Developing Scale Scores and Cut Scores for On-Demand Assessments of Individual Standards

Working Paper

April, 2018

Nathan Dadey[1], Shuqin Tao[2], and Leslie Keng[1]

[1]The National Center for the Improvement of Educational Assessment

[2]Curriculumn Associates

Suggested Citation:

Dadey, N., Tao, S., & Keng, L. (2018, April). *Developing Scale Scores and Cut Scores for On Demand Assessments of Individual Standards*. Paper to be presented at the annual meeting of the National Council of Measurement in Education: New York, NY.

Introduction

Now, more than ever, assessments are being asked to fulfill an ever broadening range of purposes. Often, test users want an overall scale score, fine grained information on specific standards, as well as information on growth. Clearly, no one assessment can be expected to provide all such information with the same level of precision, but a combination of assessments, carefully tailored, could. Using multiple assessments, however, poses a completely new challenge – integrating the results of multiple assessments into a coherent narrative about student learning. We believe there are multitude of ways of doing so. The goal of this work is to examine a particular set of assessments to see whether such a narrative can be told using one or more unidimensional reporting scales.

Specifically, this work examines two types of interim assessments – a "general" assessment that broadly covers the fourth grade common core state standards (CCSS) in mathematics and a set of 31 short "mini-assessments", each of which covers a single[1] fourth grade math CCSS standard or sub-standard (e.g., 4.NF.B.4.C). These assessments differ not only in terms of content, but also in terms of administration and use. In one simple use case, students take the general assessment, receive instruction across multiple weeks, and then are assessed using one or more of the mini-assessments. However, this is just one of a variety of possible use cases, the uses of assessments, as well as the timing of their administration, vary across classroom, schools and districts. Given that these assessments are part of an operational testing program that spans multiple states, variation in use and administration is substantial. However, this is just the type of challenge systems of assessment are now facing – a wide variety of potential use cases paired with a diverse pattern of administrations, but still requiring sound measurement.

The purpose of this paper is examine ways in which the results from the mini-assessments can be modeled using psychometric methods – with an emphasis on the creation of one or more unidimensional latent scales as well as associated cut points. Given the relatively novel nature of this work, we integrate considerations of various design decisions throughout the work. This work is guided by two research questions:

1. In what ways can the mini-assessments be scaled? Specifically, can and should the 31 mini-assessments be:
    a. Placed onto a single unidimensional latent scale?
    b. Divided up and placed onto four unidimensional latent scales, corresponding to four CCSS fourth grade mathematics domains?
2. How can provisional cut scores be set on the mini-assessment total score scales?

Addressing (1) entails investigating the dimensionality of the set of 31 mini-assessments. Typically, dimensionality is viewed as pertaining to a single assessment – here we extend the concept and methods to the set of mini-assessments through a concurrent calibration approach. To address (2), we

[1] Four of the mini-assessments break this rule, and cover a narrow range of standards instead of an individual standard.

draw on information provided by the general assessments to create provisional cut scores. In doing so, we are attempting to increase the agreement between the mini-assessments and general assessment. We do note, however, that these provisional cut scores are meant to be revisited by content experts and adjusted as needed. Our ultimate aim in addressing these two research questions is to provide an example that illustrates one way to tackle the thorny issues inherent in modeling results from this type of distributed system of assessment.

A System of Assessments: The Call and the State of the Art

The concept of a carefully tailored combination of assessments is best reflected in the body of literature focused on the concept of a "system of assessments". The idea for a system of assessments can be traced back at least to the seminal work of Pellegrino, Chudowsky and Glaser (2001), who outline a plan for "coordinated systems of multiple assessments that work together, along with curriculum and instruction, to promote learning" (p. 252, original emphasis). These systems of assessment are meant to operate at multiple levels, "from the classroom through the school, district, and state" (p. 256). Work detailing approaches to "balanced" (Gong, 2010), "comprehensive" (Perie, Marion, & Gong; 2009; Ryan, 2010), and "next generation" (Darling-Hammond & Pecheone; 2010; Herman, 2010) assessment systems followed. Also important is the work that examines a particular type of assessment system – the through course assessment model (Bennett, Kane & Bridgeman, 2011; Ho, 2011; Kolen, 2011; Sabatini, 2011; Valencia, Pearson & Wixon, 2011; Way, McClarty, Murphy, Keng & Fuhrken, 2011; Wise, 2011; Zwick & Mislevy, 2011).

The development and implementation of systems of assessment have been slow to start, but has been gaining traction recently. The through course model appears to have had some success, with the release of the Smarter Balanced interim comprehensive assessments and assessment blocks. Other summative assessment vendors are beginning to respond as well, by providing assessments alongside their summative offerings that are smaller in scope than the typical summative assessment and meant to be used within the academic year for purposes other than state accountability (e.g., interim assessments; cf., Perie, Marion, & Gong; 2009). These approaches do not fully meet the vision laid out by Pellegrino et al. (2001), but represent a significant improvement. Interestingly, whereas Pellegrino et al. (2001) mainly conceptualized the levels of the assessment in terms of educational units (e.g., classroom, school, district or state) the work of Smarter Balanced and others focuses on providing short assessments that can be used flexibly at multiple layers of the educational system based on user preference. In this way, that work attempts to equip users with a set of assessments so that they can then tailor the administration of a specific subset of assessments to match the theory of learning underpinning instruction. Ideally, this tailoring should allow a set of general, curriculum neutral assessments to become curriculum relevant – tied to the scope and sequence of instruction. Whether this can be done successfully will likely require sustained effort from assessment users as well as a continuous support from assessment developers.

Intrinsically tied to the relevance educators and administrators draw meaning from such assessments is how the results are reported. The Smarter Balanced interim comprehensive assessments[2] are reported on the Smarter Balanced summative scale. This practice appears to be indicative of the current trend – to report the results of interim assessments on the scale of the summative item bank, presumably by leveraging item parameters from the summative assessment (although Zwick & Mislevy's 2011 approach is a variation on this, in that they suggest creating a scale through the application of the latent regression model used by the National Assessment of Educational Progress).

Our approach departs from this trend, instead of attempting to place the mini-summative assessments on the scale of the general assessment, we investigate two ways in which scales can be developed for just the set of mini-assessments. It is worth noting, however, that even though current practice is to report interim results on the scale of the summative item bank, the results of each interim assessment are generally provided in isolation. That is, the information produced from these assessments is often left in separate silos – never integrated into a holistic picture of what students know and can do and made easily available for practical use. In our work, we touch on this issue, partially, by drawing on the results of the general assessment to set preliminary cut scores on the mini-assessments.

Methods

Measures

Mini-assessments. The mini-assessments are short assessments meant to provide school and district educators and administrators with information about student mastery on individual content standards, or a grouping of similar standards, throughout the academic year. For example, district administrators often assign sets of mini-assessments to provide aggregate information at specific points during the academic year, which they then use to drill down in order to find specific areas of weakness (e.g., specific standards, grades, teachers or schools) to which they can provide targeted support. Thus, the mini-assessments are intended to be aggregated and used at the school and district levels. Like the general assessments, the mini-assessments are computer administered, but unlike the general assessment, the mini-assessments are not adaptive and are not currently scaled using an item response theory (IRT) model. Both the assessments and mini-assessments provide instant score reporting, although the mini-assessments are not currently scaled. The mini-assessments and the general assessment do not have items in common. There are also no common items among the mini-assessments. Key design features of the mini-assessments include:

- **Configurable**. Administrators can choose to group multiple mini-assessments into an "assignment", which are then administered by educators within an administrator-defined window. Assignments often function as end-of-unit quizzes.

[2] The interim assessment blocks are not, and instead primarily reported in terms of three categories - above standard, near standard, and below standard.

- **Administered as Needed**. Administrators choose when, and how often, mini-assessments are given. There is a recommended a calendar with a suggested administration schedule that matches the scope and pacing of their provided curriculum, but users may deviate from this schedule.
- **Short**. The mini-assessments each contain 6 to 10 items, all of which are machine scorable. Each mini-assessment is made up of selected response and a variety of polytomous item types (e.g., ordered-list, cloze drop down).
- **Multiform**. Within each subject and grade-level, there are approximately 30 standards-based assessments, grouped into two forms (A and B), resulting in about 60 assessment forms per grade. There are no common items across mini- forms for the same standard(s), nor are there common items across any two mini-assessments.
- **Open.** Students, educators and administrators can see the items making up each form, as well as student responses to each item.

It is worth noting that each of the mini-assessments was developed in isolation from the others. That is, the items were written specifically for each mini-assessment and the classical test statistics used for quality control were computed using only the data from the items within the given mini-assessment. Therefore, the quality control process was not designed to ensure undimensionality at the domain or overall mathematics levels, which can be a byproduct of processes designed to increase reliability.

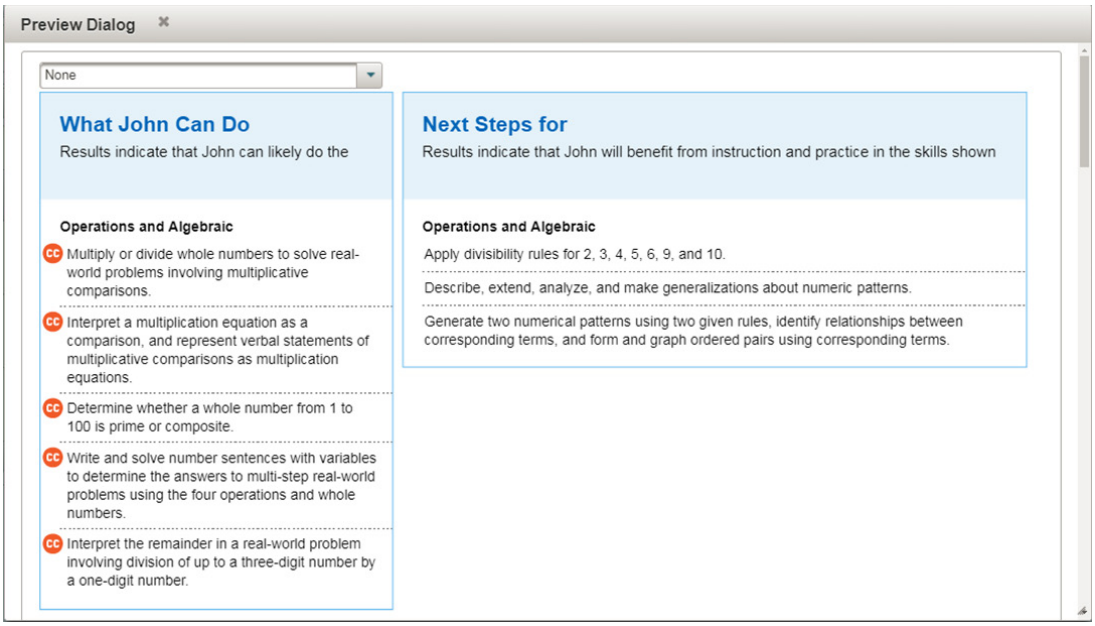
General assessment. The general assessment is a computer adaptive assessment meant to broadly assess fourth grade mathematics. Results from the general assessment are reported on a vertical scale that spans kindergarten to twelfth grade. The current instantiation of the vertical scale was created in 2015 using a concurrent calibration of approximately 9.5 million assessment administrations from operational data from a prior version of the assessment. The Rasch model was used with maximum likelihood estimation to produce parameter estimates. For fourth grade mathematics, the maximum number of possible items a student could be administered is 66 and the stopping rule is based on satisfying item minimums and maximums for each of four CCSS mathematics domains. Like the mini-assessments, the general assessment includes selected response items and a number of polytomous items, including short constructed response, text highlight, drag-and-drop, multiple-correct response, coordinate grid, and number line items. The recommended administration pattern of the general assessment is three times a year, once in fall, then again in winter and finally in spring. However, the assessment can be given as frequently as educators and administrators choose.

In addition to the vertical scale score, student reports on the general assessment also have number of other scores, including four CCSS domain subscores and a set of “indicator classifications”. The four domains upon which subscores are reported are: Operations and Algebraic Thinking, Number and Operations (which includes both Numbers and Operations in Base 10 and Numbers and Operations - Fractions), Geometry, and Measurement and Data. Each subscore is created by using the item parameter estimates from the overall vertical scale, but just for the items aligned to the given domain. A student’s domain subscore on Operations and Algebraic Thinking, for example, is based just on the

items that he or she took that are aligned to standards within the Operations and Algebraic Thinking domain.

The domain subscores also serve as the basis for set of reported scores – the indicator classifications. These dichotomous indicator classifications are meant to signal whether a student needs additional instruction on a given standard, sub-standard or grouping of standards or sub-standards. There are approximately 30 indicators for fourth grade mathematics and generally align to the same standards assessed by the mini-assessments. As explained in detail in Appendix A, the indicator classifications are defined for each student by, essentially, comparing his or her relevant domain score to the difficulty the items aligned to that indicator's standard(s). Since each student can receive a different set of items, indicators are only reported when a student receives six or more items within an indicator. Students identified as needing additional instruction are provided with content based recommendations for improving, which were generated by content experts through examination of each indicator's items. An example of the reporting of indicator classifications is provided below in Figure 1, in which each row corresponds to an indicator classification.

Figure 1. *Example Student Report of Indicator Classifications.*



Data

During the 2016-2017 academic year (August 2016 to July 2017) 101,966 students had a score on one or more of the mini-assessment forms. There are 31 mini-assessments in two forms (A and B), for a total of 62 assessment forms. The number of administrations per mini-assessment form ranges from approximately 3,000 to 47,000 with a mean of approximately 12,000 and a median of approximately 8,000. There is no mini-assessment form that all students take. The Form A mini-assessments of earlier standards (within a suggested instructional sequence) are taken by larger numbers of students than the other mini-assessments. The number of mini-assessment administrations, including re-tests, taken per

student range from 1 to 80, with a median of 6 and a mean of 7.6. There were 667 unique assignments (i.e., combinations of mini-assessments) administered 461,299 times. The number of mini-assessments within each assignment ranged from 1 to 9, with a median of 3. The number of times an assignment was administered ranged from 1 to 15,886 with a median of 93 and a mean of 692.

We are able to match general assessment data to 91,440 of the students taking mini-assessments. Our matching process uses the results from the general assessment that was administered closest to each mini-assessment in question. For example, if a student took the general assessment twice – once in the fall and again in the spring – we would use the fall administration for a mini-assessment taken in the fall. However, if the student took another mini-assessment in the spring, we would use the spring administration of the general for the analysis of that mini-assessment. Note, however, that we implemented this matching approach using the actual date of the administration, instead of the administration window.

Analytic Approach

Research question 1. Research question one asks whether the set of fourth grade mathematics mini-assessments can be placed onto a single unidimensional latent scale, or can be divided into CCSS domains and then scaled to produce four unidimensional latent scales[3]. We also entertained the idea of creating separate scales for each mini-assessment, but maintaining so many scales across multiple years seemed unfeasible. In addition, a finding that the mini-assessments could be scaled unidimensionally at the overall or domain level would render the need to investigate the individual assessment level moot.

The key criterion we use to examine whether or not any particular scaling is defensible is that of unidimensionality, as examined through a principle components analysis (PCA) of the standardized item residuals produced from a concurrent calibration of the mini-assessments. To conduct this concurrent calibration, we create a single person by item response matrix across all 62 mini-assessment forms and then apply the Rasch model (Rasch, 1960). This process produces a single scale spanning the 62 assessments, and we repeat this process within each domain to produce the domain-level scales (scaling each domain separately, ignoring the items from other assessments that are not within the domain). We conduct these calibrations in WINSTEPS (Linacre, 2016.) We then examine the standardized item residuals produced using these scales to determine whether the unidimensional scaling has adequately captured variation in student responses – that is, there are no remaining secondary dimensions present within the residuals. We also examine person and item fit using unweighted and weighted mean squared fit statistics.

To create the matrices for concurrent calibration, we pool across testing occasions. In instances were a student took a mini-assessment more than once, we use the item responses from the final adminstration. In creating this pooled item response matrix, we are eschewing traditional approaches to scale creation – we have neither common items nor common persons. In terms of the latter, although we have the same students, these students are generally not taking the mini-assessments at the same

[3] Another alternative we have not explored is to apply a multidimensional IRT model to the data.

points in time - posing questions around the applicability of a common person design. Our approach is to ignore differences in administration and scale the item responses across the 62 mini-assessment forms. This dataset represents a best-case scenario for detecting dimensionality - if we do not detect dimensionality in data where time is clearly a factor, we suspect we would not detect multidimensionality elsewhere (e.g., under an ideal common person design).

Research question 2. Research question two asks about the ways in which cut scores can be set on the mini-assessments. Specifically, the goal is to develop two cut scores for each mini-assessment form that classify students into three categories – Beginning, Progressing and Proficient. *A priori* qualitative descriptions of these categories are:

- **Beginning**. The student is not progressing well in the standard and would most likely benefit from review of concepts and skills that are prerequisite to understanding the concepts embodied in the standard.
- **Progressing**. The student is progressing towards mastery of the concepts embodied in the standard, but would most likely benefit from more practice on these concepts.
- **Proficient**. The student has mastered the concepts embodied in the standard – therefore he or she needs little or no additional instruction on the concepts within that standard.

With a single test or limited number of assessments, cut scores are generally set using judgmental procedures involving panels of experts (cf., Cizek & Bunch, 2007). However, with 62 mini-assessment forms, such a standard setting process is daunting. Moreover, the classifications from the mini-assessments are meant to signal whether students need additional instruction on a given standard, as are the indicator classifications from the general assessment. Thus, there is the potential for disagreement from the two different types of assessment – for example a mini-assessment could indicate that a student is does not need instruction on the standard(s), but a later administration of the general assessment could indicate that the student does need instruction.

For these reasons we use a method for setting cut scores that relies on the indicator classifications. Specifically, we use quantile regression to predict performance on the relevant general assessment indicator using the total scores from a mini-assessment, controlling for the difference in administration (in days). We then evaluate the resulting regression function to select a cut point for each assessment that is meant to differentiate between the Progressing and Proficient levels. Doing so entails making a number of decisions, decisions for which there is little empirical guidance. Below we list the steps we used to create the preliminary cut scores, as well as detailing the underlying motivation for the steps and decision points:

1. **Create the probability of mastering the corresponding indicator.**
   For each student, create the probability of mastering the corresponding indicator based on his or her domain score from the closest administration of the general assessment. This probability is computed as,

$$P(X_{ij} = 1) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \quad (1)$$

where $P(X_{ij} = 1)$ is the probability of student *j* mastering indicator *i*, θ_j is student *j*'s theta estimate for domain *i* from the closest general assessment administration and b_i is, essentially, an aggregate item difficulty that determines whether a student is classified as mastering or not mastering the given indicator. This difficulty value is derived through a multistep process, as described in Appendix A, and represents the domain theta value associated with obtaining 67% of the possible raw points on the items aligned to the indicator on the general assessment – adjusting θ_j for the aggregate difficulty of the items within the indicator.

This application of (1) treats each indicator as an item within the familiar Rasch function. This application also departs slightly from the way in which the indicator classifications are reported on the general assessment – that is, as dichotomous statements about whether students have mastered a particular standard and thus do, or do not, need instruction. Instead, we use a probability of mastery – to avoid losing information on student performance on the indicator. An alternative would be to use the indicator classifications directly and therefore predict the classifications via logistic regression.

2. **Conduct quantile regression.**
   Perform a quantile regression in which the probabilities of mastery produced in step 1 are predicted by the mini-assessment raw scores, controlling for the difference in administrations between the mini-assessments and the general assessments (in days)[4]. The quantile regressions are implemented using the quantreg R package (Koenker, 2017) in R version 3.4.0 and estimated for the 10th, 20th, … 90th quantiles.
3. **Evaluate the quantile regression.**
   Select a cut point by evaluating the quantile regression functions, with an emphasis on determining what mini-assessment raw score corresponds to $P(X_{ij} = 1)$ = 0.67, which is the value that is used to define the indicator classifications on the general assessment, and $P(X_{ij} = 1)$ = 0.50, which mirrors the approach in Rasch modeling to report item difficulties in terms of a response probability of 0.50. In addition, we also focus on the 50th quantile, which provides one indication of how the "typical" student performs. However, we do not treat these values as set in stone, nor do we *a priori* define the specific quantile to be evaluated.

These resulting cuts are meant to be provisional and subject to content expert review.

[4] There is a linear negative association between $P(X_{ij} = 1)$ and the difference between administrations between the assessment administrations (ranging from -0.29 to -0.04 across mini-assessment forms, with a mean of -0.18). The difference variable is defined as Date of Mini-Assessment minus the Date of the General Assessment, so the later the general assessment is administered after the mini-assessment, the higher $P(X_{ij} = 1)$ is (and vice versa). This association is not present between the difference variable and the mini-assessment scores. Likely, this pattern can be explained by the fact that scores on the general assessment domain subscales generally increase across the year, whereas the scores on the mini-assessment generally do not (barring the very beginning and end of the year).

Results

Research Question 1.

The results of the principle components analyses of the residuals from the Rasch model are summarized in Table 1 below and graphically in Appendix C. At most, the largest principle component accounted for about 2.0% of the unexplained variance – indicating that there are no sizable factors unaccounted for by the model. In addition, for the domain scaling, the percentages of items displaying misfit (values less than 0.75 and greater than 1.33) based on the unweighted mean squared fit statistics (i.e., infit) ranged from 1% to 6% across the domains. Similarly, the percent of items displaying misfit based on the unweighted mean squared fit statistics ranged from 11% to 22%.

Table 1. *Eigenvalues and Corresponding Percentages of Variance Accounted for From the Principle Components Analyses of Rasch Residuals (Domain Scaling)*.

Domain	Component				
	1	2	3	4	5
Operations & Algebraic Thinking	1.35 (1.1%)	1.31 (1.0%)	1.24 (1.0%)	1.21 (0.9%)	--
Numbers & Operations - Base Ten	1.51 (1.2%)	1.46 (1.2%)	1.34 (1.1%)	1.30 (1.0%)	1.27 (1.0%)
Numbers & Operations - Fractions	1.71 (0.9%)	1.54 (0.8%)	1.48 (0.8%)	1.44 (0.8%)	1.37 (0.7%)
Measurement & Data	1.59 (1.1%)	1.53 (1.1%)	1.47 (1.0%)	1.43 (1.0%)	1.36 (0.9%)
Geometry	1.42 (1.8%)	1.40 (1.8%)	1.32 (1.6%)	1.24 (1.6%)	1.17 (1.5%)

Table 2. *Summary of Mean Squared Fit Statistics (Domain Scaling).*

Domain	Fit Statistic	Median	Mean	Standard Deviation	% of Items <0.75	% of Items > 1.33	# Items
Operations & Algebraic Thinking	Infit	0.98	0.99	0.12	0%	1%	72
	Outfit	0.98	1.02	0.22	3%	8%	72
Numbers & Operations - Base Ten	Infit	0.98	0.99	0.12	0%	0%	72
	Outfit	0.98	1.02	0.25	7%	10%	72
Numbers & Operations - Fractions	Infit	0.97	0.99	0.12	0%	0%	108
	Outfit	1.00	1.01	0.20	3%	7%	108
Measurement & Data	Infit	1.00	1.00	0.13	0%	2%	84
	Outfit	1.00	1.02	0.32	12%	7%	84
Geometry	Infit	0.96	0.99	0.13	3%	3%	36
	Outfit	1.04	1.04	0.24	11%	11%	36

Research Question 2.

For most of the mini-assessments and quantile functions, the possible cut points were quite high. For example, Figure 2 provides multiple plots for the first mini-assessment, 1A, which is aligned to 4.NBT.A.1. The first plot is across all of the data available and shows that only the higher quantiles (60, 70, 80 and 90) intersect the line that corresponds to a Probability of Indicator Mastery of 67%. This result is partially an interaction between the changing student mastery probabilities across the year and

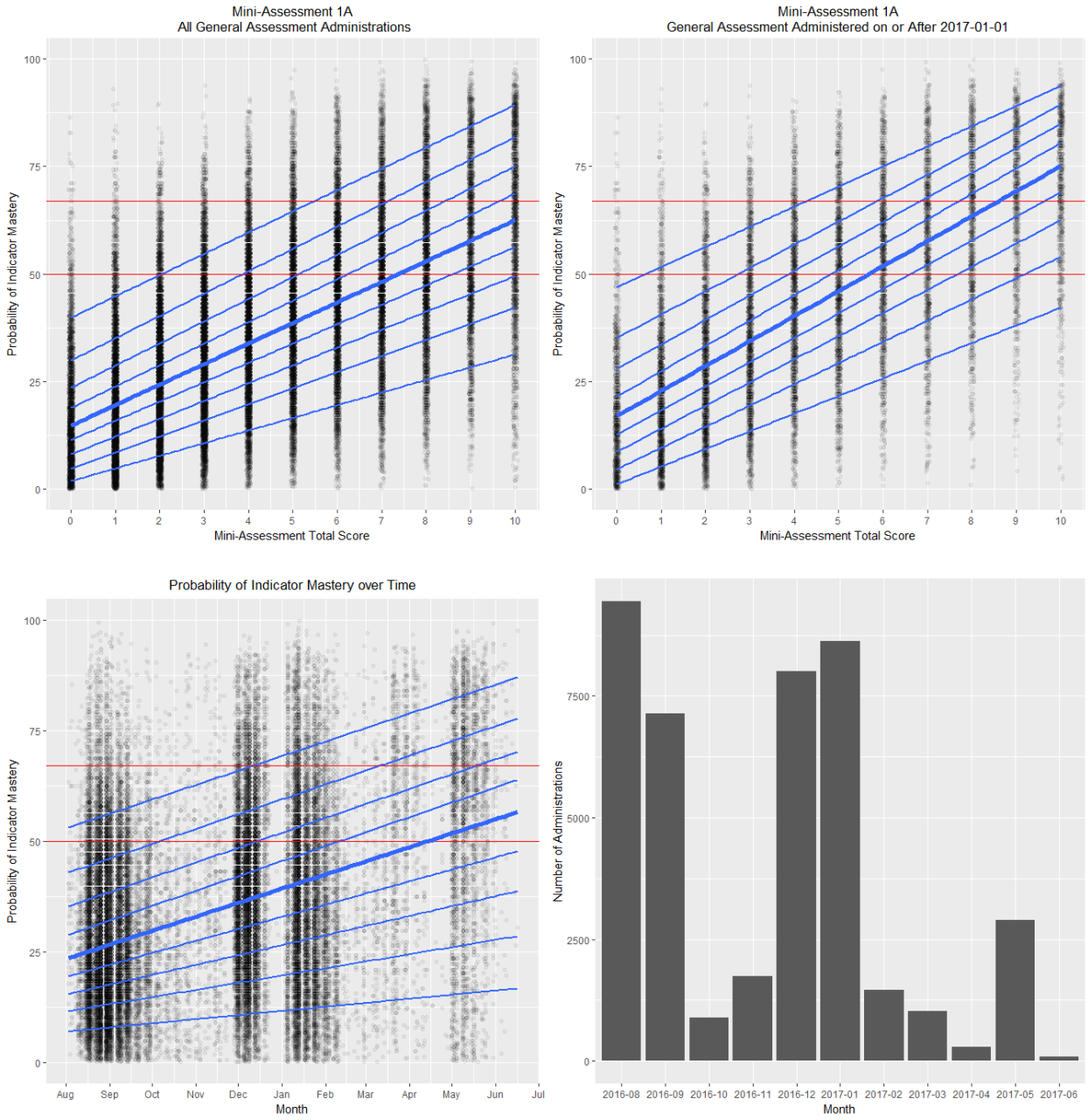
the patterns of administration of the mini-assessment[5]. That is, student mastery probabilities are generally lower at the beginning of the year than later on, and when the matching mini-assessments are generally administered towards the earlier parts of the year, the resulting regression relationships will show the total score that corresponds to a given Probability of Indicator Mastery is higher than if the mini-assessments are generally administered evenly across the year or towards the end of the year. The second plot in Figure 2 attempts to illustrate this point by using only data from the second half of the year. Whereas the 50th quantile regression line did not reach a Probability of Indicator Mastery of 67% when computed using all the data, it did when based on data from the latter half of the year.

Given these trends, the question of how to set the cut score becomes a question of what data should be used. One option is to simply use the data as is, producing the patterns mentioned above. A second option is to produce a different pattern of administrations through weighting, resampling or restricting the data window. For example, one might re-sample so that each of the recommended major administration windows (beginning of fall, middle of year and end of spring) are equally represented. Such an approach may be preferable, as the administration patterns of the mini-assessments do vary quite a bit from assessment to assessment, and thus doing so could insure some uniformity across the cuts. However, this approach is also not without issue – some assessments of standards that come later in the recommended instructional sequence (e.g., mini-assessments 32A and 32B) have almost no administrations within the first window.

We provide evaluations of the 50th quantile at $P(X_{ij} = 1)$ = 0.50 and 0.67, as well as 0.25, in Appendix D for both the overall sample as well as sample that attempts to balance the three administration windows (we do so by binning the month of administration into three categories and then sampling from within those categories). We note that this binning approach did not have a very large impact on the evaluations of the 50th quantile – likely because we were able to retain data from each binned range of administration dates. The results would have likely shifted more drastically if we had completely removed one or more of the binned ranges, as suggested in the second plot of Figure 2 (i.e., through the removal of data from the first half of the academic year).

[5] This finding is also due to the fact that the total scores on the mini-assessments, on average, remain stable across the year. So the mini-assessment scores are generally stable across the year while student mastery probabilities increase, meaning that the distribution of the matched administrations across the year impacts the cut. In addition, some general assessment indicators are much more difficult than others.

Figure 2. *Plots of General Assessment Indicator for 4.NBT.A.1 and Corresponding Mini-Assessment, 1A (the total number of students with both a general assessment score and a score on Mini-Assessment 1A was 41,581).*



Notes: The blue trend lines are regression lines for the 10th, 20th, … 90th quantiles. The heavier line in the first three plots is the 50th quantile (i.e., the median regression line). The Probability of Indicator Mastery is expressed as a percentage.

Discussion

The pooled dataset we used for the IRT calibrations showed unidimensionality at both the domain- and overall-levels (although we have focused our presentation on the domain-level). This finding suggests that establishing reporting scales for distributed systems of assessment can be achieved on data from students taking differing assessments at different points in time - although such work should be sensitive to the way the assessments are being used, as suggested below. This finding could also suggest, at least preliminarily, that the unidmensionality of these types of assessments are robust to the variations in patterns of administration found across schools, districts and states. In future research, we plan to investigate whether this is the case directly by examining the undimensionality of the scale and the invariance of the parameters across time and educational-levels. One way to do so would be to conduct multigroup IRT analysis using time (e.g., month of administration) as the grouping parameter. Such an approach, however, would have to contend with the fact that some mini-assessments have sparse data in particular months. We could also employ similar grouping techniques to examine essential dimensionality through procedures like DIMTEST and DETECT.

The way in which the mini-assessments are used – administered once as quick check proximal to instruction – could influence the results of our dimensionality analysis. This pattern of administration also affected the results of our cut-score analysis, but more clearly. The mean total scores across the academic year were relatively stable on the mini-assessment (shown in Appendix D), while the mean scores on the general assessment increased across the year. Given this, there is an open question as to what window of general assessment results should be used to define the cut scores. The total sample of students may weigh the beginning of the year too heavily, when students have not yet mastered the standards measured on the general assessment and each mini-assessment. However, the mini-assessments are meant to be used flexibly throughout the year, so any approach to creating cut scores cannot simply exclude the beginning of the year. Our first pass at adjusting the sample appeared unsuccessful and will bear additional investigation. These investigations will involve both additional modeling and work with content experts – to insure that the cut scores do indeed provide information on mastery in ways that are consistent with content knowledge. There is also an open question of whether cut-scores should be set at the level of each mini-assessment, at the domain-level, or both. Mastery of a domain is logically the composite of the standards within it, and we hope to also address this point in future work.

Finally, our working hypothesis around the stability of the total scores on the mini-assessments is rooted in the idea that the educators and administrators choose to administer each mini-assessment based on the scope and sequence of instruction of their classroom, school or district. Thus while the mini-assessments are administered at different points in time during the year, the groups of students assessed have roughly the same level of learning on the particular standard.

References

Bennett, R. E., Kane, M. T., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Cizek, G. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Ho, A. D. (2011). *Supporting growth interpretations using through-course assessments*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Kolen, M. J. (2011). Generalizability and reliability: Approaches for through-course assessments. . Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Linacre, J. M. (2016). Winsteps Rasch measurement computer program [Computer Software]. Beaverton, Oregon: Winsteps.com.

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

Ryan, J. (2010). Envisioning a state educational System: Improving learning through a comprehensive assessment system. Olympia, Washington: Office of the Superintendent of Public Instruction.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Valencia, S., Pearson, P. D., & Wixson, K. (2011) *Tracking progress in reading: The search for keystone elements in predicting college and career readiness*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Way, McClarty, Murphy, Keng & Fuhrken (2011, April). *Through-course common core assessments in the united states: Can summative assessment be formative*? Paper presented at the annual meeting of the American Educational Research Association New Orleans, LA.

Wise, L. (2011). *Picking up the pieces: Aggregating results from through-course assessments*. Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments.* Paper presented at the Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA.

Appendix A: Indicator Classifications on the General Assessment

The indicator classifications are created through a multi-step process in which:

- a test characteristic curve (TCC) for the items aligned to the standard(s) that define the indicator is created, using the item parameters from the General's unidimensional vertical scale;
- the TCC is evaluated at the point that corresponds to an expected true score of 67% of the total possible points for the set of aligned items (referred to as the response probability of 67% or "RP67") to obtain the scale score value at RP67;
- the RP67 is then used to create indicator classifications – specifically, the RP67 is compared against the General domain score that encompasses the indicator standard. For example, if the indicator is aligned to the standard 4.MD.C.5.a, the RP67 is compared to the student's MD (Measurement and Data) domain score. If the student's domain score is greater than or equal to the RP67 value, they are identified as not needing instruction.

Appendix B: Mini-Assessment Descriptive Statistics

No.	Standard	RP50	RP67	Form A		Form B	
				# Admin.	Max Points	# Admin.	Max Points
1	4.NBT.A.1	489	509	42082	10	6924	9
2	4.NBT.A.2	431	453	29838	8	4368	10
3	4.NBT.B.4	437	458	29243	10	5999	11
4	4.NBT.A.3	458	479	26071	10	4037	9
5	4.OA.A.1	476	499	20592	9	3820	9
6	4.OA.A.2	469	491	27362	8	6536	9
7	4.OA.B.4	481	500	21085	10	5712	9
8	4.OA.C.5	473	495	10994	10	4143	9
9	4.OA.A.3	485	504	13742	9	3825	9
10	4.OA.A.3	485	504	13310	9	3860	9
11	4.NBT.B.5	472	492	29956	10	8467	10
12	4.NBT.B.6	461	481	21461	9	3989	8
13	4.NF.A.1	459	479	15918	8	6938	7
14	4.NF.A.2	469	489	13348	10	5213	8
15	4.NF.B.3a	460	485	13020	10	4942	11
16	4.NF.B.3c*	472	497	11223	11	5343	7
18	4.NF.B.4a*	472	497	9198	9	2809	9
19	4.NF.B.4c*	472	497	10324	8	3406	7
20	4.NF.C.5	475	494	6199	8	1523	7
21	4.NF.C.6	461	481	7773	6	1154	9
22	4.NF.C.7	456	483	6624	8	635	7
23	4.MD.A.1	488	515	6246	9	2685	11
24	4.MD.A.2	518	540	4100	10	721	8
25	4.MD.A.2	518	540	5414	10	832	10
26	4.MD.A.3	474	497	9194	8	3090	9
27	4.MD.B.4	503	524	5232	8	1458	7
28	4.MD.C.5*	466	495	5637	9	2474	10
29	4.MD.C.7	488	510	5826	8	1996	10
30	4.G.A.1	438	461	8738	9	3123	9
31	4.G.A.2	489	511	7680	10	1646	11
32	4.G.A.3	411	441	4917	11	941	9

Notes: *The standards for four mini-assessments (#16, 18, 19 and 28, aligned to standards 4.NF.B.3c, 4.NF.B.4a, 4.NF.B.4c and 4.MD.C.5, respectively) are not covered directly by any of the indicators. In these cases, we used the weighted average RP67 from the available standards that were covered by the general indicators, albeit at the next grain-size up. For example, to create an RP67 for the iSM #16, aligned to standard 4.NF.B.3c, we took the average of the RP67 values from 4.NF.B.3a and 4.NF.B.4b, weighted by the number of items aligned to each standard on the general assessment.

Appendix C: Dimensionality Study of iSM Domains

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	124.0749	100.0%		100.0%
Raw variance explained by measures =	52.0749	42.0%		41.9%
Raw variance explained by persons =	31.1200	25.1%		25.0%
Raw variance explained by items =	20.9549	16.9%		16.9%
Raw unexplained variance (total) =	72.0000	58.0%	100.0%	58.1%
Unexplned variance in 1st contrast =	1.5369	1.2%	2.1%	
Unexplned variance in 2nd contrast =	1.4616	1.2%	2.0%	
Unexplned variance in 3rd contrast =	1.3436	1.1%	1.9%	
Unexplned variance in 4th contrast =	1.2988	1.0%	1.8%	
Unexplned variance in 5th contrast =	1.2689	1.0%	1.8%	

STANDARDIZED RESIDUAL VARIANCE SCREE PLOT

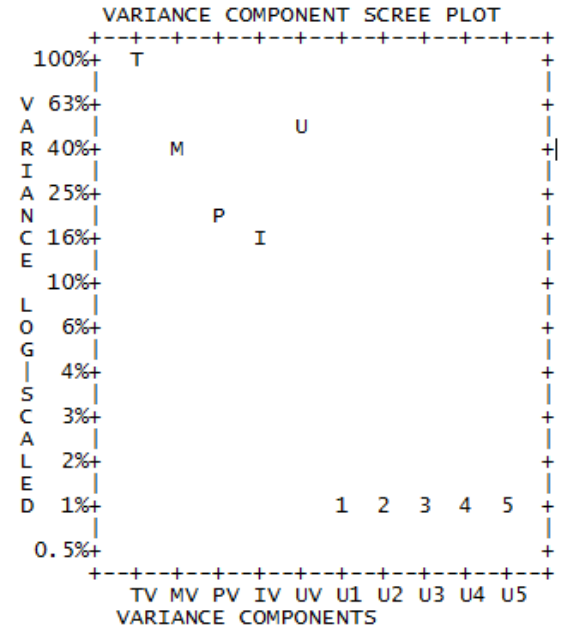


Figure A.1. Principal Component Analysis of Residuals of Math Grade 4 Domain 1: Number & Operations in Base Ten

```
     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                              Eigenvalue   Observed   Expected
Total raw variance in observations     =     127.5017 100.0%          100.0%
  Raw variance explained by measures   =      55.5017  43.5%           43.1%
    Raw variance explained by persons  =      31.4373  24.7%           24.4%
    Raw variance explained by items    =      24.0644  18.9%           18.7%
  Raw unexplained variance (total)     =      72.0000  56.5% 100.0%    56.9%
    Unexplned variance in 1st contrast =       1.3464   1.1%   1.9%
    Unexplned variance in 2nd contrast =       1.3019   1.0%   1.8%
    Unexplned variance in 3rd contrast =       1.2543   1.0%   1.7%
    Unexplned variance in 4th contrast =       1.2038    .9%   1.7%
```

STANDARDIZED RESIDUAL VARIANCE SCREE PLOT

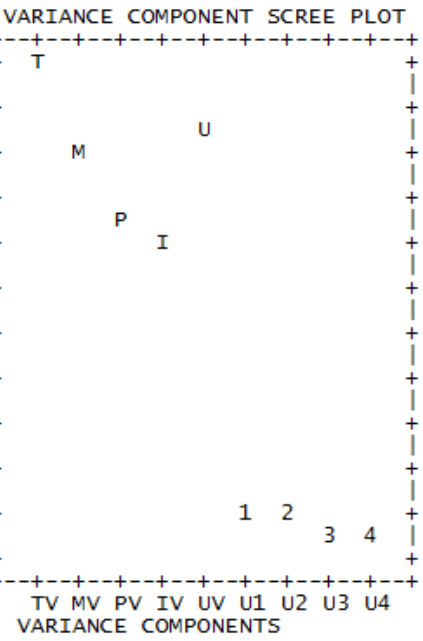


Figure A.2. Principal Component Analysis of Residuals of Math Grade 4 Domain 2: Operations & Algebraic Thinking

```
     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                          Eigenvalue   Observed   Expected
Total raw variance in observations     =     184.1389 100.0%         100.0%
  Raw variance explained by measures   =      77.1389  41.9%          42.3%
    Raw variance explained by persons  =      51.1226  27.8%          28.0%
    Raw variance explained by items    =      26.0163  14.1%          14.3%
  Raw unexplained variance (total)     =     107.0000  58.1% 100.0%   57.7%
    Unexplned variance in 1st contrast =       1.7106    .9%   1.6%
    Unexplned variance in 2nd contrast =       1.5454    .8%   1.4%
    Unexplned variance in 3rd contrast =       1.4821    .8%   1.4%
    Unexplned variance in 4th contrast =       1.4441    .8%   1.3%
    Unexplned variance in 5th contrast =       1.3706    .7%   1.3%
```

STANDARDIZED RESIDUAL VARIANCE SCREE PLOT

Figure A.3. Principal Component Analysis of Residuals of Math Grade 4 Domain 3: Number & Operations–Fractions

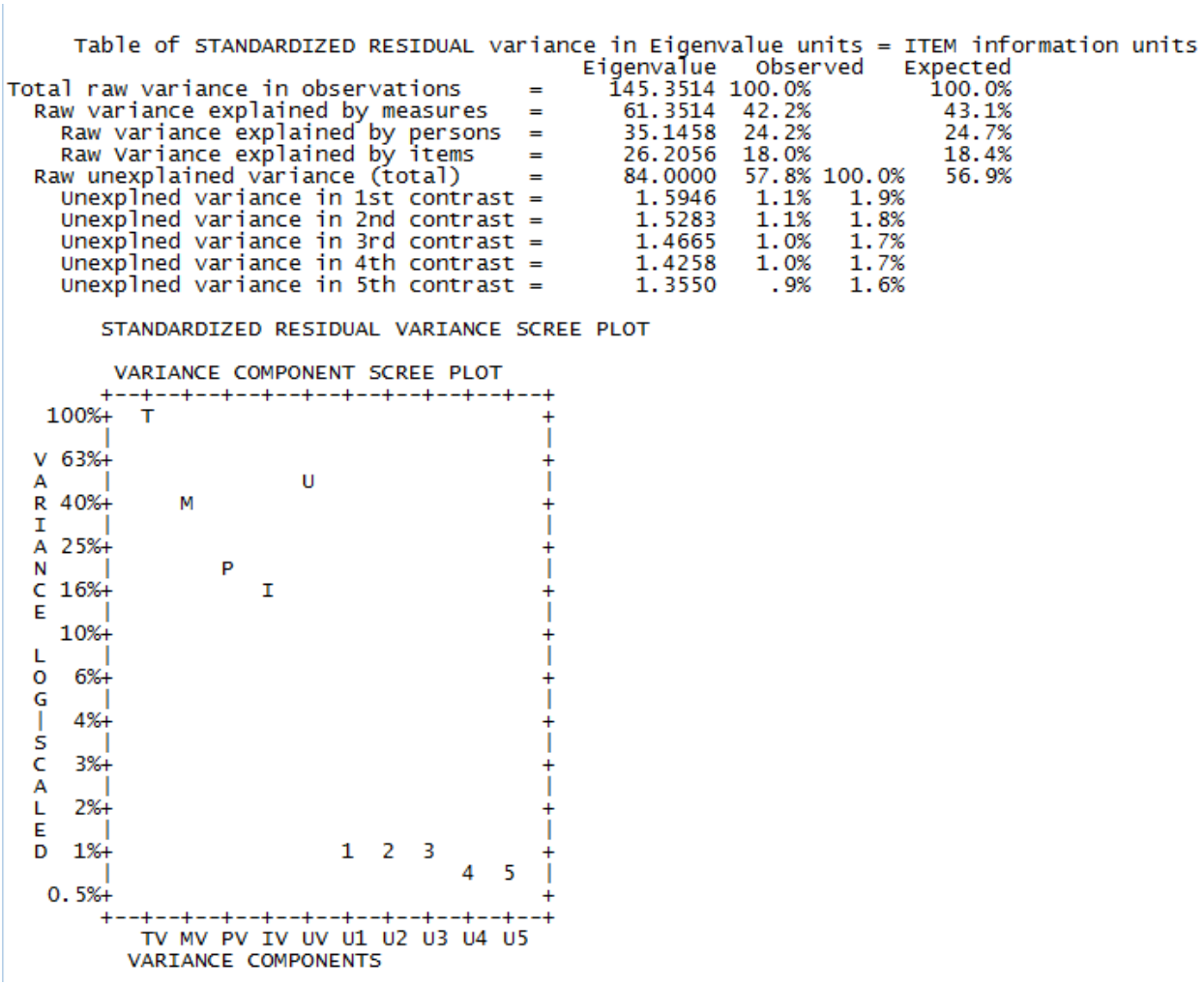


Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	145.3514	100.0%		100.0%
Raw variance explained by measures =	61.3514	42.2%		43.1%
Raw variance explained by persons =	35.1458	24.2%		24.7%
Raw Variance explained by items =	26.2056	18.0%		18.4%
Raw unexplained variance (total) =	84.0000	57.8%	100.0%	56.9%
Unexplned variance in 1st contrast =	1.5946	1.1%	1.9%	
Unexplned variance in 2nd contrast =	1.5283	1.1%	1.8%	
Unexplned variance in 3rd contrast =	1.4665	1.0%	1.7%	
Unexplned variance in 4th contrast =	1.4258	1.0%	1.7%	
Unexplned variance in 5th contrast =	1.3550	.9%	1.6%	

STANDARDIZED RESIDUAL VARIANCE SCREE PLOT

Figure A.4. Principal Component Analysis of Residuals of Math Grade 4 Domain 4: Measurement & Data

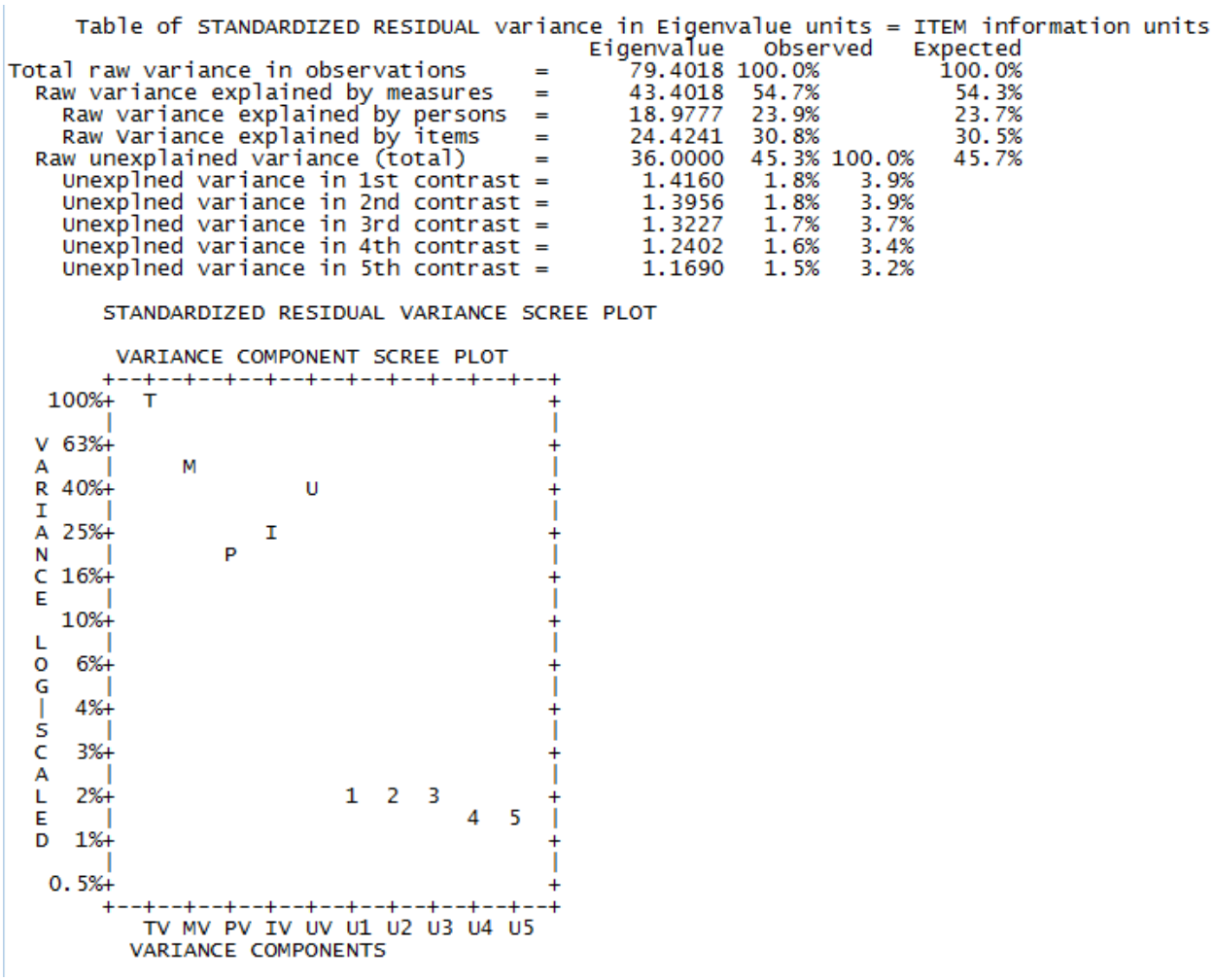


Figure A.5. Principal Component Analysis of Residuals of Math Grade 4 Domain 5: Geometry

Appendix D: Quantile Regression Results

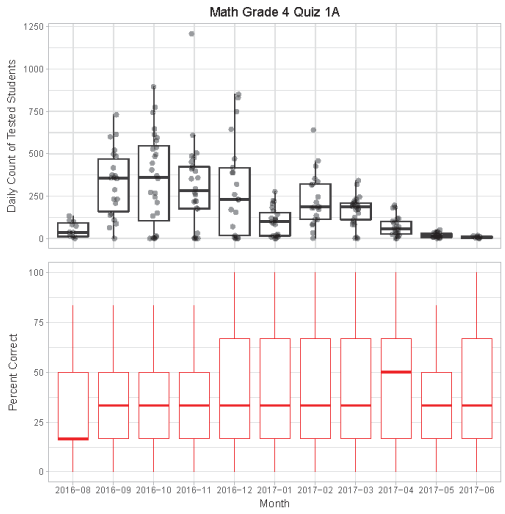
Table 1D. *Quantile Regression Results - Total Scores Resulting from the Evaluation of the 50th Quantile at Response Probabilities of 0.25, 0.50 and 0.67 .*

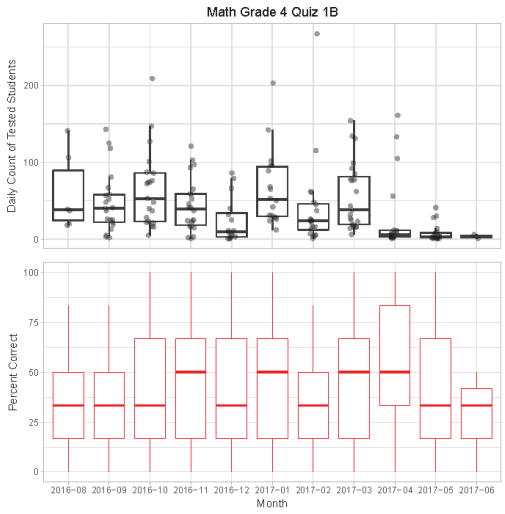
	All Data			Matched			Max Possible Score
	RP25	RP50	RP67	RP25	RP50	RP67	
1A	2.44	7.78	11.41	2.22	7.10	10.43	9
1B	2.61	7.24	10.38	2.52	6.79	9.69	9
2A	1.74	5.14	7.45	1.31	4.73	7.06	10
2B	-0.41	4.53	7.90	-0.04	4.87	8.20	10
3A	2.00	6.47	9.52	1.62	6.04	9.04	9
3B	1.95	6.36	9.36	2.13	6.57	9.60	8
4A	3.18	8.18	11.58	2.92	7.58	10.75	8
4B	1.90	6.85	10.22	1.84	6.86	10.28	7
5A	7.48	14.35	19.02	7.13	13.20	17.32	10
5B	6.35	12.94	17.43	6.56	13.49	18.21	8
6A	2.89	6.25	8.53	2.86	6.12	8.33	10
6B	2.16	6.27	9.06	2.19	6.46	9.37	11
7A	3.08	7.04	9.73	2.77	6.92	9.73	11
7B	3.28	7.25	9.96	3.34	7.26	9.93	7
8A	4.69	10.13	13.84	4.65	10.74	14.88	9
8B	4.49	8.67	11.52	4.67	9.02	11.98	9
9A	1.95	5.80	8.42	2.10	6.11	8.85	8
9B	1.47	5.67	8.52	1.94	6.39	9.42	7
10A	4.07	9.27	12.80	4.20	9.64	13.33	10
10B	4.44	10.40	14.46	5.34	12.41	17.22	9
11A	2.10	7.57	11.28	1.79	7.26	10.97	8
11B	1.41	6.31	9.64	1.53	6.58	10.01	7
12A	-0.59	4.53	8.01	-0.44	4.72	8.23	6
12B	0.03	4.52	7.57	0.05	4.60	7.69	9
13A	3.44	7.98	11.07	3.24	7.38	10.20	8
13B	3.02	7.20	10.04	2.67	6.36	8.86	7
14A	2.87	8.27	11.95	2.68	7.78	11.25	9
14B	2.37	6.50	9.31	2.17	5.95	8.52	11
15A	3.77	8.99	12.54	3.78	8.72	12.07	10
15B	5.41	10.50	13.95	5.25	10.11	13.41	8
16A	3.74	10.17	14.55	3.67	9.73	13.84	10
16B	2.70	7.32	10.46	3.20	8.18	11.56	10
18A	8.45	18.61	25.53	7.72	16.57	22.59	8
18B	8.70	19.59	26.99	9.77	21.41	29.32	9
19A	3.47	7.98	11.05	3.32	7.49	10.32	8

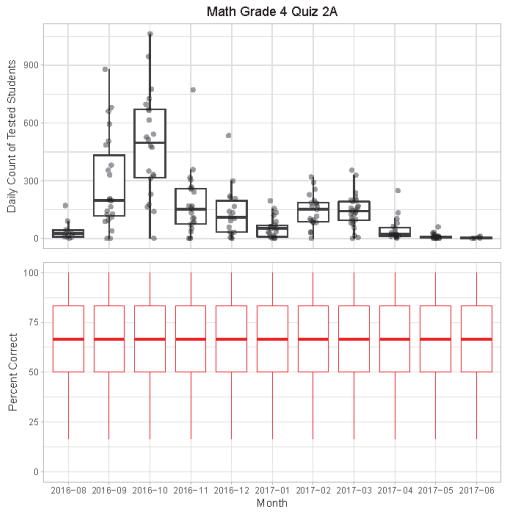
19B	2.36	6.91	10.01	2.29	6.63	9.59	7
20A	3.44	8.65	12.19	3.30	8.35	11.78	9
20B	3.92	9.15	12.71	4.18	9.87	13.75	10
21A	1.48	4.43	6.43	1.47	4.40	6.40	8
21B	5.31	9.82	12.89	5.12	9.30	12.15	10
22A	-0.24	3.71	6.39	-0.23	3.68	6.34	8
22B	-0.94	3.02	5.71	-0.75	3.11	5.74	10
23A	7.90	16.55	22.44	7.84	16.39	22.21	9
23B	9.81	18.99	25.23	11.52	22.78	30.44	9
24A	-0.19	3.84	6.59	-0.09	3.89	6.60	10
24B	-1.28	2.28	4.70	-1.46	2.27	4.80	11
25A	0.63	4.16	6.56	0.69	4.20	6.58	11
25B	-0.08	3.57	6.05	-0.09	3.75	6.37	9
26A	1.57	4.57	6.62	1.65	4.73	6.82	10
26B	2.45	6.54	9.32	2.41	6.51	9.29	11
27A	4.22	10.27	14.38	4.40	10.72	15.02	10
27B	3.54	8.87	12.49	3.82	9.76	13.81	9
28A	1.11	4.56	6.91	1.12	4.69	7.11	9
28B	2.52	7.32	10.58	2.63	6.89	9.79	9
29A	2.99	6.92	9.60	3.14	7.36	10.24	8
29B	5.11	11.24	15.41	4.58	9.47	12.80	9
30A	0.53	3.96	6.29	0.56	4.01	6.36	10
30B	1.48	5.02	7.43	1.61	5.00	7.31	9
31A	5.09	10.94	14.92	5.05	10.88	14.85	10
31B	4.60	10.64	14.75	4.90	11.80	16.49	9
32A	-7.79	-1.56	2.68	-7.53	-1.41	2.75	9
32B	-5.40	-0.82	2.29	-5.97	-1.22	2.01	9

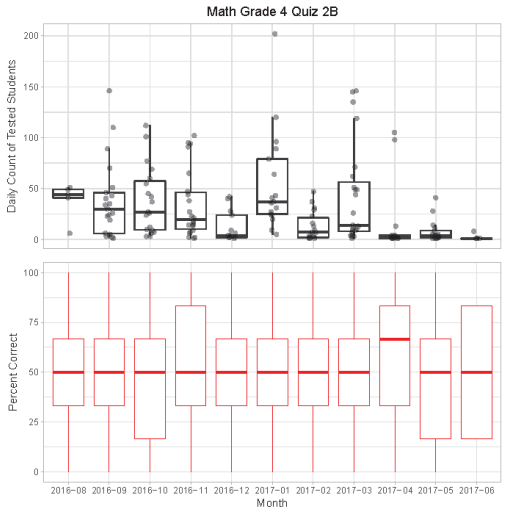
Notes: Values greater than 9 or less than 0 have been highlighted (typically the range of possible points on a mini-assessment).

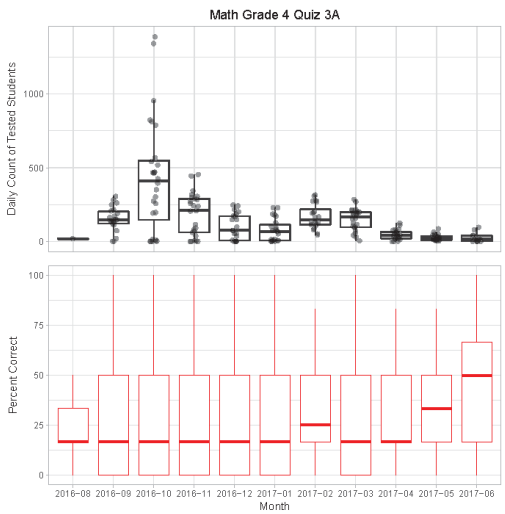
Appendix E: Smoothed Daily Administration Counts and Daily Mean Total Scores

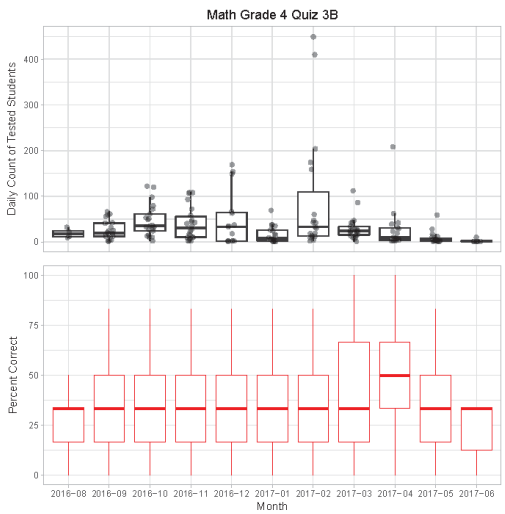

Math Grade 4 Quiz 1A
Daily Count of Tested Students
Percent Correct
Month

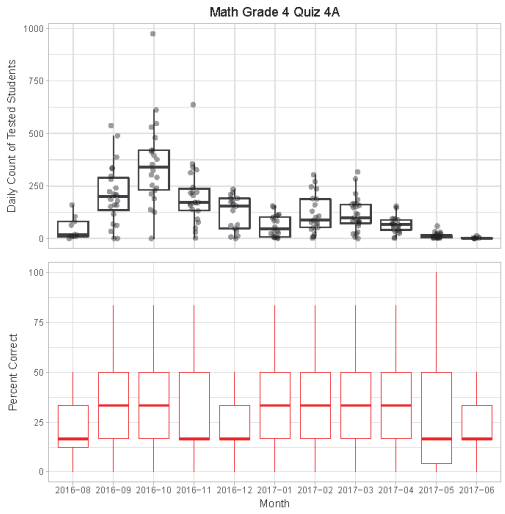

Math Grade 4 Quiz 1B
Daily Count of Tested Students
Percent Correct
Month

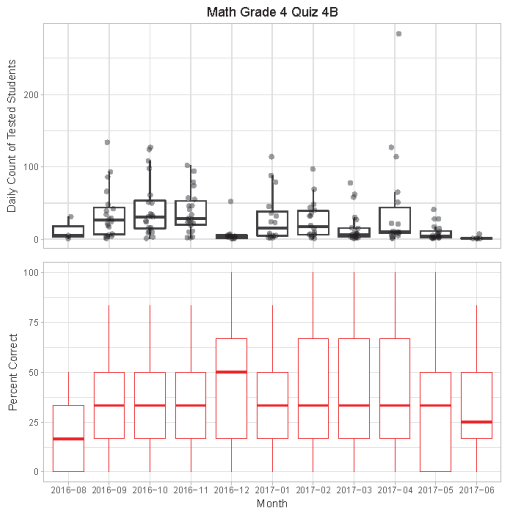

Math Grade 4 Quiz 2A
Daily Count of Tested Students
Percent Correct
Month

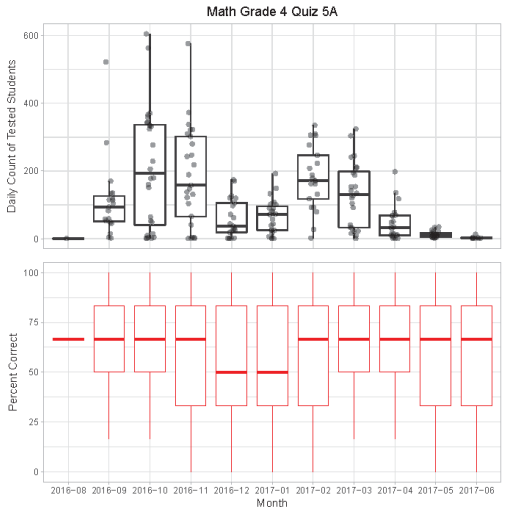

Math Grade 4 Quiz 2B
Daily Count of Tested Students
Percent Correct
Month

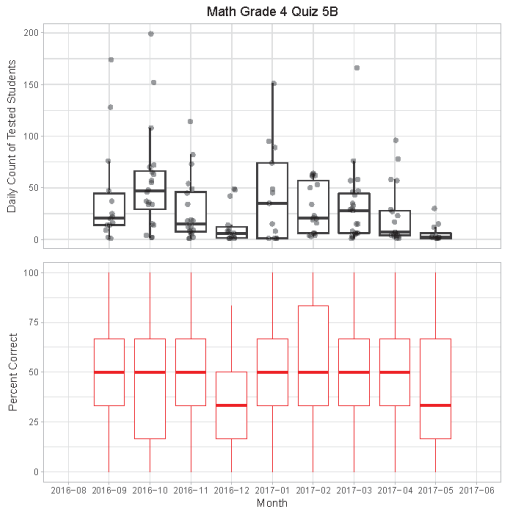

Math Grade 4 Quiz 3A
Daily Count of Tested Students
Percent Correct
Month

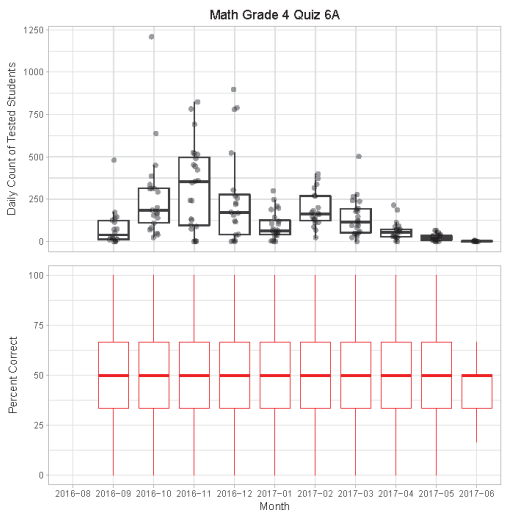

Math Grade 4 Quiz 3B
Daily Count of Tested Students
Percent Correct
Month

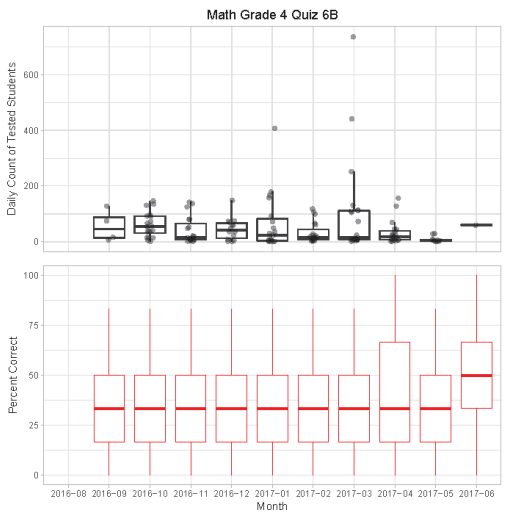

Math Grade 4 Quiz 4A
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 4B
Daily Count of Tested Students
Percent Correct
Month

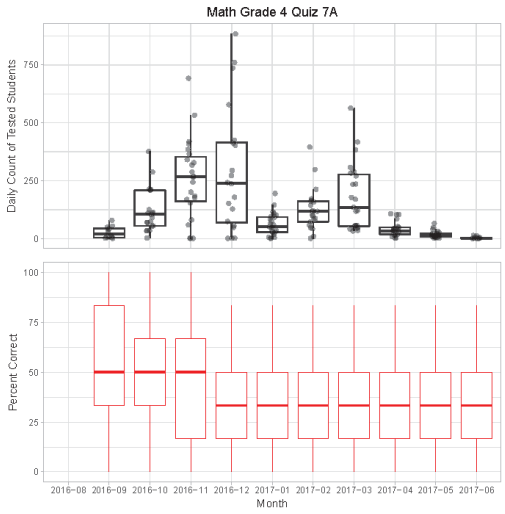

Math Grade 4 Quiz 5A
Daily Count of Tested Students
Percent Correct
Month

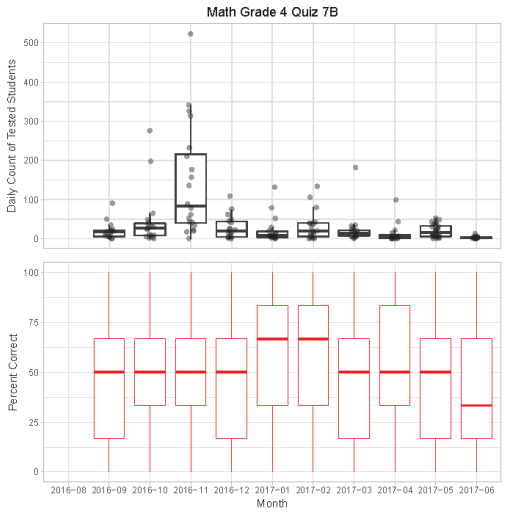

Math Grade 4 Quiz 5B
Daily Count of Tested Students
Percent Correct
Month

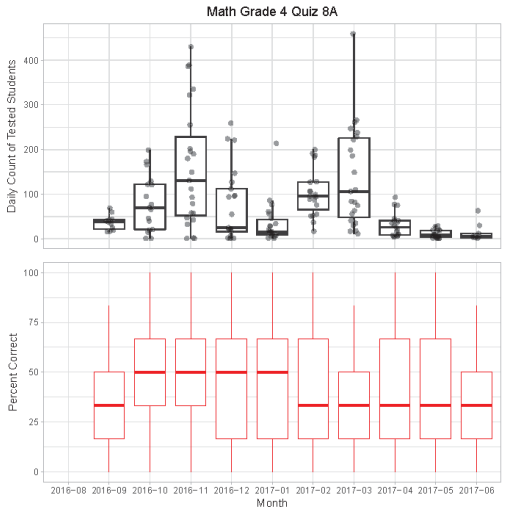

Math Grade 4 Quiz 6A
Daily Count of Tested Students
Percent Correct
Month

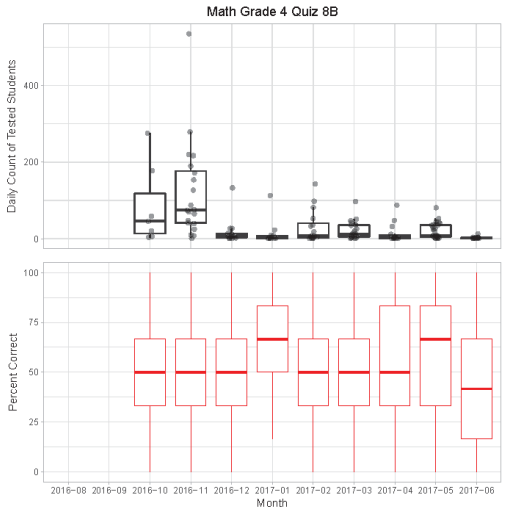

Math Grade 4 Quiz 6B
Daily Count of Tested Students
Percent Correct
Month

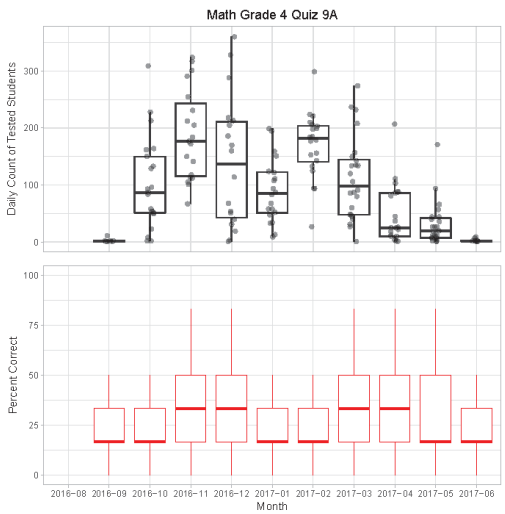
On Demand Assessments of Individual Standards

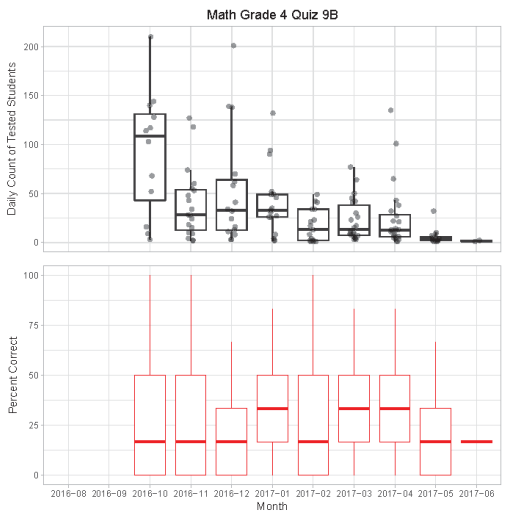


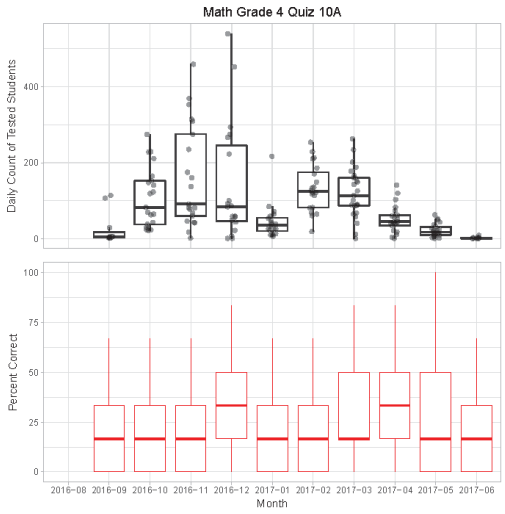


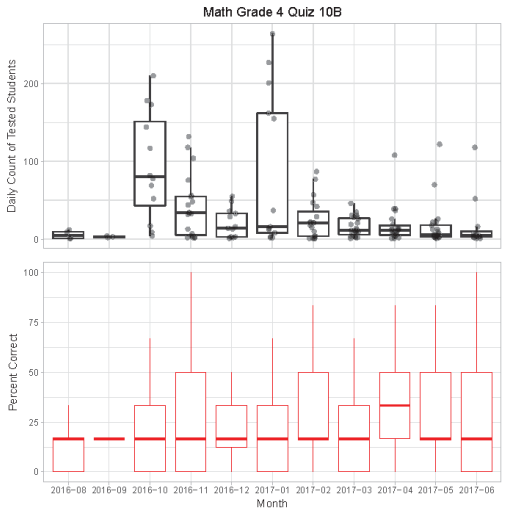


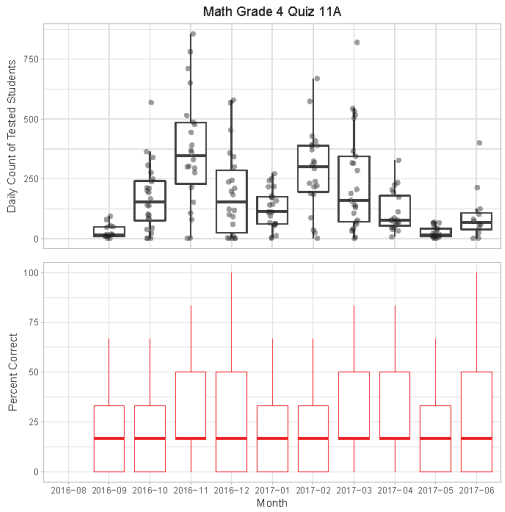


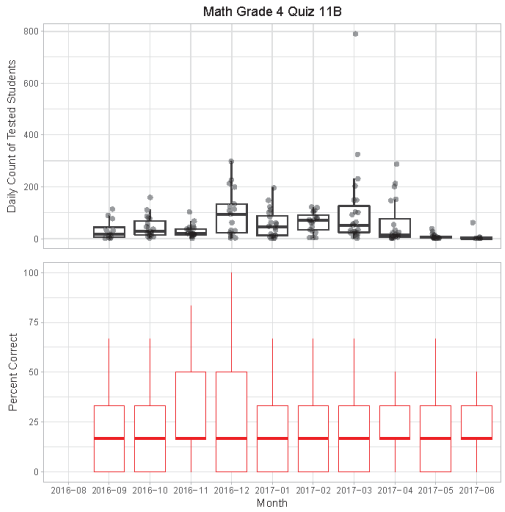


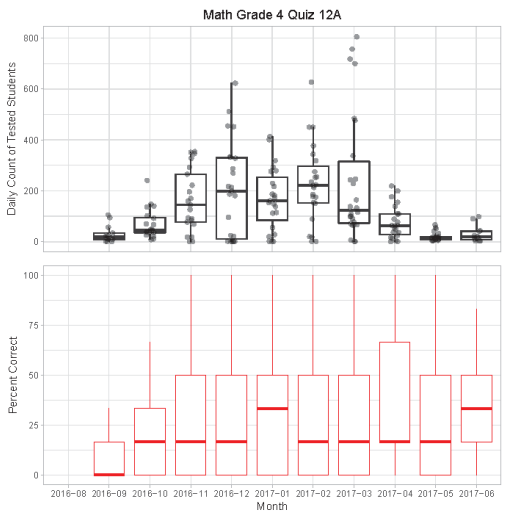


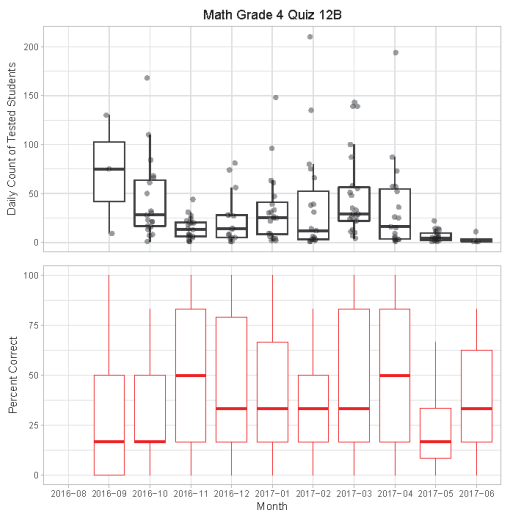

Math Grade 4 Quiz 10A
Daily Count of Tested Students
Percent Correct
Month

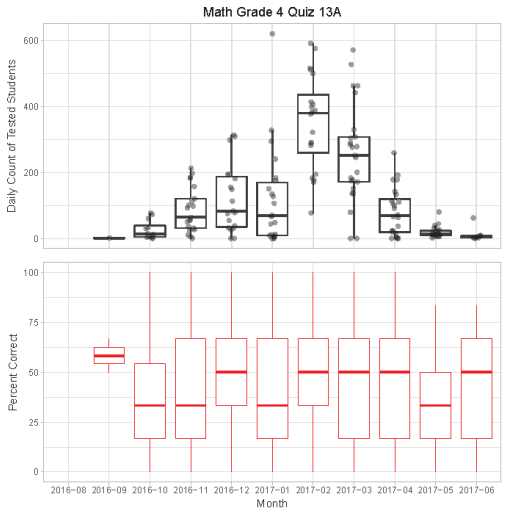

Math Grade 4 Quiz 10B
Daily Count of Tested Students
Percent Correct
Month

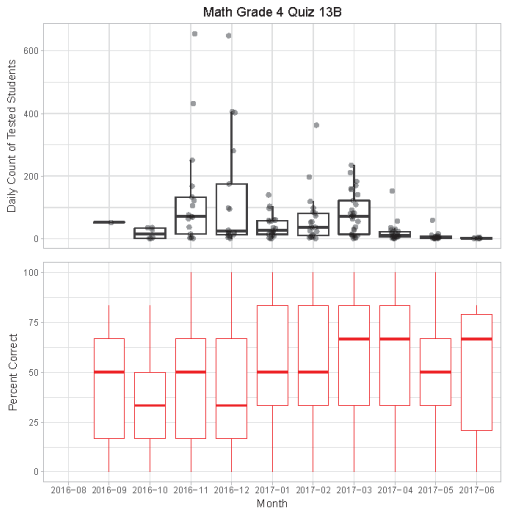

Math Grade 4 Quiz 11A
Daily Count of Tested Students
Percent Correct
Month

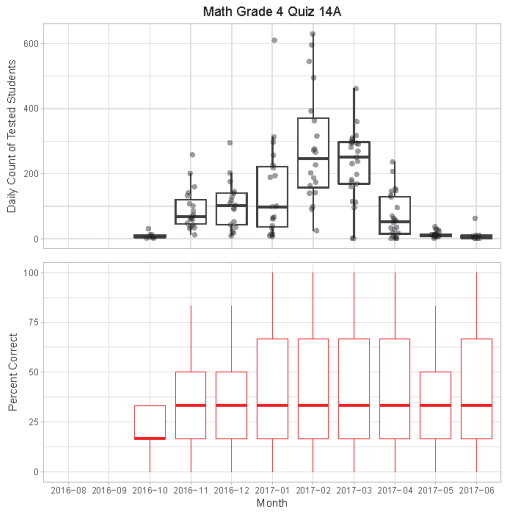

Math Grade 4 Quiz 11B
Daily Count of Tested Students
Percent Correct
Month

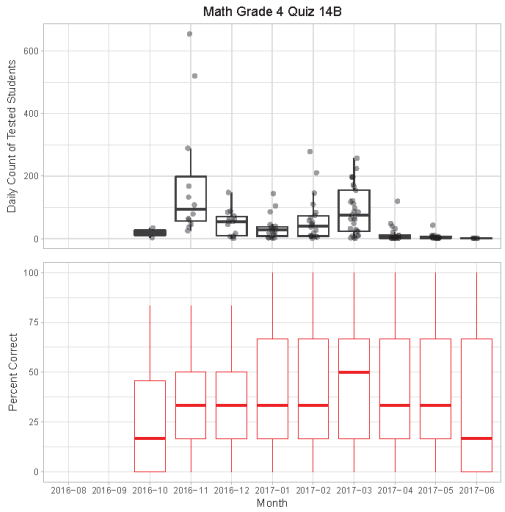

Math Grade 4 Quiz 12A
Daily Count of Tested Students
Percent Correct
Month

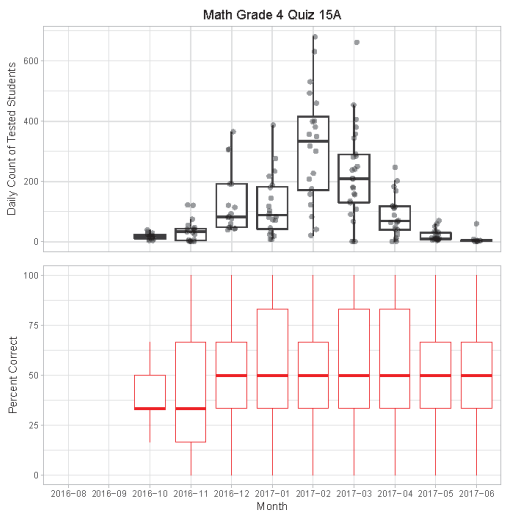

Math Grade 4 Quiz 12B
Daily Count of Tested Students
Percent Correct
Month

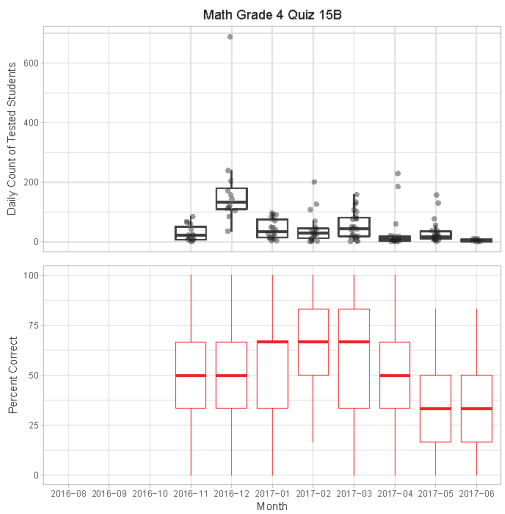

Math Grade 4 Quiz 13A
Daily Count of Tested Students
Percent Correct
Month

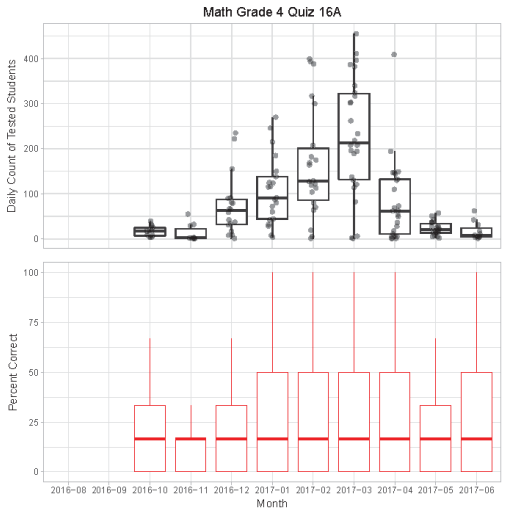

Math Grade 4 Quiz 13B
Daily Count of Tested Students
Percent Correct
Month

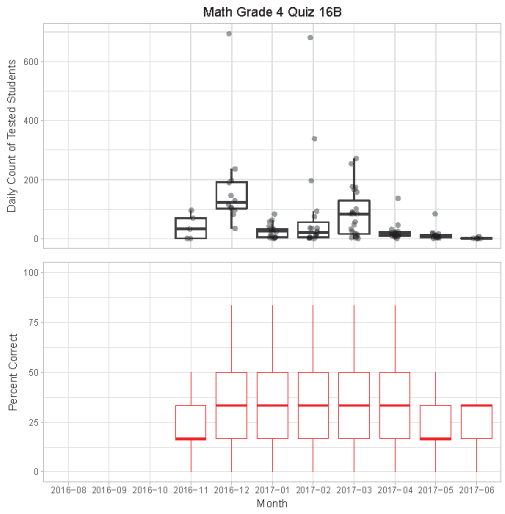

Math Grade 4 Quiz 14A
Daily Count of Tested Students
Percent Correct
Month

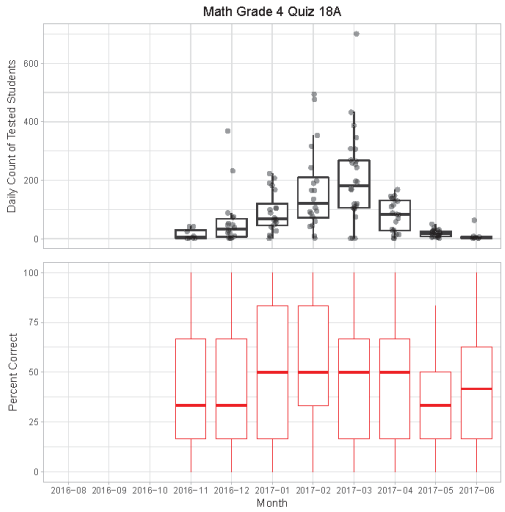

Math Grade 4 Quiz 14B
Daily Count of Tested Students
Percent Correct
Month

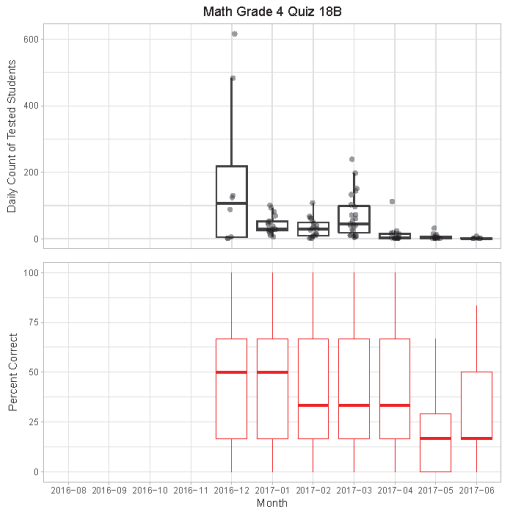

Math Grade 4 Quiz 15A
Daily Count of Tested Students
Percent Correct
Month

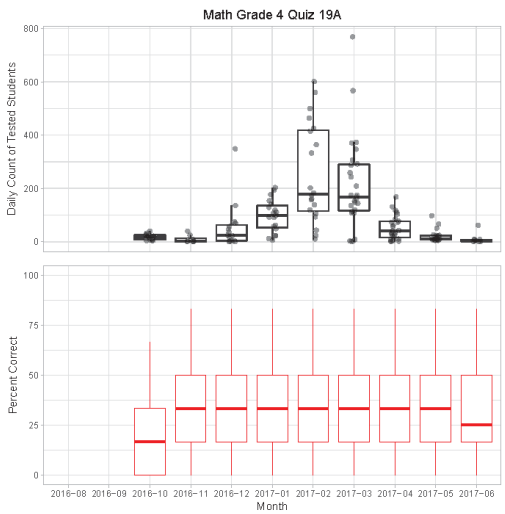

Math Grade 4 Quiz 15B
Daily Count of Tested Students
Percent Correct
Month

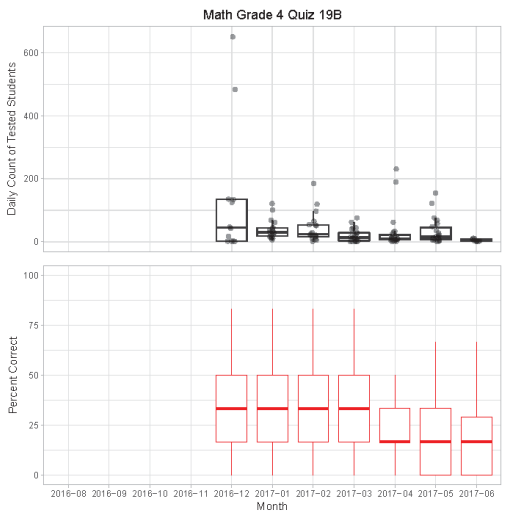

Math Grade 4 Quiz 16A
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 16B
Daily Count of Tested Students
Percent Correct
Month

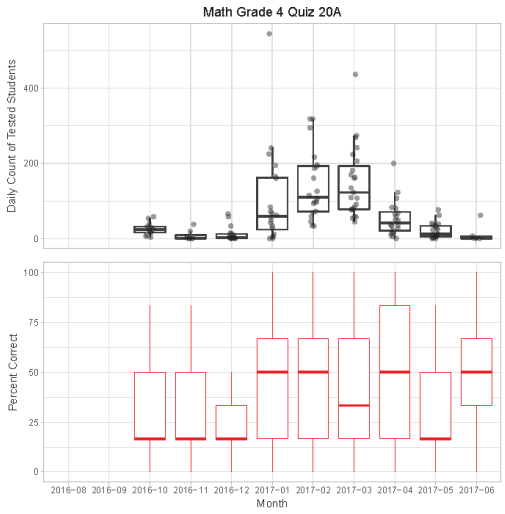

Math Grade 4 Quiz 18A
Daily Count of Tested Students
Percent Correct
Month

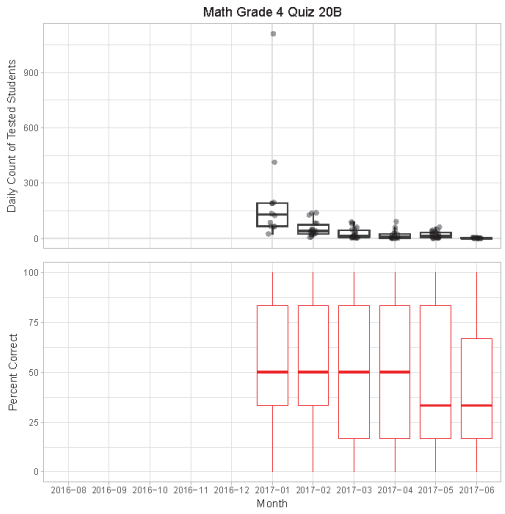

Math Grade 4 Quiz 18B
Daily Count of Tested Students
Percent Correct
Month

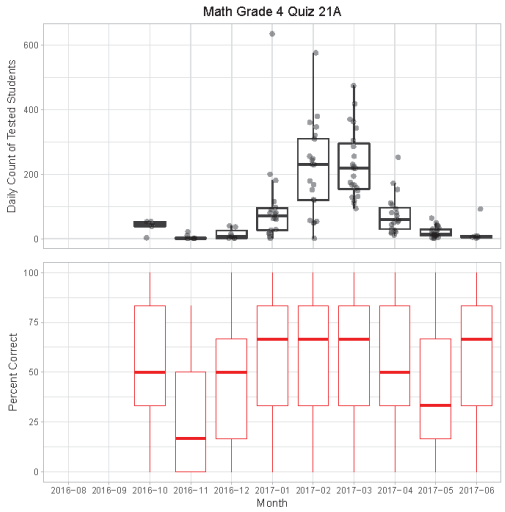

Math Grade 4 Quiz 19A
Daily Count of Tested Students
Percent Correct
Month

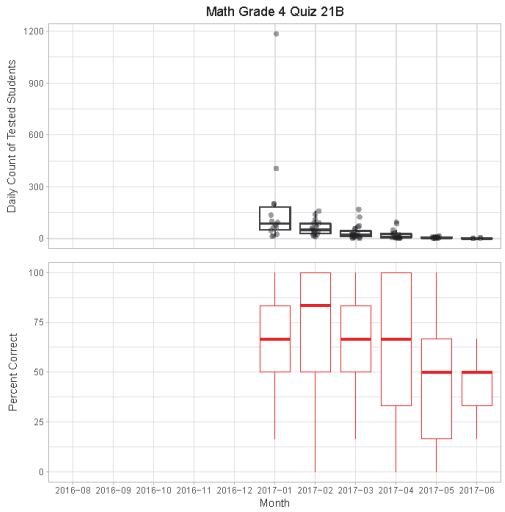

Math Grade 4 Quiz 19B
Daily Count of Tested Students
Percent Correct
Month

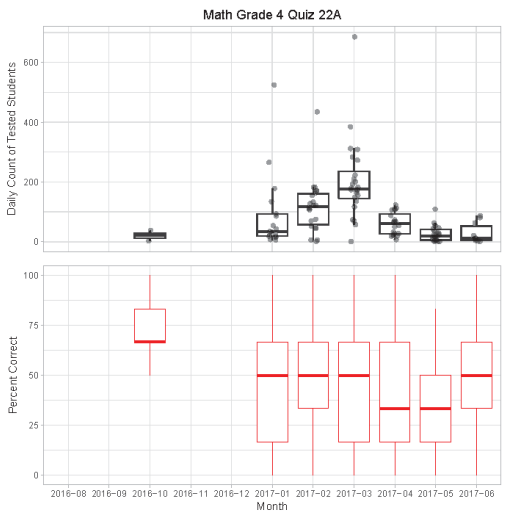
On Demand Assessments of Individual Standards

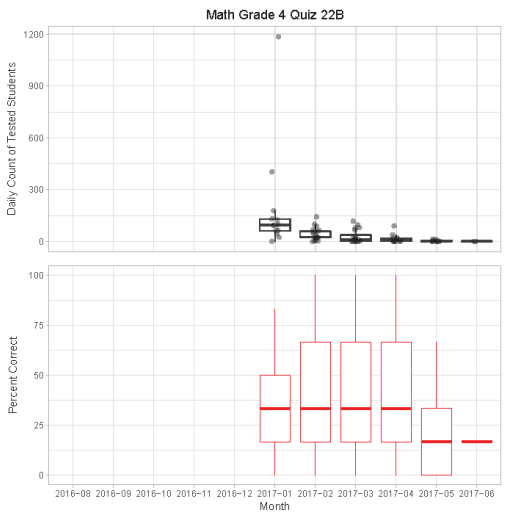


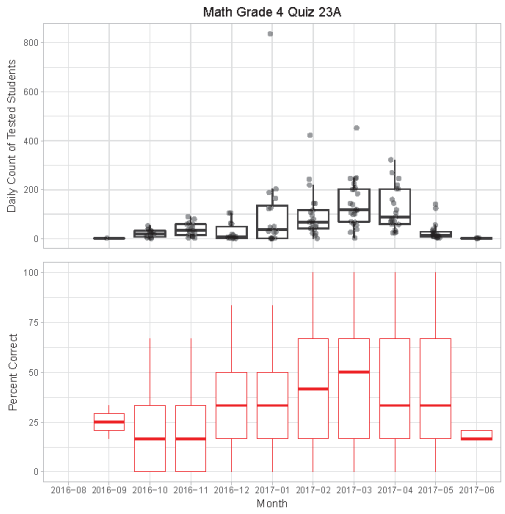


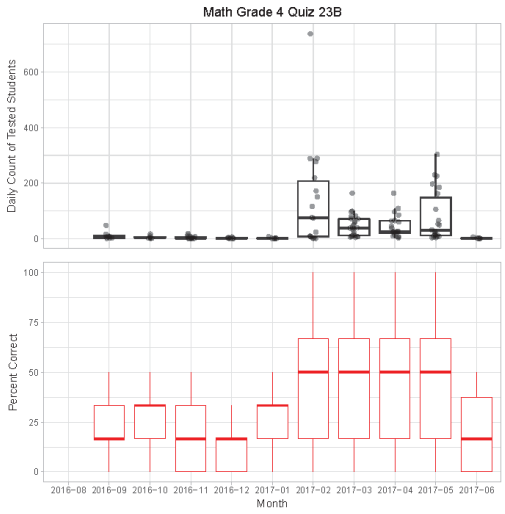


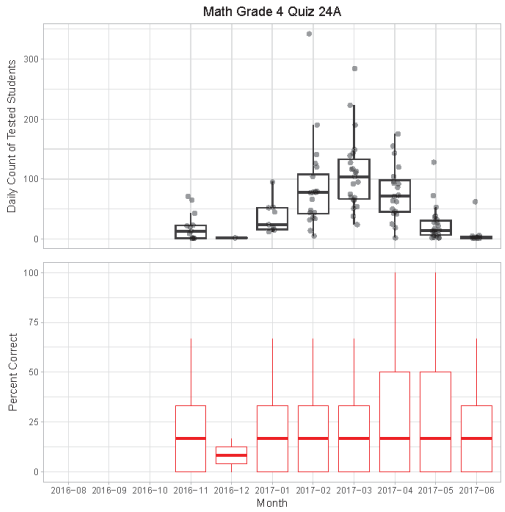


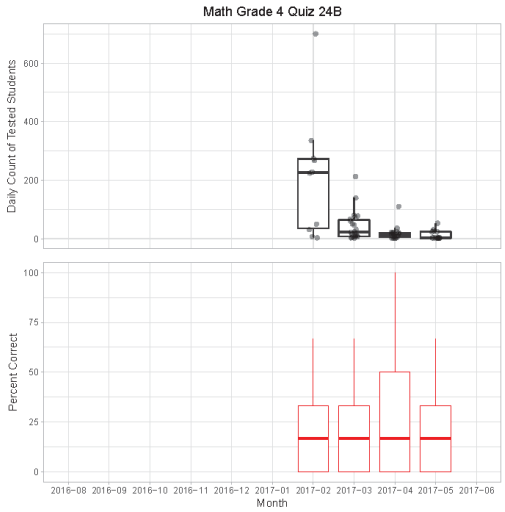


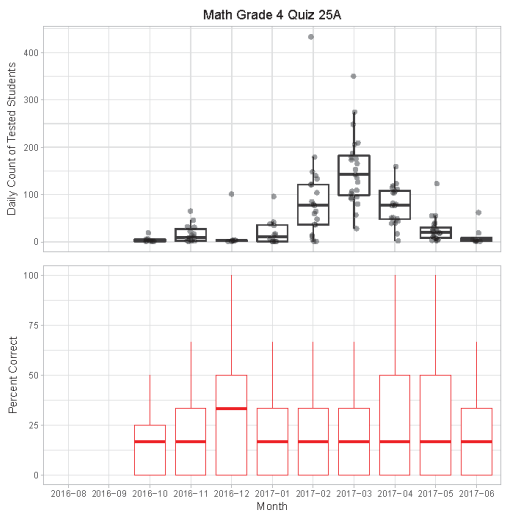


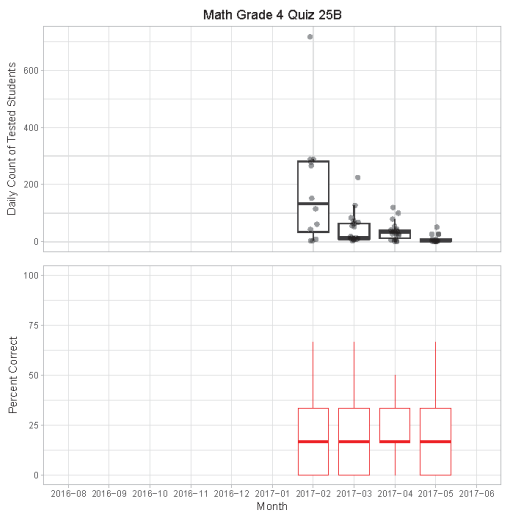

Math Grade 4 Quiz 23A
Daily Count of Tested Students
Percent Correct
Month

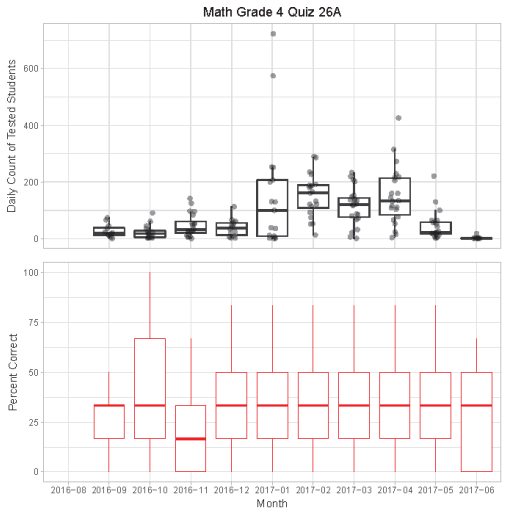

Math Grade 4 Quiz 23B
Daily Count of Tested Students
Percent Correct
Month

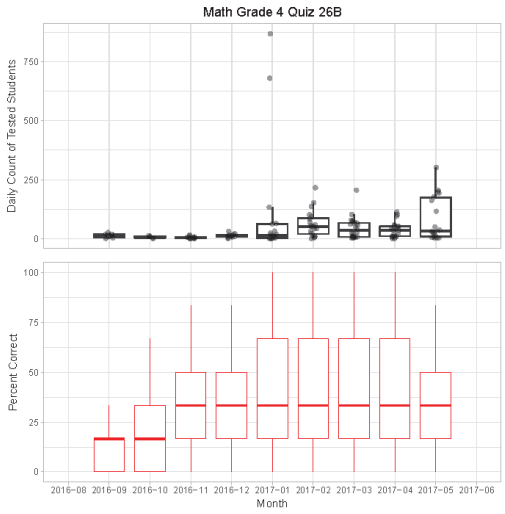

Math Grade 4 Quiz 24A
Daily Count of Tested Students
Percent Correct
Month

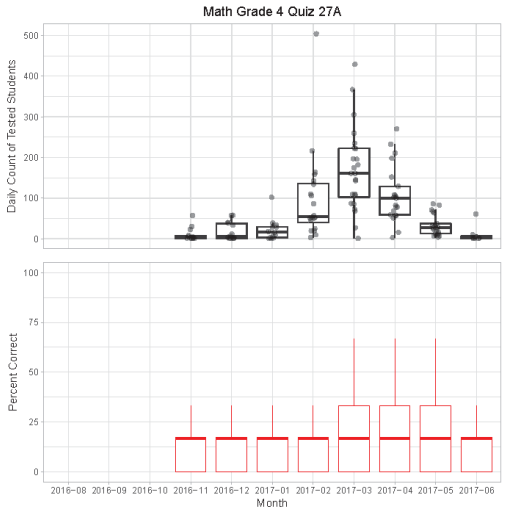

Math Grade 4 Quiz 24B
Daily Count of Tested Students
Percent Correct
Month

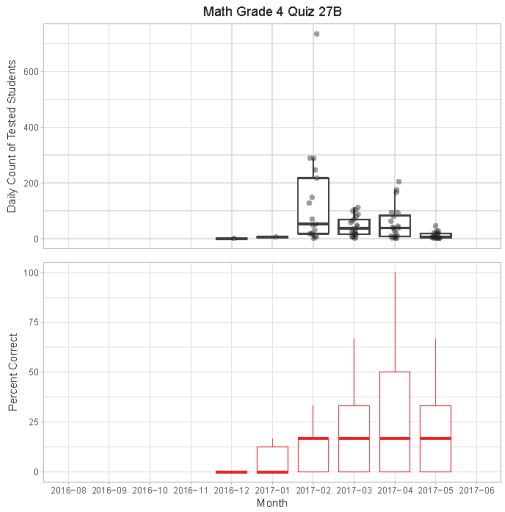

Math Grade 4 Quiz 25A
Daily Count of Tested Students
Percent Correct
Month

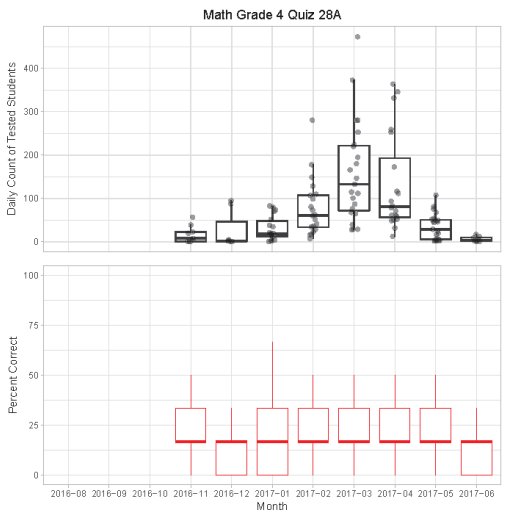

Math Grade 4 Quiz 25B
Daily Count of Tested Students
Percent Correct
Month

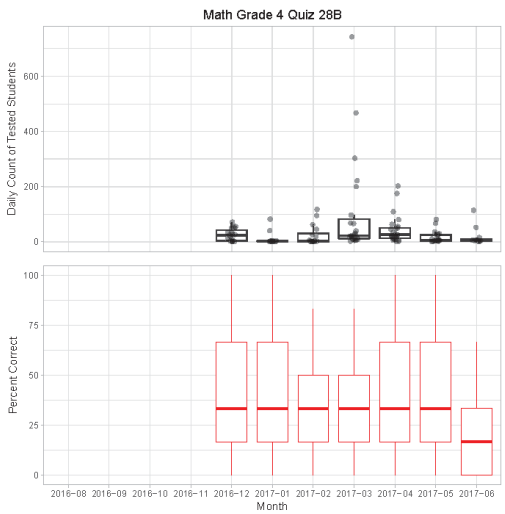

Math Grade 4 Quiz 26A
Daily Count of Tested Students
Percent Correct
Month

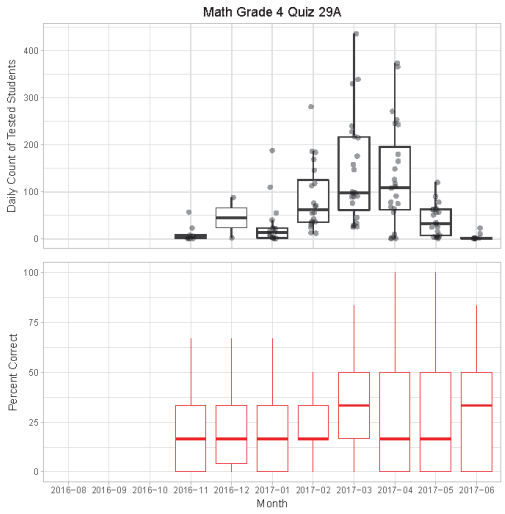

Math Grade 4 Quiz 26B
Daily Count of Tested Students
Percent Correct
Month

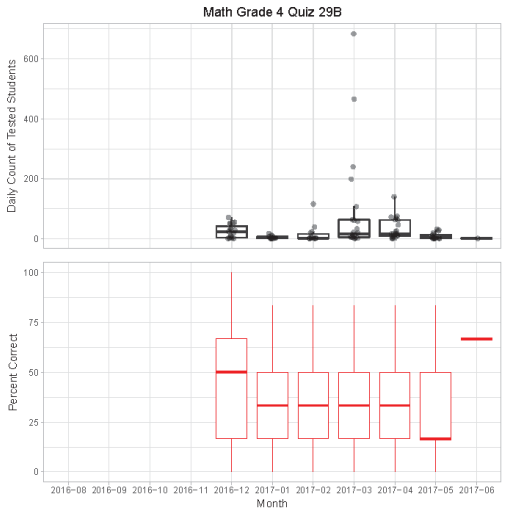

Math Grade 4 Quiz 27A
Daily Count of Tested Students
Percent Correct
Month

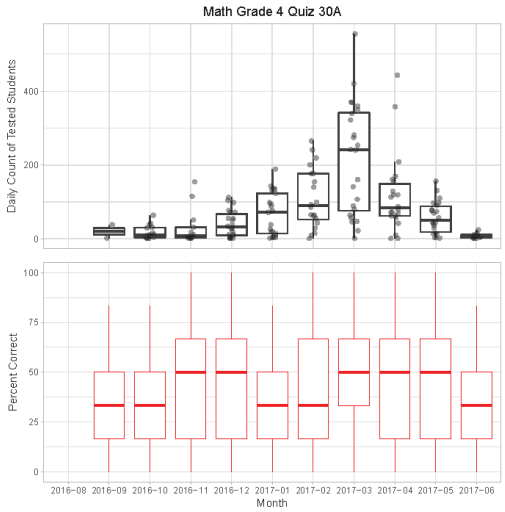

Math Grade 4 Quiz 27B
Daily Count of Tested Students
Percent Correct
Month

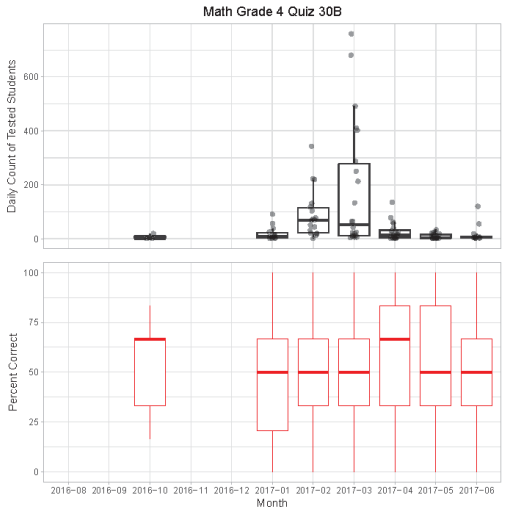

Math Grade 4 Quiz 28A
Daily Count of Tested Students
Percent Correct
Month

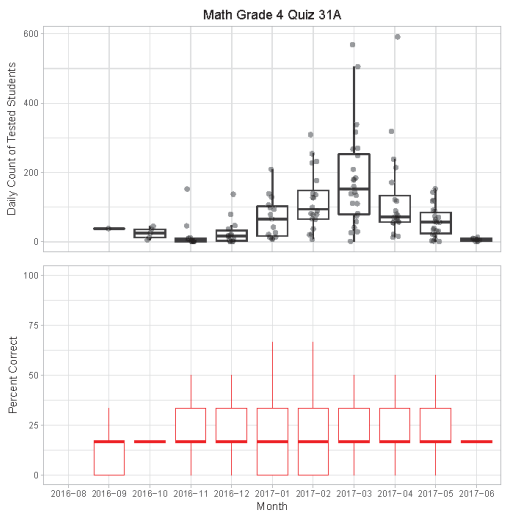

Math Grade 4 Quiz 28B
Daily Count of Tested Students
Percent Correct
Month

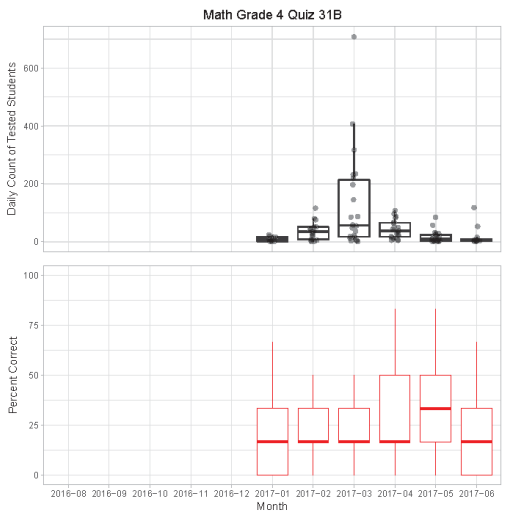

Math Grade 4 Quiz 29A
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 29B
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 30A
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 30B
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 31A
Daily Count of Tested Students
Percent Correct
Month


Math Grade 4 Quiz 31B
Daily Count of Tested Students
Percent Correct
Month

On Demand Assessments of Individual Standards