Using Learning Progressions to Design Vertical Scales that Support Coherent Inferences about Student Growth

Derek C. Briggs

Frederick A. Peck

University of Colorado, Boulder

April 2, 2015

ABSTRACT

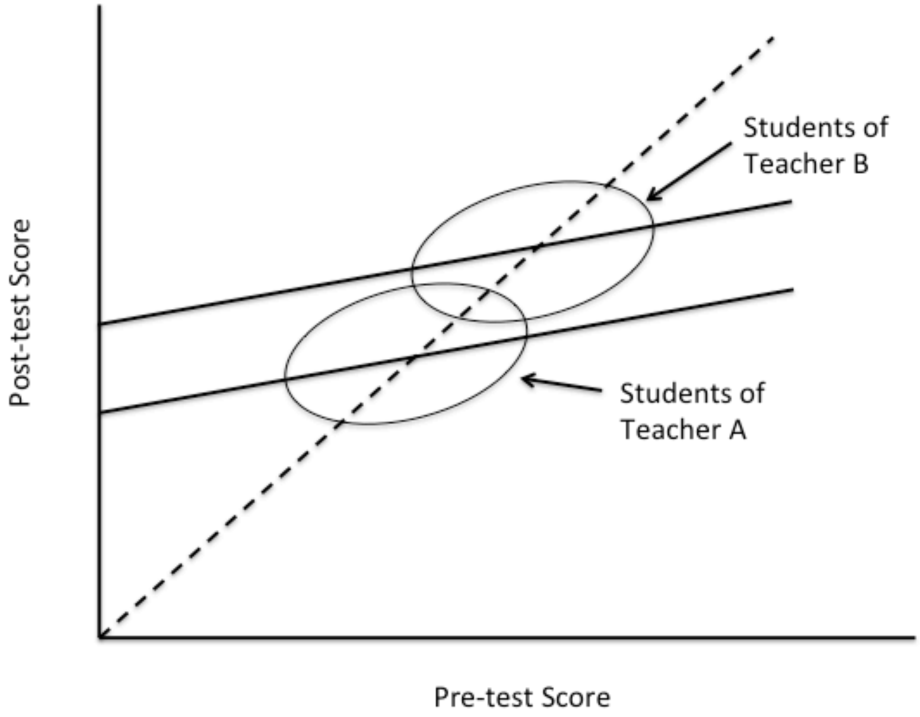
The concept of growth is at the foundation of the policy and practice around systems of educational accountability. It is also at the foundation of what teachers concern themselves with on a daily basis as they help children learn. Yet there is a disconnect between the criterion-referenced intuitions that parents and teachers have for what it means for students to demonstrate growth, and the primarily norm-referenced metrics that are used to infer growth. One way to address this disconnect would be to develop vertically linked score scales that could be used to support both criterion-referenced and norm-referenced interpretations, but this hinges upon having a coherent conceptualization of what it is that is growing from grade to grade. In this paper a learning progression approach to the conceptualization of growth and the subsequent design of a vertical score scale is proposed and illustrated in the context of the Common Core State Standards for Mathematics.

KEYWORDS: Growth, Vertical Scaling, Learning Progressions, Educational Accountability

Introduction

More than 10 years have passed since the advent of No Child Left Behind, and if anything has changed about the nature of educational accountability it is the increasing emphasis on using evidence of growth in student learning to evaluate the efficacy of teachers and schools. To a great extent this represents an improved state of affairs, since it implicitly recognizes that it is unfair to compare teachers on the basis of what their students have achieved at the end of a school year without taking into consideration differences in where the students began at the outset. Yet when researchers build models to quantify the contribution of teachers to growth in student learning, growth does not always mean what laypeople naturally think it means. This can lead to fundamental misunderstandings.

Figure 1. Growth, Effectiveness and Two Hypothetical Teachers



To appreciate why, consider the graphic shown in Figure 1. The axes of the plot represent scores from the *same* test given at the beginning of a school year (pre-test on the horizontal axis) and the end of a school year (post-test on the vertical axis). The ellipses within the plot capture different collections of data points corresponding to the students of two different teachers, Teacher A and Teacher B. The dashed line at the 45 degree angle indicates a score on the posttest that is identical to a score on the pretest. To keep the scenario simple, assume each teacher has the same number of students. On this basis of this data collection design, two researchers are asked to compare the teachers and make a judgment as to which one is better. Researcher 1 computes the average test score gains for both groups of students and gets identical numbers. This researcher concludes that students in each classroom have grown by the same amount, hence neither teacher can be inferred to be better than the other. This can be seen in Figure 1 by noticing that each teacher's class of students has about the same proportion of data points above the dashed line (indicating a pre to post gain) as they do below (indicating a pre to post loss). Researcher 2 takes a different approach. This researcher takes all the available data for both classes of students, and proceeds to regress post-test scores on pre-test scores and an indicator variable for Teacher B. The parameter estimate for the Teacher B indicator variable is large and statistically significant. This can be seen in Figure 1 by noticing that the regression line (solid black line) passing through the data ellipse for teacher B is higher (has a larger y-intercept) than the regression line for teacher A. The second researcher concludes that B is the better teacher because given how they scored on the pre-test, the students of teacher B scored higher than the students of teacher A. Who is right?

Many readers will have immediately recognized the example above as a retelling of *Lord's Paradox* (Lord, 1967) with the classrooms of Teachers A and B substituted for males and females, and test scores substituted for weight. Holland and Rubin (1983) reconciled Lord's Paradox by essentially pointing out that the two cases involved analyses pertaining to fundamentally different causal inferences. The same logic can be used for the example above. Researcher 1 is inferring the effect of Teacher B relative to Teacher A through a comparison of average score gains. Researcher 2 is inferring the effect of Teacher B relative to Teacher A by comparing the average difference in post-test scores for those students *with the same pre-test scores*. Both researchers could argue that they are making comparisons on the basis of student growth. Researcher 1 defines growth as the change in magnitude from pre-test to post-test. Researcher 2 defines growth as the increment in achievement we would predict if two students with the same pre-test score had Teacher B instead of Teacher A. Which one has come to the right conclusion about the effect of one teacher relative to the other?

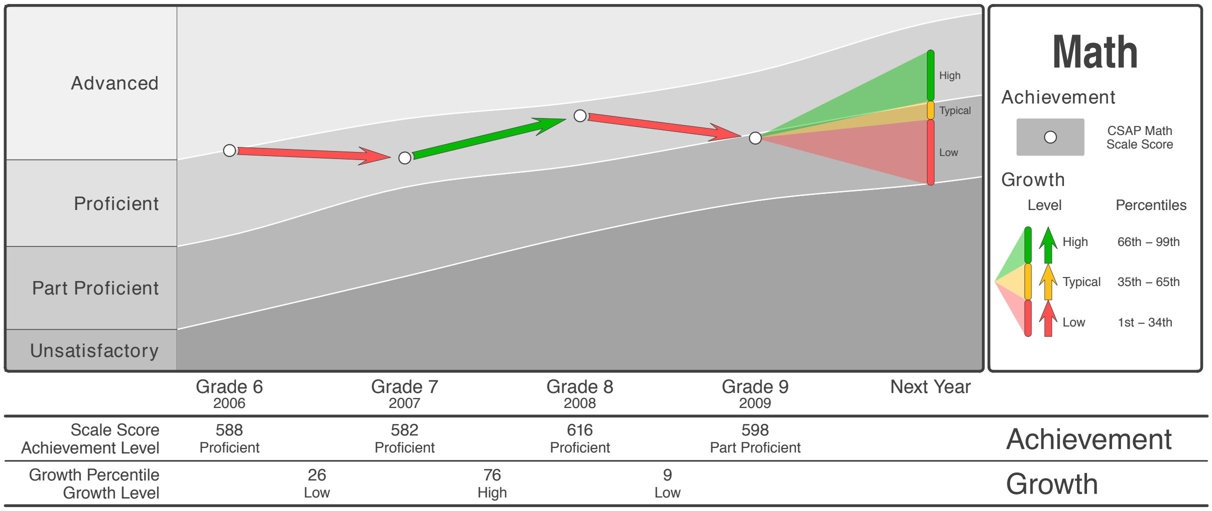
Most of the growth and value-added models that play a central role in teacher evaluation follow the approach of Researcher 2 (c.f., Chetty, Friedman & Rockoff, 2014; Kane & Staiger, 2008; McCaffrey, Lockwood, Koretz & Hamilton, 2003). A root of considerable confusion in the interpretation of estimates from such models is that important stakeholders in K-12 education—teachers, parents, the general public—assume that inferences about effectiveness derive from the sort of approach taken by Researcher 1. Put differently, judgments about the quality of a student's schooling are not based on direct estimates of the amount that a student has learned, but rather, on how well a student has performed relative to peers who are comparable with respect to variables such as prior achievement, free and reduced lunch status, race/ethnicity, etc. Yet while econometricians and statisticians may notice and appreciate the distinction between growth as measured by differences in quantity vs. growth as inferred by normative comparison, teachers, parents and the general public do not. And to some extent, this misconception is encouraged by the way results from these models are presented. Consider, for example, the Policy and Practitioner Brief released by the *Measures of Effective Teaching Project* entitled "Ensuring Fair and Reliable Measures of Effective Teaching." In the Executive Summary, the first key finding is presented as follows:

> Effective teaching can be measured. We collected measures of teaching during 2009–10. We adjusted those measures for the backgrounds and prior achievement of the students in each class. But, without random assignment, we had no way to know if the adjustments we made were sufficient to discern the markers of effective teaching from the unmeasured aspects of students' backgrounds. In fact, we learned that the adjusted measures did identify teachers who produced higher (and lower) average student **achievement gains** following random assignment in 2010–11. **The data show that we can identify groups of teachers who are more effective in helping students learn**. Moreover, the **magnitude** of the **achievement gains that teachers generated** was consistent with expectations. (MET Project, 2013, pp. 4-5, emphasis added)

The Measures of Effective Teaching policy brief was very intentionally written for a general audience of policymakers and practitioners in education. Note that in the passage above "learning" is equated to "achievement gains." Since student achievement is typically inferred from test performance, most readers of this policy brief would interpret achievement gains as implying test score gains. The larger the magnitude of test score gains, the more that a student has learned. However, this reading of the passage above would be incorrect. The MET study was able to show that differences in prior estimates of teacher value-added was strongly predictive of differences in *relative* student achievement following random assignment. Teachers flagged as effective only produced "gains" in the sense that their students scored higher, on average, than they would have had they instead been assigned to a less effective teacher. In this context then, a score "gain" could plausibly mean a true decrease in learning that was less than expected. Notions of growth in such contexts are fundamentally normative; effective and ineffective teachers are *guaranteed* to be found in any population of teachers – whether the actual amount of student learning is high, low, or even nonexistent.

For another example, this time with individual students as the units of analysis, consider the way that growth is communicated in Colorado (and many other states) using student growth percentiles (SGPs) computed using the Colorado Growth Model (CGM; Betebenner, 2009). The publicly available tutorial about the CGM can be found at http://www.cde.state.co.us/schoolview/growthmodeltutorials. Also see Castellano and Ho (2013a, 2013b). In a nutshell, an SGP attempts to show how a student's achievement at the end of the year compares with that of other students who started the year at the same level. SGPs can be interpreted as indicating the probability of observing a score as high or higher than a student's current score, given what has been observed on *all* of the student's prior scores. A student with an SGP of 75 has a current year test score that is higher than 75% of peers with a comparable test score history. It follows that the probability of observing a score this high or higher for any student with a comparable test score history is 25%. An SGP supports inferences about growth in the sense that if two students started at the same achievement level at the beginning of the year and one scores higher than the other on a test at the end of the year, it seems reasonable to infer that the student with the higher score has demonstrated more growth. Betebenner and colleagues have also made it possible to weave criterion-referenced information into the CGM by comparing each student's SGP to her adequate growth percentile—the growth percentile that would be needed to achieve a desired performance level on a test. This makes it possible to answer the question, is the growth a student has demonstrated good enough relative to the standards that have been established and enacted by the state?

Figure 2. Example of a Student Growth Report in Colorado



Source: http://www.schoolview.org/documents/ISR_explanation.pdf

Yet results from the Colorado Growth Model are also easy to misinterpret. Many teachers and parents are likely to equate a student's score with "math knowledge." Teachers and parents with this interpretation would think that a student's score should be steadily increasing across grades. If presented with a scenario in which a student has a score of 500 across grades 6-8, it would be natural for a parent to think that the student has not "learned anything" during these years. However, if the meaning of a score of 500 changes every year, this would not be a correct inference.

SGPs can easily be misunderstood as "changes in math knowledge", i.e., "amount of learning". For example, if a student has an SGP of 90, 75 and 60 across grades 6 through 8, it would be natural for a parent to interpret this to mean that the student is learning less in the grade 8 than in grade 7, and less in grade 7 than in grade 6. But such an inference would be impossible to support on the basis of SGP comparisons alone. For all its advantages, the CGM cannot be used to infer whether the amount a student has learned in the most recent year is significantly more or less than the amount a student learned in the past year.

Nonetheless, there is good reason to suspect that parents and teachers are implicitly encouraged to use it in this manner, as illustrated by the plot in Figure 2. This exemplar plot is made available to parents in order to help them interpret their child's SGP. The vertical consists of scale scores in mathematics, organized into proficiency levels. The thresholds for these levels change from grade to grade because Colorado has standards that become more and more difficult to reach as students enter middle school. Grade levels are shown along the *x*- (horizontal) axis.. Below the horizontal axis are scale scores and SGPs. The body of the plot includes small circles that indicate the student's scale score and a gray gradation that indicates proficiency levels, shown along the *y*- (vertical) axis. The location of the small circle thus indicates a student's scale score in a given grade and where that scale score is located relative to three proficiency level thresholds. Note that in addition to the circles there are color-coded arrows that indicate whether a student's SGP in a given grade is "low" (1-34), "typical" (35-65), or "high" (66-99).

The "next year" spot on the x-axis is meant to reflect the most likely proficiency levels of the student if the student were to have a low, typical or high SGP in the following year. Visually, the first thing a parent is likely to interpret is the trajectory implied by the collective slopes of the individual arrows, and the height of the bar segments in the "next year" prediction. The visual interpretation suggests that the student represented in this plot showed flat or slightly negative growth from grade 6 to 7, positive growth from 7 to 8, and negative growth from 8 to 9. If the student has positive growth from grade 9 to 10, she will fall within the proficient performance level; if the student has flat or negative growth the student will fall within the partially proficient performance level. Across grades 6 through 9, the overall growth trajectory appears relatively flat. Since the likely interpretation of this trajectory is “change in knowledge”, it appears that the student has learned nothing between grades 6-8. This inference is supported by the direction and color of the arrows that constitute the trajectory: the downward pointing red arrows support the inference that the student endured two years of negative growth, which was compensated by one year of positive growth indicated by the green, upward pointing arrow. The overall picture is that the student’s knowledge has not really changed.

In education and in life there is a constant tension between norm and criterion-referenced interpretations. Neither can be sustained in perpetuity without eventually encountering the need to invoke the other. In this article we argue that normative interpretations about student growth and teacher effectiveness need to be complemented by criterion-referenced interpretations about *how much* and *of what*? *How much* has my child grown this year? How much more has she grown relative to last year? *What* did my child learn and how can the effectiveness of my child’s teacher be quantified relative to the amount that was learned? In theory, the best way to answer such questions would be through the development of tests that could be expressed on vertically linked scales. In the next section we explain why, to date, *vertical scaling* appears to have been unsuccessful at meeting such ambitions. In the section that follows, we propose a new approach to the design of vertical scales that is premised upon a priori hypotheses about growth in the form of a learning progression hypothesis. In a nutshell, our argument is that meaningful criterion-referenced interpretations of growth magnitudes can only be supported when they follow from a coherent conceptualization of what it is that is growing over time. To speak of a student’s growth in “mathematics” is incoherent, because mathematics is just a generic label for the content domain of interest, and not an attribute for which it makes sense to speak of a student having more or less. A benefit of designing a vertical scale according to a learning progression is that it becomes possible to speak about growth in terms of specific knowledge and skills that are hypothesized to build upon one another over time. We illustrate this using a learning progression that shows how students develop the knowledge and skills necessary to be able to analyze and reason about proportional relationships.

Some Background on Conventional Vertical Scaling Methodology

The conventional method for creating a vertical scale is documented in books[1] such as *Educational Measurement* (4th Edition), *Test Equating, Linking and Scaling*, and *The Handbook of Test Development*. Although there are a number of different ways to create a vertical score scale, the approach generally consists of two interdependent stages: a *data collection stage* and a *data calibration stage*. In the data collection stage, the key design principle is to select a set of common test items (also known as "linking" items) that will be administered to students across two or more adjacent grade levels (e.g. grades 3 through 4, grades 3 though 8, etc.). This is in contrast to a unique test item, which would only be administered to students at any single grade. In some designs, the common items consist of an external test given to students across multiple grades; in others they consist of an external test given only across adjacent grades; and in others they consist of items embedded within operational test forms. Once item responses have been gathered for representative students at each grade level, the next task is to analyze differences in performance on the common items. These differences become the basis for the data calibration stage. In order to calibrate the responses from students at different grade levels onto a single scale, either the ability of the students, or the characteristics of the items (e.g., difficulty) needs

[1] Kolen, 2006, pp. 171-180; Kolen & Brennan, 2004, pp. 372-414; Young, 2006; pp. 469-485.

to be held constant across grades. Since growth in student ability across successive grades is the underlying basis for the vertical scale, the only the only reasonable option is to hold the item characteristics constant. There are two known approaches for accomplishing this, Thurstone Scaling (Thurstone, 1925; 1927) and Item Response Theory scaling (IRT; Rasch, 1960; Lord & Novick, 1968). IRT-based methods are by far the predominant approach and have been since the mid 1980s. The selling point of IRT is the property of parameter invariance, which will hold so long as the assumption of local independence has been satisfied, and the data can be shown to fit the item response function that has been specified. Parameter invariance is the critical property of IRT models that makes it possible to establish values for the characteristics of common items that do not depend on the particular group of students responding to them. When parameter invariance holds, the same difficulty parameter will be estimated for an item whether it is administered to a 3rd grade student or an 8th grade student. An even stronger invariance property, that of invariance of *comparisons* (i.e., specific objectivity) must hold when specifying the Rasch Model, and this can have implications for claims that a scale has equal intervals (Briggs, 2013).

Much of the research literature on vertical scales has focused on choices that must be made in the calibration of the scale (c.f., Skaggs & Lissitz, 1986). Two choices in particular have received considerable attention: the functional form of the IRT model, and the manner in which tests scores across grades are concatenated. The first choice is typically a contrast between the use of the three parameter logistic model (3PLM; Birnbaum, 1968) or the Rasch Model (Rasch, 1960). The second choice is a contrast between a separate or concurrent calibration approach. In the separate approach item parameters are estimated separately for each grade-specific test. Then a base grade for the scale is established and then other grades are linked to the base grade after estimating linking constants for each grade-pair using the Stocking-

Lord approach (Stocking & Lord, 1983). In the concurrent approach, all item parameters are estimated simultaneously. Although there is very little in the way of consensus in the research literature about the best way to calibrate a vertical scale, when different permutations of approaches have been applied to create distinct scales from the same data, this has been shown to have an impact on the magnitudes of grade to grade growth (Briggs & Weeks, 2009). One message that has been communicated by this research base is that there is no "right answer" when it comes to creating a vertical scale. If this message is taken to its extreme, it implies that nonlinear transformations can be employed to the scale following the calibration stage to produce whatever depiction of growth is most desirable to stakeholders, since no one depiction can be said to be more accurate than the other.

In a review of the vertical scaling practices among states as of 2009, Dadey and Briggs (2012) found that 21 out of 50 states had vertically scaled criterion-referenced assessments spanning grades 3 through 8. Notably, Dadey and Briggs found no evidence that those states with vertical scales used their scales to make inferences about criterion-referenced growth at the student, school or state levels. In many cases, it appears that states did not actually trust the aggregate inferences about student growth implied by their vertical scales. For one of the more ironic examples, Colorado, the originators of the norm-referenced Colorado Growth Model, also expressed its criterion-referenced tests in math and reading along a vertical scale. This fact would come as a surprise to most Colorado educators[2], because grade to grade scale score gains are never emphasized in conjunction with the reporting of SGPs. In another instance, as part of the process of designing their vertical scale, contractors for the state of Arizona applied a

[2] Indeed, the second author of this paper, who taught high school mathematics in Colorado as recently as 2012-13, was completely unaware that math scale scores in Colorado had been linked vertically until informed of this by the first author. This even extends to personnel at the Colorado Department of Education who work in the educational accountability group, who on one occasion in correspondence with the first author insisted that Colorado's tests were not vertically scaled.

nonlinear transformation to ensure that grade to grade reading score scale means would increase monotonically, even though the empirical evidence prior to applying the transformation indicated that students in some upper grades had performed slightly *worse* on items that were common to the lower grade.

One possible explanation for the reluctance of states to use their vertical scales to report growth in terms of grade-to-grade changes in magnitudes is that there is a disconnect between the information about growth that such scales imply, and the intuitive expectations about growth common among teachers, parents and the public—the primary audience for the communication of growth. Namely, the intuitive expectation is that as students learn they build a larger and larger repertoire of knowledge and skills that they can use to navigate the world around them. As such, irrespective of the subject in which this repertoire of knowledge and skills is to be measured, from year to year one would expect to see significant evidence of growth. In contrast to this intuition, many vertical scales show evidence of a large deceleration of growth, particularly as students transition from the elementary school grades to the middle school grades (Tong & Kolen, 2007, Dadey & Briggs, 2012). In addition, because the concept of growth borrows so heavily from the analogy of measuring height, it is intuitive to believe that the interpretation of gains from one to grade to another along a vertical scale do not depend upon a student's initial location on the scale. Indeed, Briggs (2013) argues that the premise of equal-interval interpretations has been central to the way that some testing companies have marketed the advantages to creating a vertical scale.

One response to this disconnect between intuition and practice is to say that both of the intuitions described above are wrong or at least in some sense misguided (Yen, 1986). For example, it could be argued that if students were tested repeatedly across grades to make inferences about their ability to decode and extract meaning from selected vocabulary words in a reading passage, that larger gains would be observed in the early grades of a child's schooling when decoding is a focus of instruction, and that these gains would be smaller in the later grades when the instructional focus shifts from "learning to read" to "reading to learn." Similarly, it could be argued that there is nothing inherent to the process of creating a vertical scale that would guarantee the scale has equal-interval properties. Because of this, statements along the lines of student X has grown twice as much as student Y are meaningless unless both students started at the same baseline—which brings us back normative growth inferences.

The problem with the approach of discrediting "faulty" intuitions in this manner is that it defeats the purpose of creating a vertical scale in the first place. In the first example we have an a clear instance of construct underrepresentation if a test claims to measure "reading" or "English Language Arts" yet really only measures the decoding of words. This would explain why growth decelerates, but would certainly not validate the inferences about growth that were purported. In the second example, if a vertical scale can only support inferences about ordinal differences among students, why create the vertical scale at all? As Briggs (2013) argues, the purpose of vertical scales is to facilitate inferences about changes in magnitude with respect to a common unit of measurement. The warrant behind this use is the assumption that changes along any point of the scale have an equal-interval interpretation. Therefore, to validate that a given vertical scale can be used for its intended purpose, evidence must be presented to support the equal-interval assumption.

We take the position that the best way to move the science behind vertical scaling forward is to place a greater emphasis on design issues. In making this case we are essentially sounding the same drum that was first pounded in the National Research Council's 2001 report

Knowing What Students Know (Pellegrino, Chudowsky & Glaser, 2001), a report which emphasized that principled assessment design always involves an implicit model of cognition and learning. Yet while this message has resulted in some important improvements in assessment design over the past decade (e.g., the application of "Evidence Centered Design" principles; Mislevy, Steinberg & Almond, 2002), it is less clear that the message has had much influence on the design of vertical scales. In the next section we use the subject area of mathematics to illustrate an approach to vertical scale design that is premised on what we call a *learning progression* conceptualization of growth.

Using Learning Progression Hypotheses to Design Vertical Scales

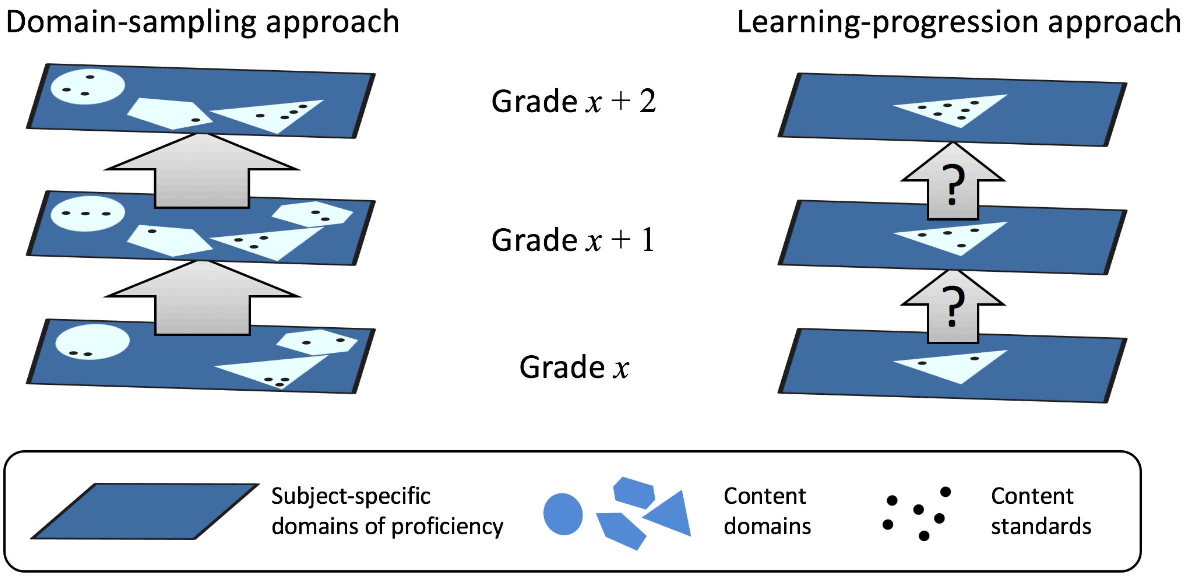
Domain-Sampling vs. Learning Progression Conceptualizations of Growth

Fundamental to the development of large-scale assessments for use in systems of educational accountability is a collection of content-specific targets for what students are expected to know and be able to do within and across grades. At present, through their participation in one of the two large-scale assessment consortia (the Partnership for Assessment of Readiness for College and Careers [PARCC] and the Smarter Balanced Assessment Consortium [SBAC]), many American states are using the Common Core of State Standards for Mathematics and English Language Arts (CCSS-M & CCSS-ELA) as the basis for these targets. A good case can be made that the Common Core of State Standards are especially amenable to the creation of vertical scales to support inferences about growth because these standards were written with any eye toward how students' knowledge and skills in mathematics and English

Language Arts would be expected to become more sophisticated over time. However, there are still two different ways that the concept of growth could be conceptualized before choosing a data collection design that could result in the calibration of a vertical scale. These different conceptualizations are illustrated in Figure 3 in the context of mathematics.

The left side of Figure 3 contains planes that are intended to encompass what it means to be "proficient" or "on track for college and career readiness in mathematics" at a given grade level (e.g., grade 3). Within each plane are light-colored shapes, and within each shape is a series of dots. The shapes are meant to represent different "content domains" (e.g., Numerical Operations, Measurement & Data, Geometry); the dots represent domain-specific performance standards that delineate grade-level expectations for students (e.g., within the domain of Measurement & Data: "Generate measurement data by measuring lengths using rulers marked with halves and fourths of an inch."). This sort of taxonomy has traditionally been used in the design of large-scale assessments to deconstruct the often amorphous notion of "mathematical ability" into the discrete bits of knowledge, skills, and abilities that should, in principle, be teachable within a grade-level curriculum. Such an approach facilitates the design of grade-specific assessments because test items can be written to correspond to specific statements about what students should know and be able to do. The growth target in such designs is not a cognitive attribute of the test-taker, but a composite of many, possibly discrete pieces of knowledge, skills, and abilities. We refer to the assessment design implied by the left side of Figure 3 as the *domain-sampling* approach.

Figure 3. *Different Construct Conceptualizations and Implications for Growth*



Under the domain-sampling approach, the intent is for growth to be interpreted as the extent to which a student has demonstrated increased mastery of the different domains that comprise mathematical ability. This is represented by the single arrow indicating movement from the plane for a lower grade to the plane for a higher grade. Note that if both the domains and the content specifications within each plane change considerably from grade to grade, then it becomes possible for students to appear to "grow" even if entirely different content is tested across years. This is represented in Figure 3 by the fact that two domains (circles and triangle shapes) are shown in each grade while one domain (hexagon shape) is only present in grades x and $x + 1$ and another (pentagon shape) is only present in grades $x + 1$ and $x + 2$. In the best case scenario for growth inferences, considerable thought has been put into the vertical articulation of the changes among content domains and standards from grade to grade. For example, according to the CCSS-M, a composite "construct" of mathematical ability could be defined from grade to grade as a function of 5 content domains and 6 skill domains (i.e., mathematical practices). Yet this leaves ample room for growth in terms of the composite to have an equivocal interpretation depending upon the implicit or explicit weighting of the domains in the assessment design and scoring of test items. Furthermore, the number of items required to make inferences about *all* CCSS domains at one point in time in addition to change over time is likely to be prohibitive. The problem of changing domains over time has been described as the problem of "construct shift" in context of research conducted by Joseph Martineau (Martineau, 2004; 2005; 2006). The basic argument is that most achievement tests are only unidimensional to a degree. At one point in time for specific grade level, ignoring minor secondary dimensions is unlikely to cause large distortions in inferences about student achievement. However, when the nature of the primary and second dimensions and their relative importance are themselves itself changing over time, the calibration of a single undimensional vertical has much greater potential to lead to distortions about student growth.

A different basis for a growth conceptualization comes from what we refer to as the *learning progression* approach. Learning progressions have been defined as empirically grounded and testable hypotheses about how students' understanding of core concepts within a subject domain grows and become more sophisticated over time with appropriate instruction (Corcoran, Mosher, & Rogat, 2009). Learning progressions provide "likely paths" (Confrey, 2012, p. 157) for learning, along with the instructional activities that support this path. The key feature of learning progressions is that they are developed by coupling learning theories with empirical studies of student reasoning over time. This is in contrast to some curricula that are developed based on disciplinary logic, or "reductionist techniques to break a goal competence into subskills, based on an adult's perspective" (Clements & Sarama, 2004, p. 83). Therefore, while there are many ways that understanding can develop over time, learning progressions capture particularly robust pathways that are supported by both learning theory and empirical studies of learning *in situ* (Daro, Mosher, & Corcoran, 2011; Sarama & Clements, 2009). As Daro et al. (2011, p. 45) explain,

> Evidence establishes that learning trajectories are real for some students, a possibility for any student and probably modal trajectories for the distribution of students.

At the same time, learning progressions are always somewhat hypothetical, and should be refined over time (Shea & Duncan, 2013).

This key idea is shown in the right panel of Figure 3, which depicts a hypothesis about the nature of growth: the way that students' understanding of some core concept or concepts *within the same domain* is expected to become qualitatively more sophisticated from grade to grade. The notion that this constitutes a hypothesis about growth to be tested empirically is represented by the question marks placed next to the arrows that link one grade to the next. In contrast to inferences about growth based on domain-sampling, changes in a student's depth of knowledge and skills within a single well-defined domain over time are fundamental to a learning progression conceptualization.

In mathematics, the distinction between across and within domain inferences about what students know and can do is evident in the fact that the CCSS-M makes it possible to view standards by grade (across domain emphasis, single point in time) or by domain (within domain emphasis, multiple points of time)[3]. Importantly, when math standards from the CCSS are viewed by domain and by grades 3 through 8, as in Table 1, it becomes evident that there is in fact good reason to be concerned about the potential for construct shift in how "mathematics" is

[3] http://www.corestandards.org/Math

being defined from grade 3-5 (elementary school) to grades 6-8 (middle school). Notice that the only content domain that remains present across the all six grades is geometry. This is why, well before worrying about technical issues in calibrating a vertical scale, it is important to first ask whether the vertical scale would allow for inferences about growth over time that are conceptually coherent. If all the content domains shown in Table 1 were to be the basis for a domain-sampling approach to the creation of a single vertical scale, what would it mean if a student grew twice as much from grade 4 to 5 as from grade 5 to 6? At least on the basis of the CCSS-M content domains, this would seem to be an apples to oranges comparison.

Table 1. Math Content Domains Associated with Grades 3 to 8 in the CCSS

Content Standards by Domain	Grade in Which CCSS Include Domain					
	3	4	5	6	7	8
Operations & Algebraic Thinking	X	X	X			
Number & Operations in Base 10	X	X	X			
Number & Operations—Fractions	X	X	X			
Measurement & Data	X	X	X			
Geometry	X	X	X	X	X	X
Ratios & Proportional Relationships			X	X		
The Number System				X	X	X
Expressions & Equations				X	X	X
Functions						X
Statistics & Probability				X	X	X

When taking a learning progression approach, one would eschew the notion of representing growth with a single composite scale for mathematics across grades 3 through 8 and instead choose a cluster of standards within a given domain and across a subset of grades as candidates for quantifying growth. So for example, a single learning progression might be hypothesized with respect to how students in grades 3 through 5 become increasingly sophisticated in the way that they reason and model numbers and operations that involve fractions. After designing and calibrating a vertical scale associated with this learning progression, two different pieces of information could be provided to a fourth grade student. A number summarizing the student's *composite achievement* across all math content domains tested in grade 4, and a measure pertinent to the student's *growth* along the vertical scale for numbers and operations that involve fractions. To be clear, these two numbers would derive from two different scales for two different purposes: one scale to characterize achievement status across domains, another scale to measure growth within a single, well-defined domain.

Example: A Learning Progression for Proportional Reasoning

The content domains in the CCSS-M, and the ways they are expected to change across grades as a function of their standards, provide a starting point for math education researchers and psychometricians—working together—to flesh out learning progression hypotheses. As stated in the online introduction to the CCSS-M[4].

> What students can learn at any particular grade level depends upon what they have learned before. Ideally then, each standard in this document might have been phrased in the form, "Students who already know A should next come to learn B." But at present this approach is unrealistic—not least because existing education research cannot specify all such learning pathways. Of necessity therefore, grade placements for specific topics have been made on the basis of state and international comparisons and the collective experience and collective professional judgment of educators, researchers and

[4] http://www.corestandards.org/Math/Content/introduction/how-to-read-the-grade-level-standards

> mathematicians. One promise of common state standards is that over time they will allow research on learning progressions to inform and improve the design of standards to a much greater extent than is possible today.

The last sentence of this paragraph is important because it makes clear that within-domain content standards ("clusters" of standards) in the CCSS-M are unlikely to serve as an adequate basis for a learning progression without further elaboration, and also that the domain conceptualizations in the Common Core are by no means sacrosanct as models for student learning. Finally, this sentence explicitly calls for more research on learning progressions. An encouraging development along these lines are the recent efforts by Jere Confrey and colleagues at North Carolina State University to "unpack" the CCSS-M in terms of multiple learning progressions—18 in all (Confrey, Nguyen, Lee, Panorkou, Corley & Maloney, 2012; Confrey, Nguyen, & Maloney, 2011). Building on Confrey's work, we provide an example of a learning progression[5] for proportional reasoning that could be used to conceptualize growth along a vertical scale.

Proportional reasoning involves reasoning about two quantities, x and y, that are multiplicatively related. This relationship can be expressed formally as a linear equation in the form $y = mx$, or, two value pairs can be expressed as equivalent ratios in the form $\frac{y_1}{x_1} = \frac{y_2}{x_2}$. For example, the following questions involves proportional reasoning: (A) if 3 pizzas can feed 18 people, how many pizzas would you need to feed 30 people? (B) At one table, there are 3 pizzas for eight people. At another table, there are 7 pizzas for 12 people. At each table, the people

[5] What we show here is a snapshot view of the full learning progression which is too large to fit on a single page, but is much easier to convey on a website. A link to the full learning progression will be included once this manuscript is no longer under peer review.

share the pizza equally. Which table would you rather sit at, if you want to get the most pizza? Question (A) involves finding a missing value in a proportional situation, and question (B) involves comparing two ratios.

The first five levels of this progression[6] are based upon a detailed learning progression for equipartitioning developed by Confrey and colleagues (Confrey, Maloney, Nguyen, Mojica, & Myers, 2009; Confrey 2012), while levels 6 and 7 come from a progression developed by Peck and Matassa to extend the equipartitioning progression into Algebra I (Matassa & Peck 2012; Migozuchi, Peck, & Matassa, 2013). This progression, like all learning progressions, is grounded in studies of student learning. To develop the equipartitioning progression, Confrey et al. first engaged in a comprehensive synthesis of the literature related to student learning of rational numbers. From this, they developed a number of "researcher conjectured" learning progressions for different aspects of rational number and multiplicative reasoning. One of these aspects was equipartitioning, which Confrey et al. (2009, p. 347) describe as "behaviors to create equal-sized groups" in sharing situations. For example, students use equipartitioning to find the fair share when 7 pizzas are shared by 12 people.

To refine the progression for equipartitioning, they conducted 52 clinical interviews with students in grades K-6. Peck and Matassa's work to extend this progression into middle and high-school followed a similar path of creating a researcher-conjectured progression based on the research literature and testing and refining it through work with students (Peck and Matassa conducted classroom design studies rather than clinical interviews for this step). Because the progression is grounded in studies of student learning, it is not simply an abstract construction developed by researchers, but rather an empirically supported description of learning over time.

[6] In the mathematics education literature, the term learning trajectory is typically used in place of learning progression, and the work of Confrey and colleagues also invokes the trajectory terminology. However, for the sake of consistency, we use the term progression throughout.

The concepts that are developed in this learning progression are foundational for school mathematics. The progression begins with equipartitioning, which Confrey and colleagues (Confrey & Smith, 1995; Confrey et al., 2009) have argued ought to be considered a "primitive" (along with counting) for the development of fractions, multiplicative reasoning, and proportional reasoning. Thus, the levels in the equipartitioning portion of the learning progression (Levels 1-5) set the stage for many of the standards that students are expected to learn in elementary school (e.g., fair sharing as a basis for division and fractions, and reversing the process—i.e., re-assembling shares into a whole—as a basis for multiplication). Moreover, mastery of equipartitioning sets the stage for proportional reasoning. This is important because just as equipartitioning provides a fertile environment for so much subsequent mathematics, so too does proportional reasoning (Post, Behr, & Lesh, 1988). In fact, the National Council of Teachers of Mathematics identifies proportional reasoning as one of five "foundational ideas" (NCTM, 2000, p. 11) in mathematics (rate of change—which is also developed in the progression—is another foundational idea). Thus the progression represents what is arguably the most important thread in elementary and middle school mathematics.

Insert Figures 4 and 5 about here

Figures 4 and 5 present an overview of the seven distinct levels of the proportional reasoning learning progression. The figures are two sides of the same coin in that Figure 4 describes, for each level, the attributes students are mastering in order to demonstrate increasing sophistication in their proportional reasoning, while Figure 5 describes the essence of the instructional and assessment activities that can be used both to develop, and gather evidence of, mastery. The lowest level of the learning progression is premised on a student that has just begun to receive formal instruction in mathematics (perhaps in Kindergarten, perhaps in first grade) and is being asked to complete activities that require the first building blocks in the development of proportional reasoning—sharing collections of objects with a fixed number of people. The highest level of the learning progression represents the targeted knowledge and skills in mathematics that would be expected of a student at the end of grade 8. At this level when faced with problems that involve making predictions from linear relationships, students are able to apply modified proportional reasoning to solve for unknowns, calculate unit rates (the rate at which one quantity changes with respect to a unit change in a different quantity, e.g., "miles per hour"), and interpret the algebraic construct of "slope" flexibly as both a rate of change and steepness. The levels in-between represent intermediate landmarks for students and teachers to aim for as they move along from the elementary school grades to the middle school grades.

Note that in this learning progression, at least as it has been initially hypothesized, there is not a one-to-one relationship between the number of distinct levels of the progression and the number of grades through which a student will advance over time. It may be the case that as we gather empirical evidence about student learning along this progression that we discover additional levels, or collapse existing ones. Rather than assigning a single grade with a single level, we might instead associate grade bands with each level, recognizing that grade designations are largely arbitrary, and that a student's sophistication in proportional reasoning is likely to depend upon the quality of focused instruction they have received on this concept rather than the age they happen to be. Notice also that the levels of the learning progression are not always defined by standards pulled from a single grade of the CCSS-M. In fact, standards from grade 4 of the CCSS-M do not fit within this particular progression at all because the grade 4 standards for fractions and rational number are focused on *fraction-as-number*. This sub-construct is the focus of a separate (but related) learning progression, based on the synthesis discussed above (Confrey, 2012).

It is the key activities that have been linked to each level of the progression in Figure 5 that ground proportional reasoning within the curriculum and teaching that are expected to take place behind classroom doors. These activities also serve as a basis for the design of assessment tasks or items that could be used in support of both formative and summative purposes. This is facilitated by the construction of item design templates for each level of the progression. These item design templates are similar in nature to the design pattern templates associated with Evidence-Centered Design. However, one feature of the templates we develop that makes them unique for the context of designing a vertical scale is the specification of item design factors that could be purposefully manipulated to make any given item harder or easier to solve. To illustrate this, and more generally the way that an item design template is linked to the learning progression, we describe the attributes of level 5 in more depth, using the exemplar task given in at the bottom of Figure 6 to ground the discussion.

Insert Figure 6 about here

For attribute 1, students can name a fair share in multiple ways and can explain why the different names represent equivalent quantities. In general, this means that students can use different referent units when naming a share and can coordinate the numerical value with the referent unit. In the exemplar task, this would result in share names of “1/10 of the four pounds”, "4/10 of one pound", or "4/10 pounds per chicken". For attribute 2, students use and justify multiple strategies when sharing multiple wholes to multiple sharers. In the exemplar task, students might use a "partition-all" strategy or an "equivalent ratio" strategy (Lamon, 2012). In the partition all strategy, students would partition each pound into tenths, and then distribute one-tenth from each pound to each chicken. In the equivalent ratio strategy, students would reason that ten chickens sharing four pounds of food results in the same shares as if five chickens shared two pounds of food, and then share the food according to this reduced ratio. For attribute 3, students assert, use, and justify the general principle that whenever *p* items are shared by *n* sharers, the fair shares will have size of *p*/*n* items per sharer (or equivalent names as discussed above). In the exemplar task, students would write a correct name for the fair share and would justify this share by using a strategy as described above.

The *task family* implicit in Figure 6 is designed to help students master these attributes, and also to help test developers and teachers assess student mastery of these attributes. The task can be varied by changing the number and type of items to be shared as well as the number and type of sharer. By varying these task features, test developers and teachers can (a) create novel learning and assessment experiences, (b) vary the difficulty of the task, and (c) create conditions that are conducive to particular teaching strategies. Perhaps most obviously, for the level 5 task family the number of objects (*p*) to be shared and the number of people with whom the objects are to be shared (*n*) can be changed (e.g., chocolate bars and people, or chicken food and chickens). This does more than change the surface appearance of the task, it can also adjust the "distance" between the real-world activity and the mathematical activity. For example, in the chocolate bars and people situation, the real-world activity of breaking chocolate bars and passing out pieces is closely related to the mathematical activity of partitioning and distributing.

This is probably less true for the chicken food and chickens situation. In this way the task can become more or less abstract as the items and sharers are varied. The difficulty of the task can also be varied by changing *p* and *n* according to the schedule given in row three of Figure 6 (this progression of difficulty comes from Confrey, 2012). In classroom settings, teachers could modify *p* and *n* to create conditions that are conducive to particular strategies. For example, situations where *p* and *n* have a common divisor are more conducive to the equivalent ratio strategy than are situations where *p* and *n* are relatively prime.

A fully elaborated item design template would also include scoring rules for constructed response items and examples of student responses that would earn different scores. As evidence is gathered about the ways that students tend to respond to such items, the template could be extended to include rules or guidelines for writing selected response items. From the standpoint of extracting diagnostic information from such items, a particularly compelling feature of such items might be to give students partial credit for responses that demonstrate mastery of some, but not all, of the attributes associated with the level to which an item has been written.

Common Item Linking Designs

A challenge in designing a vertical scale is collecting data on how students at one grade level would fare when presented with items written for students at a higher or lower grade level. There is understandably some concern about overwhelming younger students with items that are much too hard, or boring older students with items that are much too easy. Adopting a learning progression as the basis for a common item linking design has the potential to lessen this concern for three reasons. First, because explicit connections are being made between the mathematical content and practices to which students are exposed from the lower (e.g. elementary school) to upper (e.g., middle school) anchors of the learning progression, it would no longer be the case that, for example, the activities at upper levels of a learning progression would be completely foreign to students at the lower levels. For example, activities at level 6 of the proportional reasoning learning progression (see Figure 5) could still involve asking students to devise fair shares using equipartitioning strategies, a common feature of activities from levels 1 through 5. Second, because the items designed for each level of the progression could be manipulated to be easier or harder, one would naturally expect to see a great deal of overlap in the ability of students to solve these different item families correctly across grade bands. For example, a very hard level 5 item might be just as challenging as a very easy level 6 item. This blurring of artificial grade level boundaries makes it possible to envision field test designs in which students in adjacent grades could be given items that span three or more hypothesized learning progression levels, because a level would not necessarily be equivalent to a grade. For example, while it would surely be unreasonable to ask first grade students to answer level 6 or 7 items, it might be entirely reasonable to pose some of these items to students in 3rd or 4th grade, just as it might be reasonable to pose level 3-5 items to students 7th or 8th grade. Third, as noted previously, there is no requirement that a vertical scale associated with any given learning progression design would need to span any set number of grades. For example, instead of building a vertical scale to represent growth in proportional reasoning across grades 3 through 8, a decision could be made to create a vertical scale that only spans grades 6 through 8. Indeed, an entirely different learning progression hypothesis might be the basis for a another vertical scale that spans grades 3 to 5, or 4 to 6, etc.

Discussion

To recap, the concept of growth is at the foundation of the policy and practice around systems of educational accountability. It is also at the foundation of what teachers concern themselves with on a daily basis as they help their students learn. Yet there is a disconnect between the criterion-referenced intuitions that parents and teachers have for what it means for students to demonstrate growth, and the primarily norm-referenced metrics that are used to communicate inferences about growth. One way to address this disconnect would be to develop vertically linked score scales that could be used to support both criterion-referenced and norm-referenced interpretations, but this hinges upon having a coherent conceptualization of what it is that is growing from grade to grade. In this paper we have proposed a learning progression approach to the conceptualization of growth and the subsequent design of a vertical score scale. We have used the context of the CCSS-M and the "big idea" of proportional reasoning to give a concrete illustration for what such a design approach would entail.

In their book *Test Equating, Scaling & Linking*, Kolen & Brennan (2004) also distinguish between two different ways that growth could be conceptualized when designing a vertical scale. They introduce what they call the "domain" and "grade to grade" definitions of growth. In what they refer to as a "domain definition" of growth, the term domain is used much more broadly than we have used it here to encompass the entire range of test content covered by the test battery across grades. In other words, the domain of a sequence of grade-specific tests of mathematics as envisioned by Kolen & Brennan would include all the shapes we defined as unique content domains in Figure 3. In contrast, Kolen & Brennan define grade to grade growth with respect to content that is specific to one grade level but which has also been administered to students at an adjacent grade level (i.e., all the shapes in Figure 3 that overlap grades). The learning progression definition of growth we have illustrated has some similarity to Kolen & Brennan's domain definition in the sense that a learning progression design focuses upon growth with respect to a common definition of focal content across grades. However, the learning progression approach departs from Kolen & Brennan's domain definition in the emphasis on (a) one concept (or collection of related concepts) at a time, and (b) how students become more sophisticated in their understanding and application of this concept as they are exposed to instruction.

A learning progression approach to design has the potential to address two of the concerns that can threaten the validity of growth inferences on existing vertical scales. The first concern is the empirical finding that growth decelerates as students enter middle school grades to the point that it appears that some students have not learned anything from one grade to the next (Dadey & Briggs, 2012; Briggs & Dadey, 2015). Although such a finding could still persist even when a vertical scale has been designed on the basis of a learning progression hypothesis, it would be easier to rule out construct shift as a plausible cause of score deceleration. If it were to be found, for example, that students grew twice as fast in their proportional reasoning from grades 5 to 6 relative to grades 6 to 7, this could be raise important questions about the coherence of curriculum and instruction in grade 7 relative to grade 6. The second concern is that gains along a vertical scale cannot be shown to have interval properties. Although there is nothing about taking a learning progression approach that guarantees a resulting scale with interval properties, there are in fact novel empirical methods that could be used to evaluate this proposition (Briggs, 2013; Domingue, 2013; Karabatsos, 2001; Kyngdon, 2011). One of the key design features that could make test data more likely to approximate the canonical example of an attribute with ratio scale properties (length) or interval scale properties (temperature), is the presence of external factors that can be used to predict the empirical difficulty of any given item, or the probability of any given person answering an item correctly. Because such factors are made explicit in the development of a learning progression hypothesis, this represents a step in the right direction. At a minimum, tests designed according to a learning progression would seem more likely to fit the Rasch family of IRT models, and thereby inherit some of the desirable invariance properties of such models (Andrich, 1988; Wright, 1997).

Another key advantage of the learning progression approach is that it can serve as a bridge between summative and formative uses of assessments. Although there is a great deal of rhetoric around the need for teachers to make "data-driven" instructional decisions, there is little reason to believe that teachers are able to extract diagnostic information from the student scores reported on a large-scale assessment, even when score are disaggregated into content-specific subscores. With respect to inferences about growth in particular, finding out that in a normative sense one's students are not growing fast enough relative to comparable peers tells a teacher nothing about what they need to be changing about their instruction. In contrast, if a normative SGP attached to each student could be accompanied by information about the change and current location of the students along a vertical scale for proportional reasoning, this would greatly expand the diagnostic utility of the results. Not only would parents and teachers have a sense for how much a student has grown, but by referencing the canonical items and tasks associated with a student's current location, they would have actionable insights about what could be done next. Further, by making item design templates associated with the learning progression publicly available, it becomes possible for teachers to create and score their own tasks to assess and monitor student progress at multiple junctures over the course of a school year.

Our focus here on the potential benefits of thoughtfully designed vertical scales is not intended as a rebuke of the normative inferences fundamental to value-added models or the Colorado Growth Model. Instead, it is a recognition that neither purely normative nor purely criterion-referenced growth interpretation are sufficient to answer all the questions parents, teachers and students have about learning in educational settings. Economists and applied statisticians have made great innovations in the development and research into models that can flag teachers and schools that appear to be excelling or struggling on the basis of normative comparisons. Similar innovations in the development and research on vertical scaling have lagged in the psychometric community. If fundamental questions about how student growth should be conceptualized and measured are not being taken up among psychometricians, they are likely to remain unanswered altogether.

Taking a learning progression approach to design one or more vertical scales within a subject area (i.e., math, English Language Arts) is not incompatible with the need to also assess the breadth of student understanding along the full range of the CCSS. Just as the salient distinction between status and growth has become clear since the advent of NCLB in 2002, so to is it possible to distinguish between the use of a large-scale assessment to produce different scale scores for different purposes. If the sole purpose is to take a grade-specific inventory of the different knowledge and skills that students are able to demonstrate from the different domains that define math and ELA, then domain sampling is an entirely appropriate method for building a test blueprint. However, if an additional purpose is to support coherent and actionable inferences of growth, this can be accomplished at the same time by adopting a stratified domain sampling approach, where one or more strata might consist of the domain within which a learning progression has been specified. Naturally it would be convenient to have a single scale that could fulfill both purposes, and this has been the impetus for conventional approaches to vertical scale design. But what does it really mean to say that a student has grown X points in math or Y points in ELA? This merely begs the next question: growth in what aspect of math or ELA? In our view the latter is a question that has a much greater chance of being answered coherently when a vertical scale is based on a learning progression hypothesis.

Challenges and Opportunities

The use of the learning progression approach within the context of large-scale assessment design and analysis comes with significant psychometric challenges. To begin with, the initial development of a learning progression hypothesis can be time-consuming process, not always amenable to the tight deadlines facing large-scale assessment programs. Fortunately, there is a considerable literature on learning progression in math education, so much of this initial work has already been started. A thornier issue is coming up with items that are rich enough to elicit information about the sophistication of student understanding without always requiring lengthy performance tasks with open-ended scoring. The problem with such tasks is that while they may be ideal as a means of eliciting the information needed to place a student at a specific location along the vertical scale, the context of the task may contribute so much measurement error that it is very hard to feel much confidence in a student's location. And if a student's location at one point in time cannot be established reliably, the reliability of gain scores across two points, or score trajectories across more than two points in time are likely to suffer even more. A possible solution to this is to attempt to break larger performance tasks into smaller set of selected response and constructed response items. This is essentially the compromise approach presently being taken for the math assessments that have been designed by PARCC and SBAC. The item template we illustrated for level 5 of our proportional reasoning learning progression also hints at this strategy, since the target item prompts could be expressed as short constructed response items, selected response items, or some combination of the two. Where this challenge is likely to be hardest to overcome would be for a learning progression that focused on increasing sophistication of a written argument.

Another significant challenge to the learning progression approach comes in heterogeneity of curricular sequences to which students are exposed across states, within the same state and even within the same school district. For example, given one state that repeatedly emphasizes the concepts underlying proportional reasoning in its K-8 curriculum relative to another state that does not, one might expect to find differential item functioning on linking items as a function of each state's enacted curriculum. Of course, this is a potential problem for the assessments being developed by PARCC and SBAC even without taking a learning progression approach.

At the same time, there is a risk that a learning progression approach to assessment will narrow and homogenize learning opportunities, and can lead to simplistic interpretations of complex processes (Sikorski, Hammer & Park, 2010). At worst, this might limit opportunities for students to bring their own heterogeneous backgrounds and ways of knowing to bear on their learning, thus "re-inscrib[ing] normative expectations in learning that have homogenizing effects" (Anderson et al., 2012, p. 15). In part, this risk derives from a tension in the research on learning progressions that we alluded to earlier, namely, that learning is a complicated process with multiple pathways, even as some pathways are more likely than others. While our focus on this paper is on learning progressions, we note in passing that some researchers, for example those in the *Dynamic Learning Maps consortium*, are exploring how psychometric techniques can be incorporated into progressions with multiple pathways[7]. The risk of homogenization is compounded to the extent that researchers who develop learning progressions do not attend to heterogeneity in students' ways of knowing, or simply account for this diversity in the "lower anchor" of a progression (Anderson et al., 2012). One response, then, is that it is the responsibility of the researchers who *create* the learning progressions to attend to heterogeneity, and to create progressions at large enough grain sizes so as to allow for diverse learning opportunities. From this perspective, learning progressions are simply the *a priori* background which inform assessments and vertical scales. However, we reject this unidirectional model, and instead suggest that assessments and learning progressions can—and should—be mutually informing.

A learning progression constitutes a hypothesis about growth, and as longitudinal evidence is collected over time, the hypotheses can be proven wrong, and at a minimum it is likely to evolve. This fact represents a challenge to conventional psychometric practices, but also an opportunity. It is an opportunity for psychometricians to partner with content specialists, cognitive and learning scientists and teachers to gain insights about not just what students know and can do, but what and how much they can learn. For more than a decade now every state has been testing its students across multiple grades in math and reading, but all this testing has generated very little insight about student learning and how it can best be facilitated. Vertical scales could provide these kinds of insights if a case can be made the growth indicated by test scores is a measure of learning. Making this case coherently could be the next frontier in educational assessment.

[7] We thank an anonymous reviewer for bringing this to our attention

References

Anderson, C. W., Cobb, P., Barton, A. C., Confrey, J., Penuel, W. R., & Schauble, L. (2012). *Learning progressions footprint conference: Final report*. East Lansing, MI: Michigan State University

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills, CA: SAGE Publications.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores (pp. 397-479). Reading, MA: Addison-Wesley.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204-226.

Briggs, D. C. & Dadey, N. (2015). Making sense of common test items that do not get easier over time: Implications for vertical scale designs. *Educational Assessment*, 20(1), 1-22.

Briggs, D. C. & Weeks, J. P. (2009) The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice,* 28(4), 3-14.

Castellano, K. E., & Ho, A. D. (2013a). *A Practitioner's Guide to Growth Models*. Washington, DC: Council of Chief State School Officers.

Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to Student "Growth" Percentiles. *Journal of Educational and Behavioral Statistics*, *38*(2), 190-214.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9): 2593-2632.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, *6*(2), 81–89. doi:10.1207/s15327833mtl0602

Confrey, J. (2012). Better measurement of higher cognitive processes through learning trajectories and diagnostic assessments in mathematics: The challenge in adolescence. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making* (pp. 155–182). Washington DC: American Psychological Association.

Confrey, J., Nguyen, K. H., Lee, K., Panorkou, N., Corley, A. K., and Maloney, A. P. (2012). Turn-On Common Core Math: Learning Trajectories for the Common Core State Standards for Mathematics. Retrieved from: www.turnonccmath.net.

Confrey, J., Nguyen, K. H., and Maloney, A. P. (2011). Hexagon map of Learning Trajectories for the K-8 Common Core Mathematics Standards. Retrieved from: http://www.turnonccmath.net/p=map.

Confrey, J., & Smith, E. (1995). Splitting, covariation, and their role in the development of exponential functions. *Journal for Research in Mathematics Education*, *26*(1), 66–86.

Confrey, J., Maloney, A., Nguyen, K. H., Mojica, G., & Myers, M. (2009). Equipartitioning/splitting as a foundation of rational number reasoning using learning trajectories. In M. Tzekaki, M. Kaldrimidou, & C. Sakonidis (Eds.), *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1). Thessaloniki, Greece: PME.

Corcoran, T., Mosher, F.A., & Rogat, A. (2009). Learning progressions in science: An evidence based approach to reform. NY: Center on Continuous Instructional Improvement, Teachers College—Columbia University.

Dadey, N. & Briggs, D. C. (2012). A meta-analysis of growth trends from vertically scaled assessments. *Practical Assessment, Research & Evaluation*, 17(14). Available online: http://pareonline.net/getvn.asp?v=17&n=14

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. CPRE Research Report #RR-68. Philadelphia: Consortium for Policy Research in Education. DOI: 10.12698/cpre.2011.rr68

Domingue, D. (2013). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika 79*(1), 1-19.

Holland, P. W. & Rubin, D. B. (1983). On Lord's Paradox. *Principals of modern psychological measurement.* H. Wainer, & S. Messick, Eds., Hillsdale, NJ: Lawrence Erlbaum Associates.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research. doi: 10.3386/w14607

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2(4), 389-423.

Kolen, M. J. (2006). Scaling and norming. In R. Brennan, (ed.) *Educational Measurement* (4th ed, pp. 155-186). Westport, CT: American Council on Education/Praeger

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Verlag.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, 64(3), 478-497.

Lamon, S. J. (2012). *Teaching fractions and ratios for understanding: Essential content knowledge and instructional strategies for teachers* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Lord, F. M. (1967) A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304–5.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability. Addison-Wesley, Reading, MA*.

Martineau, J. A. (2004). The effects of construct shift on growth and accountability models. Unpublished Dissertation. Michigan State University.

Martineau, J. A. (2005). *Un-distorting measures of growth: Alternatives to traditional vertical scales*. Paper presented at the Annual Conference of the Council of Chief State School Officers.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62.

Matassa, M. & Peck, F. (2012). Rise over run or rate of change? Exploring and expanding student understanding of slope in Algebra I. *Proceedings of the 12th International Congress on Mathematics Education*. Seoul, Korea. 7440-7445. Retrieved from: http://www.icme12.org/upload/UpFile2/WSG/0719.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. RAND Education. (Vol. 158). Research Report prepared for the Carnegie Corporation. Santa Monica, CA: RAND Corporation

MET Project. (2013). Ensuring fair and reliable measures of effective teaching. Policy and Practitioner Brief. Downloaded January 30, 2013 from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Migozuchi, T., Peck, F., & Matassa, M. (2013). Developing robust understandings of slope. *Elementary mathematics teaching today* (Journal published in Japan), 2013 no. 511. 31-32.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1, 3-67.

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

NCTM. (2000). *Principles and Standards for School Mathematics*. Reston, VA: NCTM.

Post, T. R., Behr, M. J., & Lesh, R. (1988). Proportionality and the development of pre-algebra understanding. In J. Hiebert & M. J. Behr (Eds.), *Number Concepts and Operations in the Middle Grades* (pp. 93–118). Reston, VA: National Council of Teachers of Mathematics.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York: Routledge.

Shea, N. A., & Duncan, R. G. (2013). From theory to data: The process of refining learning progressions. *Journal of the Learning Sciences*, *22*(1), 7–32.

Sikorski, T., Hammer, D., & Park, C. (2010). A critique of how learning progressions research conceptualizes sophistication and progress. In *Proceedings of the 9th International Conference of the Learning Sciences Vol 1* (pp. 1032–1039). Chicago, IL: International Society of the Learning Sciences.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56(4), 495-529.

Stocking, M. L. and Lord, F. M. (1983) Developing a common metric in item response theory. Applied Psychological Measurement, 7(2), 201-210.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7), 433-451.

Thurstone, L. L. (1927). The unit of measurement in educational scales. The Journal of Educational Psychology, 18, 505-524.

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.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, *Winter 199*, 33–45.

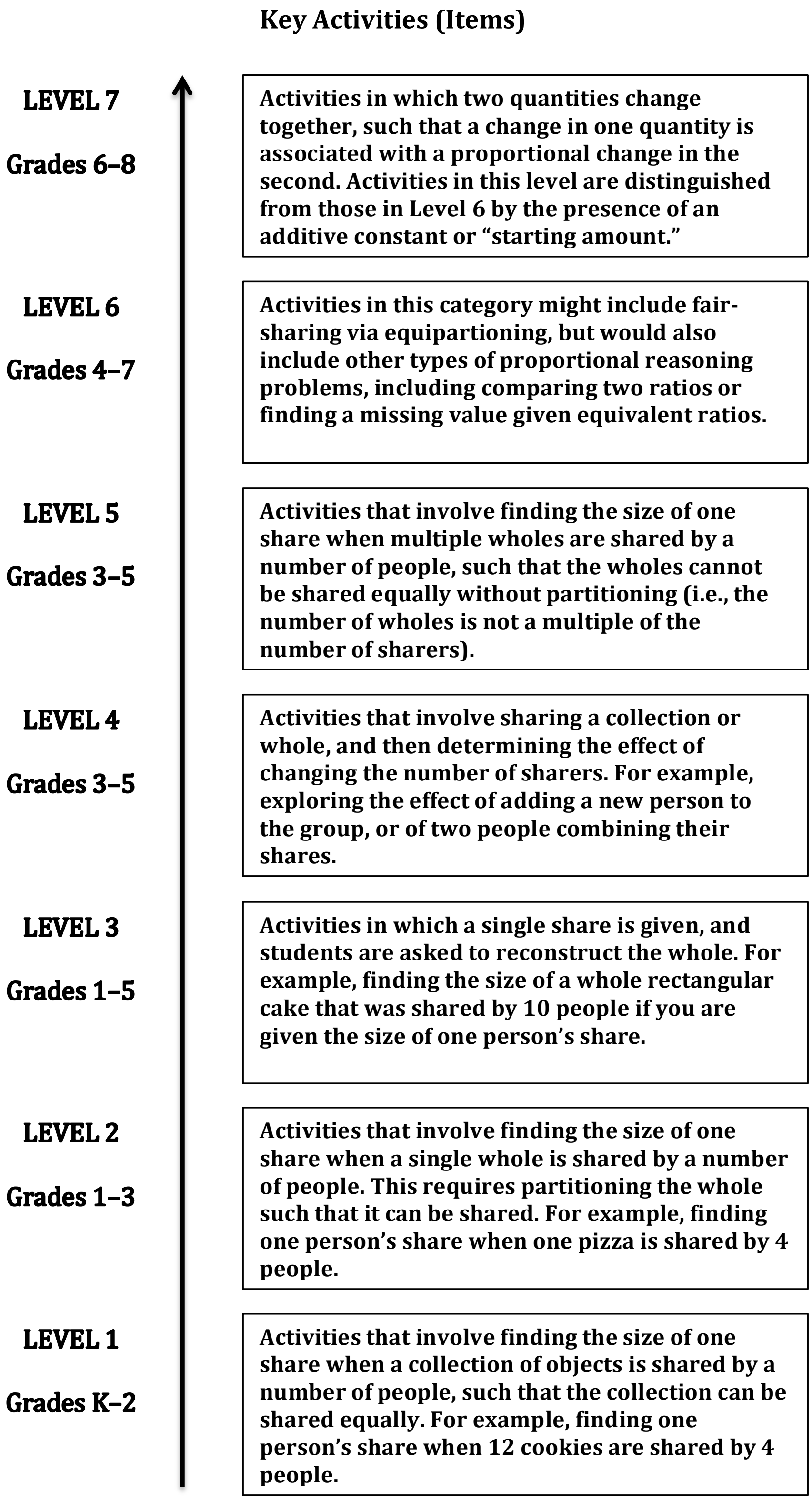
Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299–325.

Young, M. J. (2006). Vertical Scales. In S. Downing & T. Haladyna (eds). *Handbook of Test Development*, 469-485. Mahwah, NJ: Lawrence Erlbaum Associates.

Figure 4. Learning Progression for Proportional Reasoning: Student Attributes

	Attributes: What Students Know and Can Do	**CCSS-Math**
LEVEL 7 **Grades 6–8**	**1. Students can make predictions for linear relationships as *y* = *mx* + *b*.** **2. Students can calculate the unit rate given any two values.** **3. Students interpret slope as rate of change.**	**8.EE.B.5** **8.F.A.3** **8.F.B.4** **7.RP.A.2** **MP.2**
LEVEL 6 **Grades 4–7**	**1: Students can build one ratio from another.** **2: Students use multiplication or division to arrive at target ratio in single step.** **3: Students use ratio as multiplicative comparison and can name unit rates.**	**6.RP.A.1** **6.RP.A.2** **6.RP.A.3** **MP.2**
LEVEL 5 **Grades 3–5**	**1: Students can name fair shares multiple ways and explain their equivalence.** **2: Students can explain why different methods to create fair shares are equivalent.** **3: Use and justify principle: p by $n = {}^{p}/_{n}$.**	**5.NF.3** **3.NF.3c** **MP.2**
LEVEL 4 **Grades 3–5**	**1: Students recognize qualitative compensation as an inverse relationship.** **2: Distinctions made between additive and multiplicative relationships.** **3: Students can use transitivity in explanation.**	**3.NF.3b** **MP.2**
LEVEL 3 **Grades 1–5**	**1: Students master concept of "*n* times as much." If collection or whole is shared by n people, whole is *n* times single share.** **2: Can justify equivalence by reconstructing whole from a given part.**	**5.NF.5** **3.0A.1** **MP.2**
LEVEL 2 **Grades 1–3**	**1: In naming shares as counts of "one piece," students consider <u>number</u> and <u>size</u> of pieces.** **2: Students use geometric or measurement ideas to justify equivalence of shares.**	**3.OA.2** **3.G.2** **3.NF.1** **2.G.2, 2.G.3** **MP.2**
LEVEL 1 **Grade K–2**	**1: Students can name shares from a collection of objects numerically using extensive or intensive units.** **2: Students can justify the equivalence of shares by counting.**	**2.G.2** **1.G.3** **MP.2**

Figure 5. Learning Progression for Proportional Reasoning: Key Activities



Key Activities (Items)

LEVEL 7

Grades 6–8

Activities in which two quantities change together, such that a change in one quantity is associated with a proportional change in the second. Activities in this level are distinguished from those in Level 6 by the presence of an additive constant or "starting amount."

LEVEL 6

Grades 4–7

Activities in this category might include fair-sharing via equipartioning, but would also include other types of proportional reasoning problems, including comparing two ratios or finding a missing value given equivalent ratios.

LEVEL 5

Grades 3–5

Activities that involve finding the size of one share when multiple wholes are shared by a number of people, such that the wholes cannot be shared equally without partitioning (i.e., the number of wholes is not a multiple of the number of sharers).

LEVEL 4

Grades 3–5

Activities that involve sharing a collection or whole, and then determining the effect of changing the number of sharers. For example, exploring the effect of adding a new person to the group, or of two people combining their shares.

LEVEL 3

Grades 1–5

Activities in which a single share is given, and students are asked to reconstruct the whole. For example, finding the size of a whole rectangular cake that was shared by 10 people if you are given the size of one person's share.

LEVEL 2

Grades 1–3

Activities that involve finding the size of one share when a single whole is shared by a number of people. This requires partitioning the whole such that it can be shared. For example, finding one person's share when one pizza is shared by 4 people.

LEVEL 1

Grades K–2

Activities that involve finding the size of one share when a collection of objects is shared by a number of people, such that the collection can be shared equally. For example, finding one person's share when 12 cookies are shared by 4 people.

Figure 6. Item Design Template for Level 5 of Proportional Reasoning Progression

Title	Multiple people sharing multiple wholes
Overview	This family of activities involves finding equal shares when there are multiple items to be shared among multiple "sharers" (e.g., people), and the number of sharers is not a multiple of the number of items (i.e., some or all of the individual items will have to be partitioned)
Factors that change the difficulty of the task	Sharing multiple wholes [p = items; n = sharers] • $p = n + 1; p = n - 1$ • p is odd & $n = 2^j$ • $p >> n$ or p is close to n • all p; all n
Task in general form	<n> <sharers> share <p> <items> equally. Either *Representation given* The <items> are shown below. Mark the <items> to show how the <sharers> could share the <items> and shade in one <sharer>'s equal share. Explain your reasoning. or *Representation not given:* Find one <sharer>'s equal share. Explain your reasoning. How many different ways can you use to describe each <sharer>'s share numerically? Write as many ways as you can think of.
Task in Exemplar Form	Ten chickens share four pounds of food. (a) Find one chicken's equal share. Explain your reasoning. (b) How many different ways names can you use to describe each chicken's share numerically? Write as many ways as you can think of.