solve math problems that invoke them—is probabilistic, not deterministic. In the

illustrative example considered in this paper, a student with a scale score that falls within the level 3 region of the LP is one who we expect to be more likely to solve problems that involve adding or subtracting fractions with uncommon denominators than a student with a scale score that falls in the level 2 region. And from this we infer that the reason they are successful in solving these problems is that they understand that all fractions, irrespective of their numerator and denominator, can be located as measures on a number line. But we must keep in mind that the location of each student on the scale has uncertainty due to measurement error. And we should recall from Table 2 that knowledge of the LP level to which an item is aligned only explains 40% of the variance in item difficulty on the *i-Ready* assessment. It follows that teachers can expect that a student categorized at level 3 may still struggle with challenging equipartitioning tasks and may therefore benefit from practice and instruction working on these kinds of problems.

In our view, the results from a large-scale assessment that has been mapped to levels of an LP and administered on a single occasion should be taken as a starting point for inquiry rather than a final determination. This is why the focus of the CRG reporting approach is on changes across multiple occasions, and how these can be given a qualitative interpretation. When the focus is on diagnosing student understanding at one point in time, it will be important for a teacher to have triangulating evidence from tasks that do not have the same constraints as the items found on a standardized assessment. To this end, a key feature of our most recent iterations of the CRG reporting prototype involves the inclusion of at least one "follow-up" activity associated with each LP level, along with things to look for in a student's response that would be consistent with the conceptualization of fractions at that level.

Discussion

The CRG framework uses item difficulty modeling to establish that the levels of an LP can be used as reference locations to give instructionally meaningful interpretations to distances along a score scale. In the example provided with our fractions LP, knowing the level of the LP to which an item is aligned explained 40% of the variance in item difficulty. Is there a minimum R^2 necessary before it is sensible to proceed with the CRG reporting approach? It depends. The more relevant statistic is probably the root mean square error (RMSE) from the regression of item difficulty on LP level. A large R^2 paired with high variance in item locations may well produce a RMSE that is comparable to a lower R^2 with a lower variance in item locations. When an RMSE is greater than the distance between two LP-based reference locations on the scale, it may suggest the need for a larger reference distance, or at the very least, for care in generalizing the meaning of the reference distance with respect to any two items sampled at random from the two LP levels. In our fractions CRG reporting example, the item difficulty RMSE was between 28 and 22 points on the *i-Ready* score scale (depending on the model), hence only the distance between level 1 and 2 LP thresholds was greater than this criterion.

Could the CRG approach still be used if a vertical scale was calibrated using a more flexible IRT model in the context of dichotomously scored items such as the 2PL or 3PL? Yes, it could. After all, the NAEP item maps are based on a 3PL IRT model. However, in the CRG context, this would threaten the generalizability of reference distances along the scale, because these distances are defined by the location of items relative to the locations of persons on the common scale. Under the 2PL and 3PL, the ordering of items can change as a function of the response probability used to locate the items on the scale. In this sense only IRT models in the

Rasch model family are ideal because they establish a basis for invariant comparisons on an interval scale (see, e.g., Briggs, 2013; Domingue, 2013).

In the initial CRG version 1.0 reporting prototypes we provided a hypothetical example of a student for whom an observed trajectory could be plotted across nine test occasions spanning three school years (recall Figure 4). As the number of test occasions increases, the accuracy of a modeled growth parameter as an estimate for the observed trajectory will improve. But for a small number of occasions (e.g., 2-4), the measurement error associated with individual student trajectories could lead teachers to misinterpretations about student growth. One approach that we have been investigating is to use a population model for a past longitudinal cohort to produce an Empirical Bayes estimate of the trajectory for a student in a new cohort. It would then become an open question whether it is better to report an Empirical Bayes estimate of a student's growth trajectory (which might be shrunken and smoothed), the noisier observed trajectory, or both.

Finally, in the basis for the CRG reporting application illustrated here was a vertical scale that was designed with overlapping content for the full mathematics domain, even though the growth we are attempting to describe qualitatively is in terms of a subset of content (fractions) specific to a single domain (number and operations). Would the distributions of item difficulty by LP level shown in Figure 2 look the same (with approximately the same distances between levels) if we had instead created a vertical scale solely with items specific to fractions content? One way to investigate this is by making use of embedded field test items administered to students each year when they take *i-Ready Diagnostic*. As part of this design, students are randomly administered both on and off-grade level items. This makes it possible to use this data to simulate the process of creating three different vertical scales spanning grades 2 through 6. In

one, all items are used irrespective of their mathematics content domain (the current design of the vertical scale); in a second, only content from items in the number and operations domain is used; and in a third, only items with fractions content are used. The closer the linking constants from these three designs are to each other, the more defensible it will be to interpret differences in scale scores calibrated across multiple mathematics content domains in terms of the more focused content of a learning progression.

Conclusion

Research and development efforts related to score reporting of growth for large-scale assessment programs (at least that which has been publicly disseminated) have been few and far between. What research does exist on score reporting has focused on static score reports about test performance at a single point in time. But the assessment programs of the present exist in digital environments, and this naturally lends itself a conceptualization of score reporting that is interactive, not static. And the demands for future large-scale assessment systems will place an increasing premium on the ability of assessments to produce meaningful evidence about student growth. In this article, we have presented a framework and an existence proof for the development of interactive score reports that focus on qualitative interpretations of student growth, an approach we have characterized as content-referenced growth reporting. We have illustrated how this framework can be applied toward the development of an interactive reporting prototype; how it can be systematically pilot tested; and how important it is to anticipate and attempt to address threats to the validity of interpretations and uses.

Through the course of the research and development of the CRG reporting approach, we have developed four prototypes for big picture ideas in mathematics (fractions, ratios, functions and spatial measurement) and one prototype for a big picture idea in reading (phonics). The feedback we have gotten from pilot tests, advisory panels, and conference attendees to date suggest that this is a good and promising start, but more research would be valuable before (or when) the approach is taken to scale. Three questions in particular should be explored empirically. First, how does a teacher's interaction with the CRG reporting interface change when they are seeing real data from their actual students, as opposed to a hypothetical group of students (as in our pilot test)? Second, to what extent does the reporting interface lead teachers to the desired inferences and next steps? Third, does use of the CRG approach have a positive effect on teacher attitudes towards assessment, instructional practices, and student outcomes?

It is important to appreciate that no approach to reporting student-level growth will lead to definitive interpretations. Learning and growth are often messy and complicated, and attempts to model what this looks like "on average" will necessarily be imperfect. A key ethos of the CRG approach is that the questions teachers should be asking about their students upon interacting with the reporting prototype should be even more important than the answers they are getting. We suspect that the status quo of score reporting does not give teachers actionable answers, or engage them sufficiently in the content of the assessment to get them to ask the substantive questions about student learning. The CRG approach represents an attempt to break from this status quo.

References

Alonzo, A. C. & Gotwals, A. W. (2012). Learning Progressions in Science: Current Challenges and Future Directions. Springer.

Arieli-Attali, M., & Cayton-Hodges, G. (2014). Expanding the CBALPM mathematics assessments to elementary grades: The development of a competency model and a rational number learning progression. *ETS Research Report Series*, *2014*(1), 1–41.

Bailey, D. H., Hoard, M. K., Nugent, L., & Geary, D. C. (2012). Competence with fractions predicts gains in mathematics achievement. *Journal of Experimental Child Psychology*, *113*(3), 447–455. https://doi.org/10.1016/j.jecp.2012.06.004

Barnes, N., Fives, H., & Dacey, C. M. (2017). U.S. teachers' conceptions of the purposes of assessment. *Teaching and Teacher Education*, *65*, 107–116. https://doi.org/10.1016/j.tate.2017.02.017

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17(2), 191-204.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389–396.

Betebenner, D. (2009), Norm- and Criterion-Referenced Student Growth. Educational Measurement: Issues and Practice, 28: 42-51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Booth, J. L., & Newton, K. J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology*, *37*(4), 247–253. https://doi.org/10.1016/j.cedpsych.2012.07.001

Binet, A., & Simon, T. (1916). *The development of intelligence in children (the Binet-Simon scale)*. Leopold Classic Library. Translated by Elizabeth S. Kite.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318. https://doi.org/10.1080/0969594042000304609

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology*, *27*(5), 777–786. https://doi.org/10.1037/0012-1649.27.5.777

Clements, D.H., & Sarama, J. (Eds.). (2004). Hypothetical Learning Trajectories: A Special Issue of Mathematical Thinking and Learning (1st ed.). Routledge. https://doi.org/10.4324/9780203063279

Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*. http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Confrey, J., Maloney, A. P., & Corley, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM*, *46*(5), 719–733. https://doi.org/10.1007/s11858-014-0598-7

Curriculum Associates. (2018). *i-Ready Assessments Technical Manual* (Curriculum Associates Research Report RR 2018-47). Curriculum Associates.

Domingue B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*(1), 1–19. https://doi.org/10.1007/s11336-013-9342-4

Duschl, A. R., Schweingruber, A. H., & Shouse, W. A. (2007). Taking science to school: Learning and teaching science in grades K-8. Washington DC: The national Academies Press.

Empson, S. B., Levi, L., & Carpenter, T. P. (2011). The algebraic nature of fractions: Developing relational thinking in elementary school. In *Early algebraization* (pp. 409–428). Springer.

Galton, F. (1883). *Inquiries into Human Faculty and Its Development*. Macmillan, London.

Hansen, N., Jordan, N. C., Fernandez, E., Siegler, R. S., Fuchs, L., Gersten, R., & Micklos, D. (2015). General and math-specific predictors of sixth-graders' knowledge of fractions. *Cognitive Development*, *35*, 34–49. https://doi.org/10.1016/j.cogdev.2015.02.001

Hegarty, M. (2019). Advances in cognitive science and information visualization. In D. Zapata-Rivera (Ed.) *Score reporting research and applications*. Routledge.

Kieren, T. E. (1976). *On the mathematical, cognitive and instructional*. *7418491*, 101.

Kieren, T. E. (1980). The rational number construct: Its elements and mechanisms. *Recent Research on Number Learning*, 125–149.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer, 3rd Edition.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark standard setting procedure. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed., pp. 225-254). New York: Routledge.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G. N., & Forster, M. (1996). Developmental Assessment: Assessment Resource Kit. Hawthorn: ACER Press.

National Research Council. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10019

National Research Council (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press. doi: 10.17226/18290

Otero, V.K. and Nathan, M.J. (2008), Preservice elementary teachers' views of their students' prior knowledge of science. J. Res. Sci. Teach., 45: 497-523. https://doi.org/10.1002/tea.20229

Peck, F., Johnson, R., Briggs, D., & Alzen, J. (2021). Toward learning trajectory-based instruction: A framework of conceptions of learning and assessment. *School Science and Mathematics*, 121, 357– 368. https://doi.org/10.1111/ssm.12489

Perie, M., Marion, S. F., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5–13.

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science*, *23*(7), 691–697. https://doi.org/10.1177/0956797612440101

Thurstone, L. L. (1925). A method of scaling psychological and educational test. *Journal of Educational Psychology*, 16, 433-451.

Tong, Y., & Kolen, M. J. (2007). Comparisons of Methodologies and Results in Vertical Scaling for Educational Achievement Tests. *Applied Measurement in Education*, *20*(2), 227–253. https://doi.org/10.1080/08957340701301207

Torbeyns, J., Schneider, M., Xin, Z., & Siegler, R. S. (2015). Bridging the gap: Fraction understanding is central to mathematics achievement in students from three different continents. *Learning and Instruction*, *37*, 5–13. https://doi.org/10.1016/j.learninstruc.2014.03.002

Wilkins, J. L. M., & Norton, A. (2018). Learning progression toward a measurement concept of fractions. *International Journal of STEM Education*, *5*(1), 27. https://doi.org/10.1186/s40594-018-0119-2

Wilson, M. (2023). *Constructing Measures: An Item Response Modeling Approach*. Routledge, 2nd Edition.

Wright, V. (2014). Towards a hypothetical learning progression for rational number. *Mathematics Education Research Journal*, *26*(3), 635–657. https://doi.org/10.1007/s13394-014-0117-8

Yulia, Y., Musdi, E., Afriadi, J., & Wahyuni, I. (2019). Developing a hypothetical learning progression of fraction based on RME for junior high school. *Journal of Physics*, 9.

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, *21*, 442–463

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results. *Educational Assessment*, *21*(3), 215–229.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26

Zenisky, Keller & Park. (2019). {} In D. Zapata-Rivera (Ed.) *Score reporting research and applications*. Routledge.

Zwick, R., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues & Practice*, 20(2), 15-25.

Table 1. Illustration of Process Used to Assign Lessons and Assessment Items by Learning Progression Level

LP Level	*i-Ready Diagnostic* Assessed Skill	Associated Lesson	# of Items
4. Operator	Student represents the division of two fractions as a division expression or equation when a verbal description or model is provided, or as a multiplication expression or equation when a division expression or equation is provided.	6.7 Divide with Fractions	5
	Student multiplies or represents the multiplication of a fraction less than 1 by a fraction less than 1 or a whole number, presented without a real-world context and/or with the aid of a visual model.	5.13 Understand Products of Fractions	14
3. Measurement	Student adds and subtracts fractions and mixed numbers with like denominators) without composing or decomposing wholes or uses visual models to represent these problems.	4.16 Add and Subtract Fractions	11
	Student recognizes fractions equivalent to a named fraction, using a visual model showing two or more equivalent wholes partitioned into different numbers of parts (e.g., area models, fraction strips, labeled number lines).	3.16 Understand Equivalent Fractions	8
2. Quotient	Student expresses a fraction, a/b, as a division expression, a ÷ b, or a division expression, a ÷ b, as a fraction, a/b, where a and b are represented symbolically, or represented numerically with b > a.	5.12 Fractions as Division	4
1. Part-Whole	Student names part of a whole using a fraction (denominator of 2, 3, or 4). (All models show equal parts. Area is not mentioned for unit fractions.)	3.14 Understand What a Fraction Is	12

Table 2. Item Difficulty Models for Fractions LP

	(1)	(2)
Quotient Level	31.32***	23.32***
	(6.56)	(6.26)
Measurement Level	51.44***	38.36***
	(5.27)	(5.29)
Operations Level	75.93***	49.30***
	(5.35)	(6.29)
Grade Dummy Variables	N	Y
Observations	357	357
R^2	0.40	0.48
Residual Std. Error	27.8	26.08
F Statistic	78.97	79.55

Note: ****p<0.01. Omitted LP Level is Part-Whole*

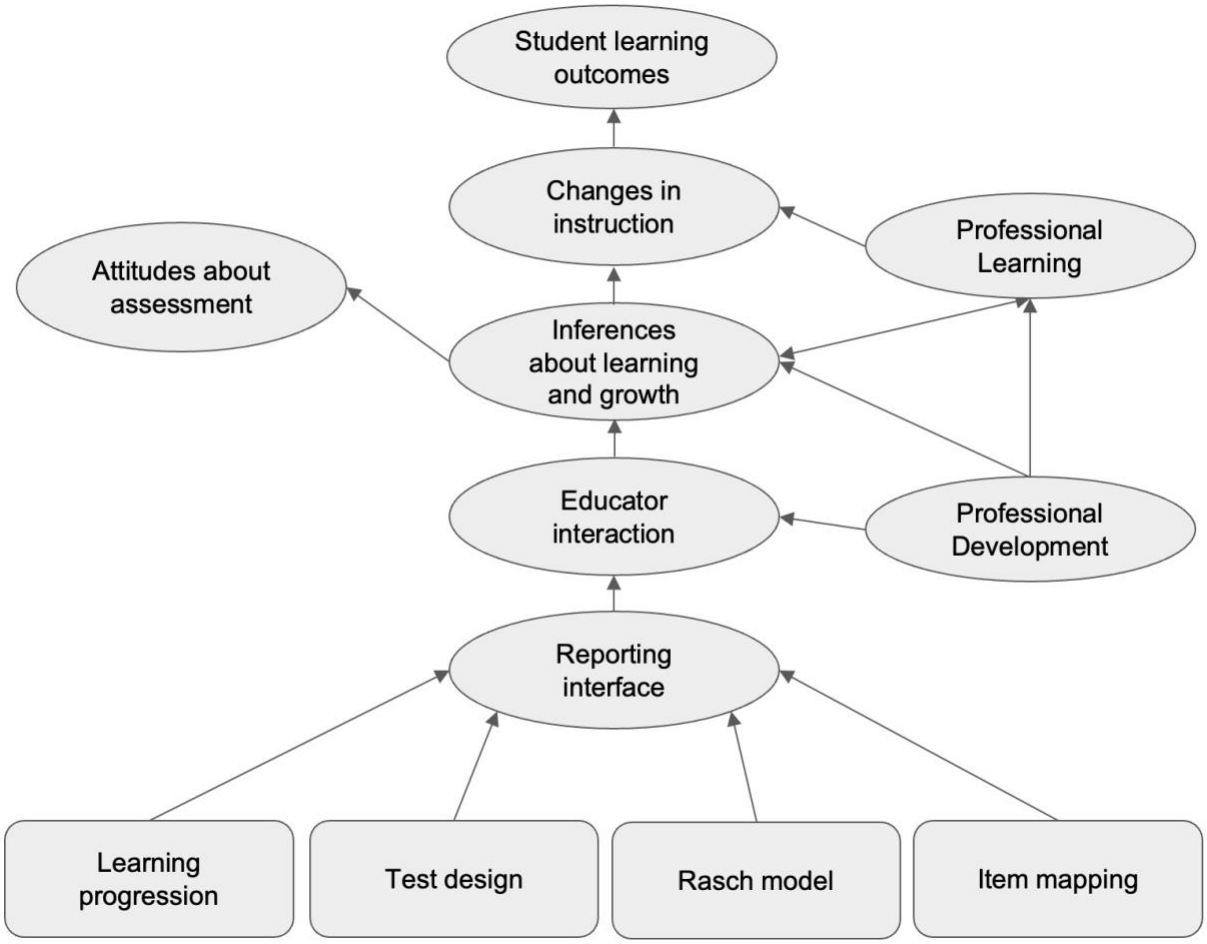

Figure 1. Theory of Action for the CRG Approach

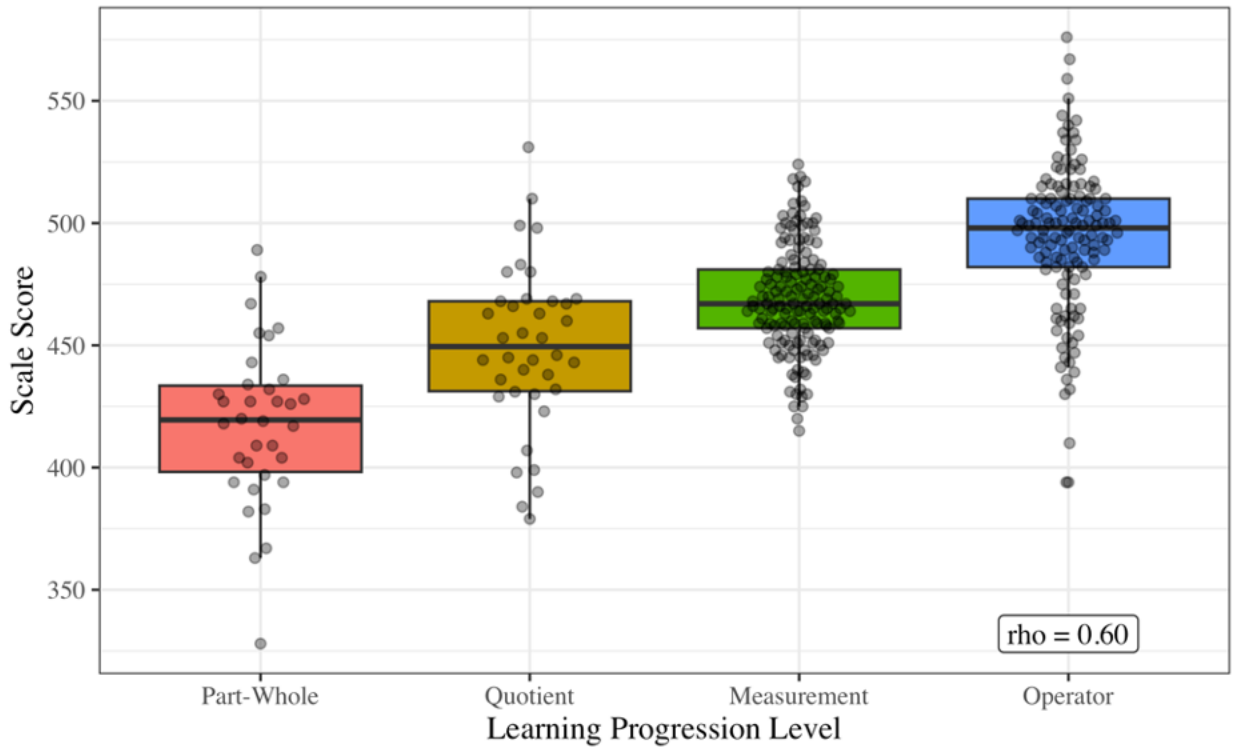

Figure 2. Comparing Distributions of Item Difficulties by LP Level

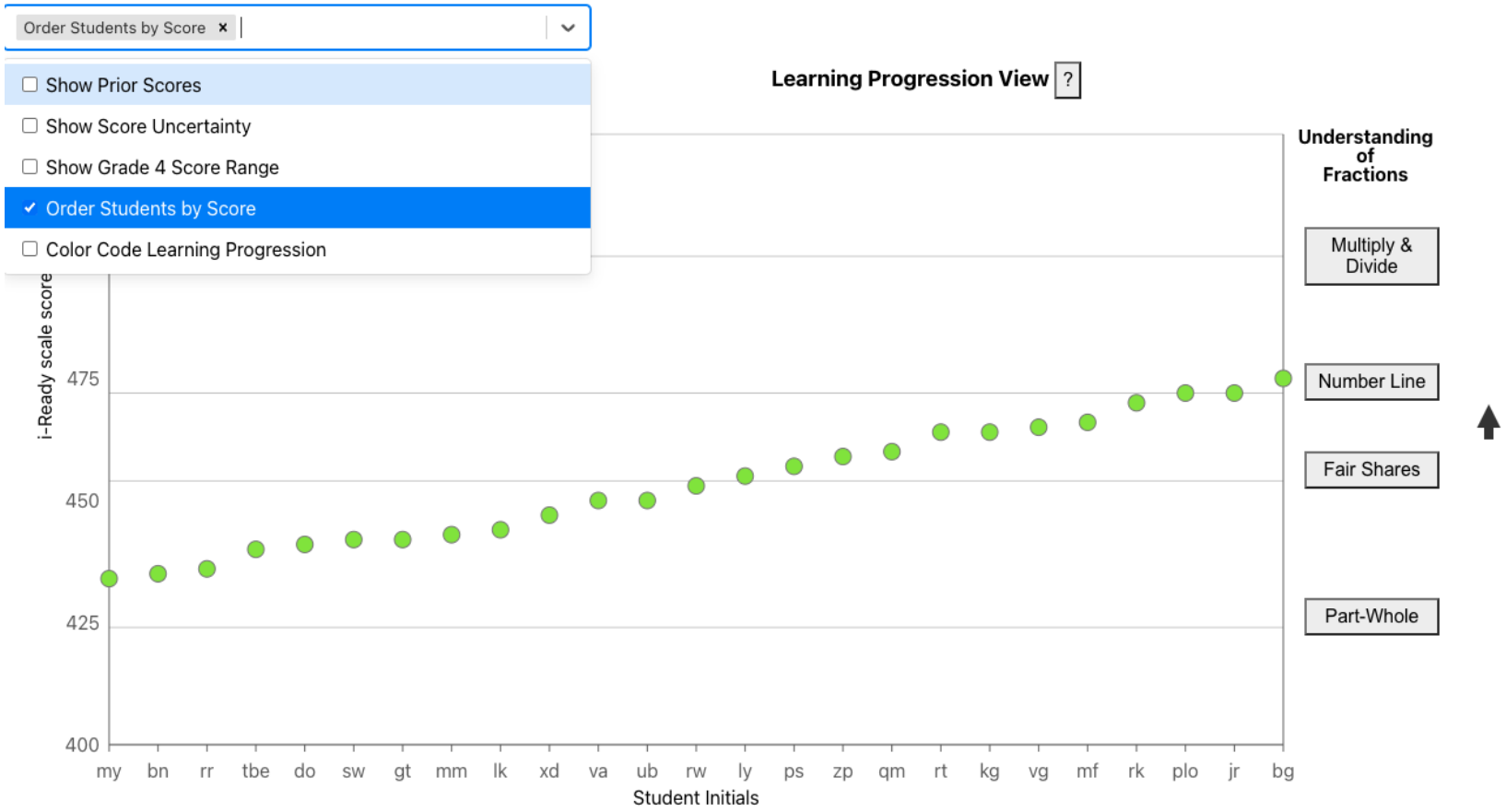

Figure 3. Screenshot of the Class View of the CRG Reporting Prototype

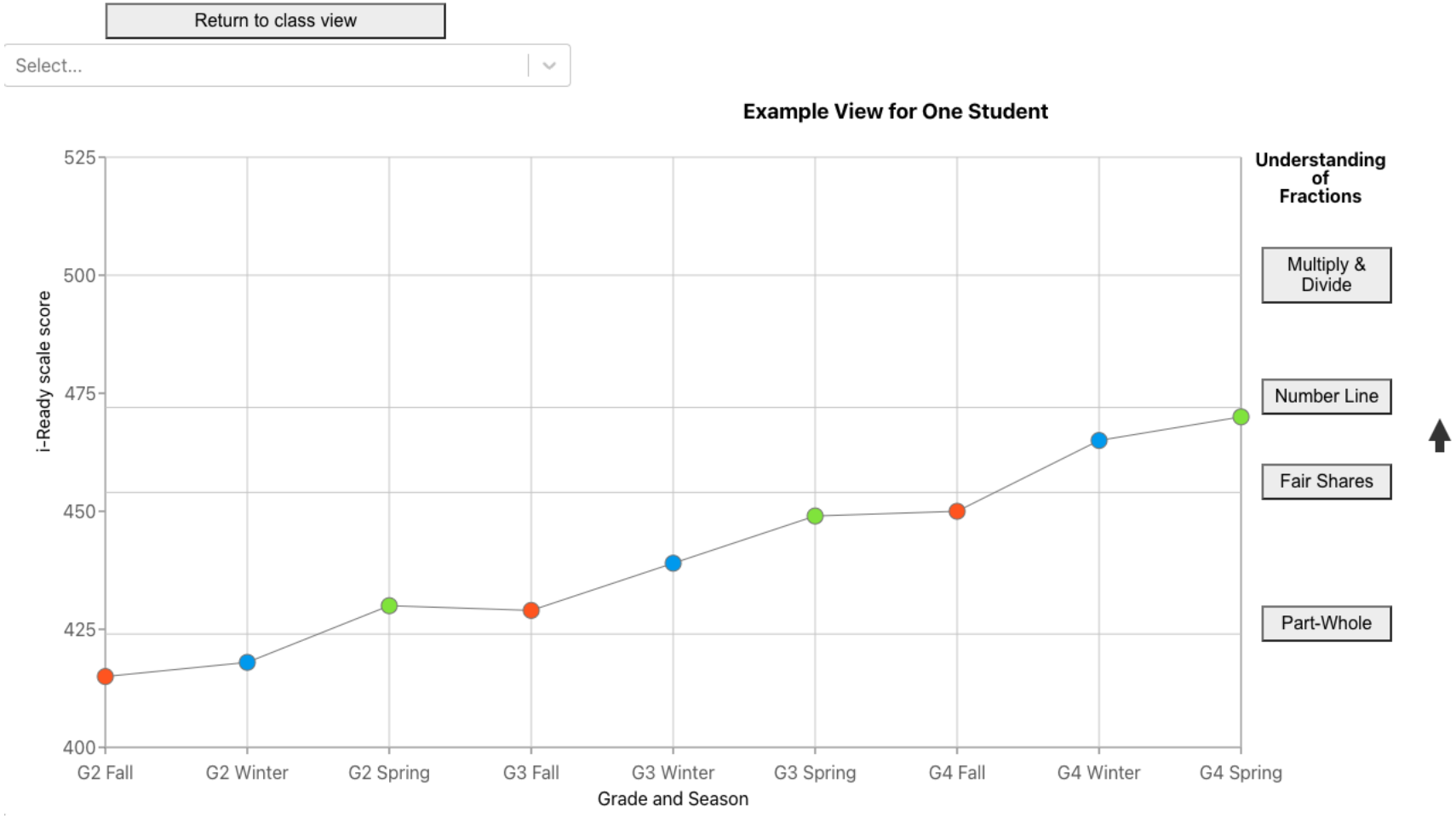

Figure 4. Screenshot of Student View of the CRG Reporting Prototype

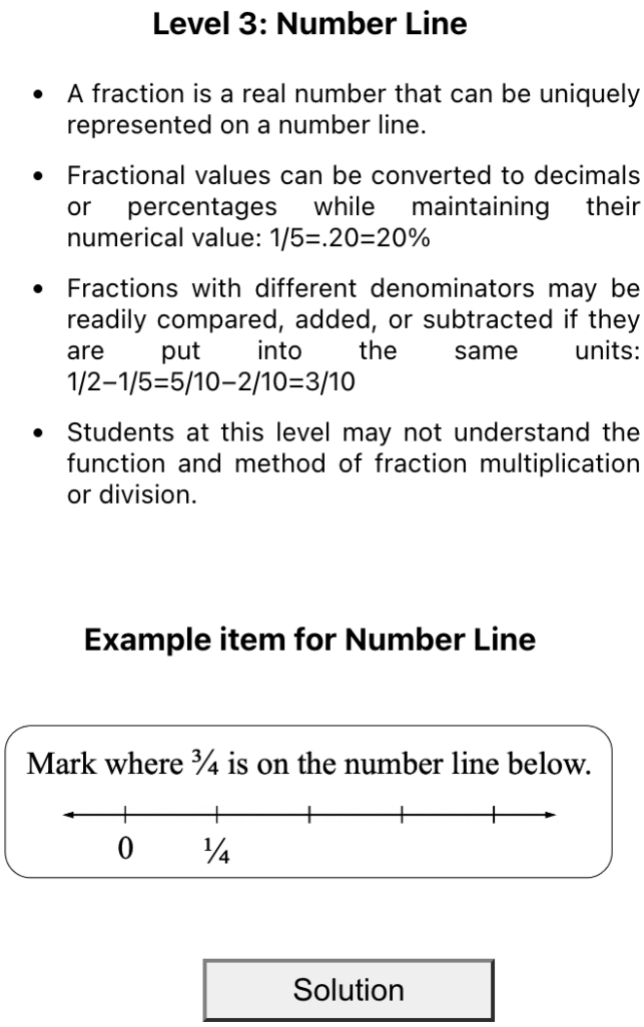

Figure 5. Screenshot of LP Level Specific Information when user clicks on gray box for "Number Line"

Example item for Fair Shares

Three friends are sharing two chocolate bars. How could they split the bars so that each friend gets the same amount of chocolate? How many bars worth of chocolate does each friend get?

Example item for Number Line

Mark where ¾ is on the number line below.

Figure 6. Exemplar Items Used to Distinguish Level 2 (Fair Shares) and 3 (Number Line) of Fractions LP

Appendix

Table A-1. Understanding Fractions Learning Progression

Interpretation	Student Characteristics	Item Responses
Operator	**Understands that:** • Multiplying a value by a fraction $\frac{a}{b}$ results in a value that is a-$b$ths of the original value • Understands the difference between multiplying and dividing fractions	**Is able to:** • Use multiplication to find a portion of a value • Determine that multiplying a value by a fraction with magnitude less than 1 will result in a value with smaller magnitude and multiplying by an improper fraction will result in a value with larger magnitude, and vice versa for division, without performing the calculations • Divide a value by a fraction
Measurement	**Understands that:** • Fractions represent unique numerical values • Two fractions are equivalent if they represent the same numerical value • Fractional values can be converted to decimals or percentages while maintaining their numerical value • Improper fractions may be rewritten as mixed numbers and vice versa • Fractions with different denominators may be compared or added if they are put into the same units **May not yet understand that:** • Fractions may be written as ratios and may represent part-part relationships or rates	**Is able to:** • Create and identify equivalent fractions, including converting between improper fractions and mixed numbers • Order fractions and mixed numbers with different numerators and different denominators • Add and subtract fractions and mixed numbers with different denominators **Common Errors:** • Treating all ratios as part-whole • Treating rates as two independent values with different units

Quotient	**Understands that:** • Fractional parts must be equal ("fair shares") but may not appear the same • The fraction $\frac{a}{b}$ represents the division of a by b • Unit fractions can be iterated to reproduce the original whole or part of the whole • Dividing the same whole into more parts (larger denominator) results in smaller unit pieces **May not yet understand that:** • A fraction has its own specific value that can be uniquely placed on a number line. • The same fractional value may be represented in multiple ways	**Is able to:** • "Share" a whole between a specified number of groups • Identify unit fractions • Use unit fractions $\left(\frac{1}{b}\right)$ to reproduce composite fractions $\left(\frac{a}{b}\right)$, including the whole $\left(\frac{b}{b}\right)$ • Compare fractions with the same numerator and different denominators • Add and subtract composite fractions with the same denominator **Common Errors:** • Misplacing a fraction on a number line • Incorrectly comparing two fractions with different numerators and different denominators • Not recognizing improper fractions as valid
Part-Whole	**Understands that:** • A fraction represents a specified number of parts out of the total number of parts **May not yet understand that:** • A whole must be partitioned equally • All parts of the whole must be used when partitioning	**Is able to:** • Identify the number of specified and total parts in an area model or in a described situation. • Compare fractions with the same denominator and different numerators **Common Errors:** • Making unequal parts or fail to exhaust the whole when attempting an equipartitioning task • Treating the numerator and denominator of a fraction as unrelated values