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

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

[1] Center for Assessment

[2] Curriculum Associates®

NCME - New York, NY

April 16th, 2018

Context

- Much work has been done on improving a **single assessment,** in terms of efficiency and information.
  - Although the definition of an "assessment" continues to blur.
- This work takes a different tack, instead examining how scale scores and cut scores can be developed for a **set of assessments**, motivated by the ideas around the concept of a system of assessments.

Context, Continued (Grade 4 Math)

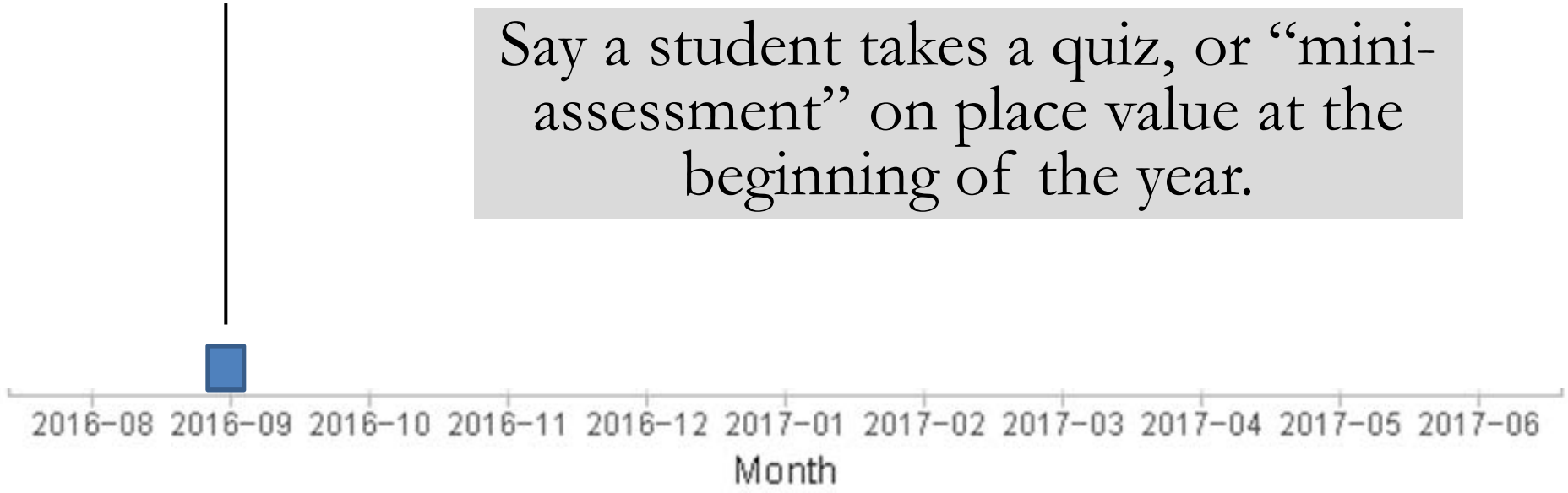
Key to this set of assessments is the idea of **modularity**.

Context, Continued (Grade 4 Math)

Key to this set of assessments is the idea of **modularity**.

Consider this hypothetical example:

1: Place Value

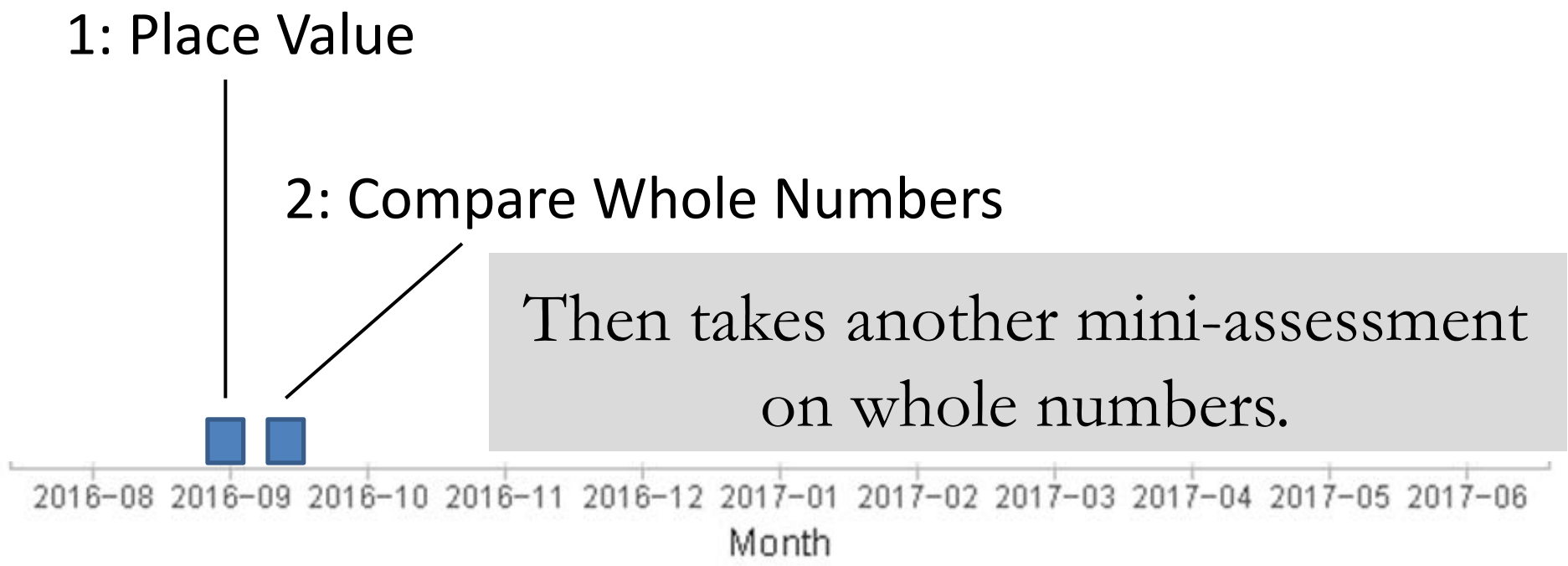


Say a student takes a quiz, or "mini-assessment" on place value at the beginning of the year.

Context, Continued (Grade 4 Math)

Key to this set of assessments is the idea of **modularity**.

Consider this hypothetical example:

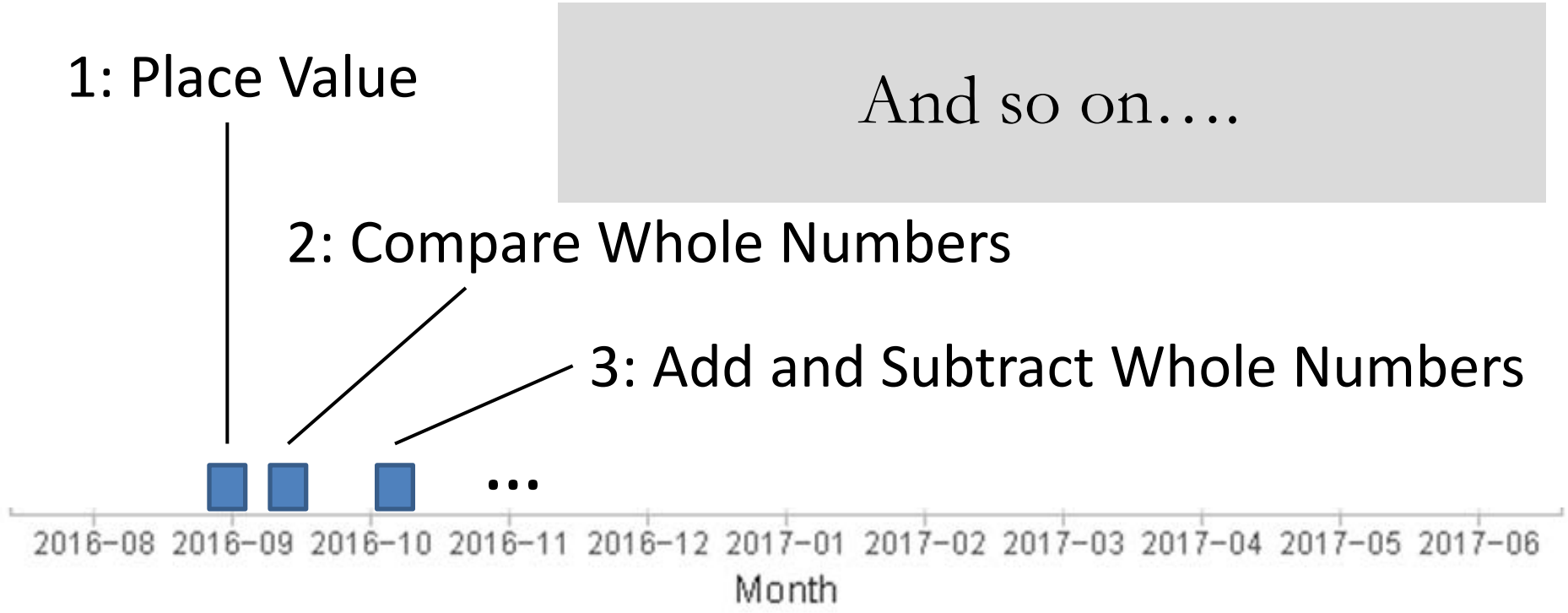


Then takes another mini-assessment on whole numbers.

Context, Continued (Grade 4 Math)

Key to this set of assessments is the idea of **modularity**.

Consider this hypothetical example:



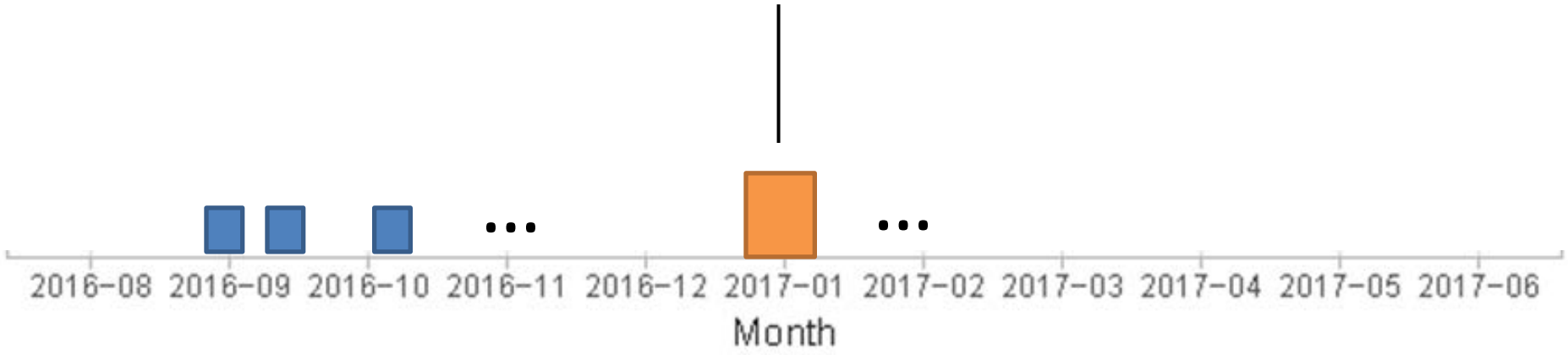
And so on….

Context, Continued (Grade 4 Math)

Key to this set of assessments is the idea of **modularity**.

Consider this hypothetical example:

Let's say the student also takes an "general" purpose assessment that surveys the full set of standards.

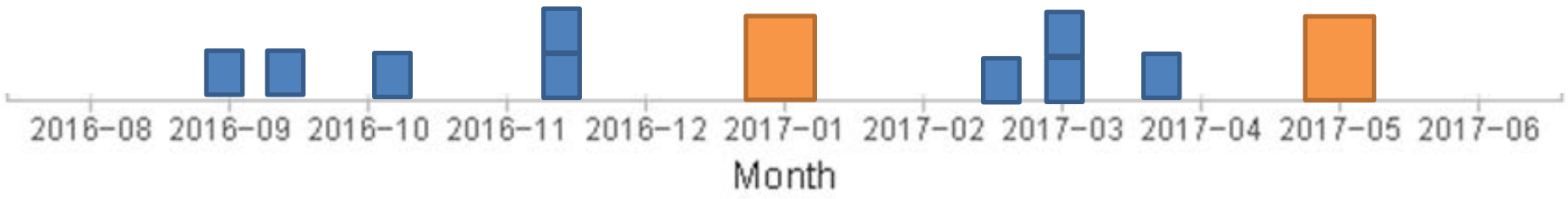


Context, Continued (Grade 4 Math)

Key to this set of assessments is the idea of **modularity**.

Consider this hypothetical example:

Then the full set of assessment this hypothetical student might look like ↓

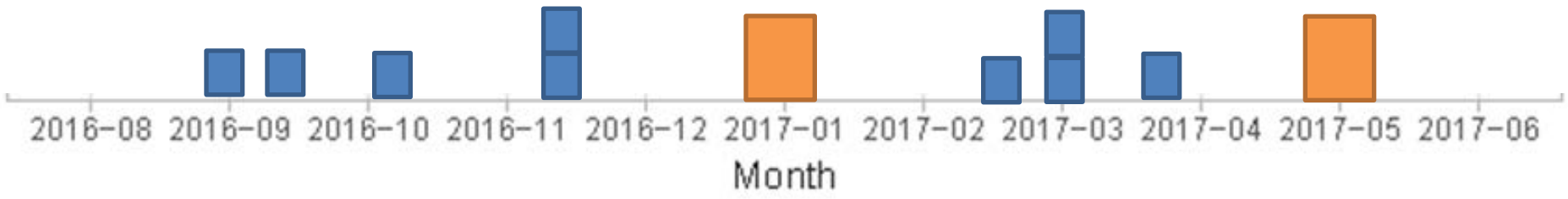


Context, Continued (Grade 4 Math)

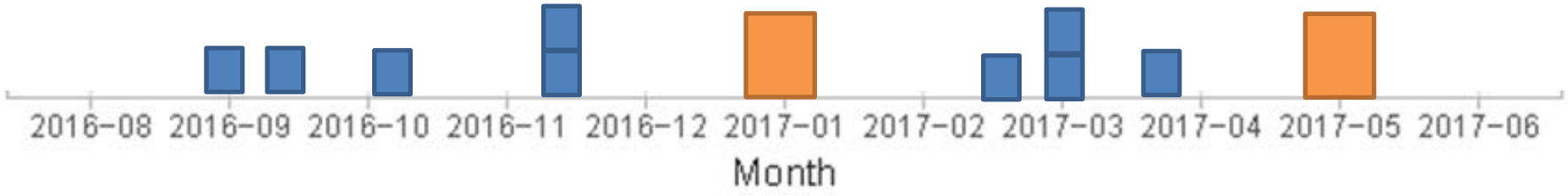
Key to this set of assessments is the idea of **modularity**.

Consider this hypothetical example:

Then the full set of assessment this hypothetical student might look like ↓



Given data like this, how can we make sense of it? In particular, how can we develop scale scores and achievement-level classifications?



Research Questions

1. In what ways can the mini-assessments be scaled?
2. How can provisional mastery classifications be created based on the results of the mini-assessment results?

This work is exploratory and presents a picture of our first efforts to tackle this unique type of assessment in the context of fourth grade mathematics.

Measures

- Assessments of Fourth Grade Mathematics based on the Common Core State Standards
- Two types of on-demand, computer administered assessments:
  - 31 "mini-assessments" aligned to individual standards
  - A "general assessment" of the standards broadly (adaptive and vertically scaled)

Mini-Assessments (31)	General Assessment
Individual standards (e.g., 4.NBT.A.1)	CCSS Fourth Grade Mathematics
Flexibly administered	
Open Access to Items	Secure
Short & Fixed Form (7 Items)	Longer & Adaptive (66 Items Max)
Machine Scored, Instant Reporting	
Non-overlapping (no common items)	Adaptive from the same item pool
--	Scale scores, CCSS domain subscores, & classifications on individual standards

Data

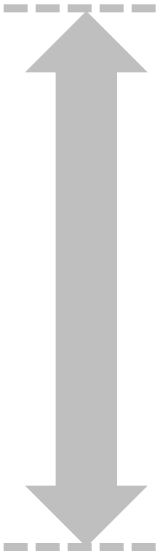
- 2016-2017 academic year
- 91,440 of the students taking at least one mini-assessment & the general assessment
- Mini-Assessments
  - Approximate number of administrations per mini-assessment: ranges from 3,000 to 47,000, mean of 12,000 and a median of 8,000
  - Approximate number of forms per student: ranges from 1 to 80, with a median of 6 and a mean of 7.6 (including re-tests)

RQ1 | Scaling the mini-assessments

One Set of Possible Approaches

Conduct Rasch scaling, place the mini-assessments onto:

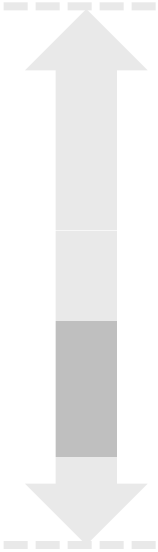
- the scale of the general assessment (via a fixed theta calibration approach).
- a single scale across all mini-assessments.
- CCSS domain specific scales (5 in all).
- individual scales for each mini-assessment.



One Set of Possible Approaches

Conduct Rasch scaling, place the mini-assessments onto:

- the scale of the general assessment (via a fixed theta calibration approach).
- a single scale across all mini-assessments.
- CCSS domain specific scales (5 in all).
- individual scales for each mini-assessment.



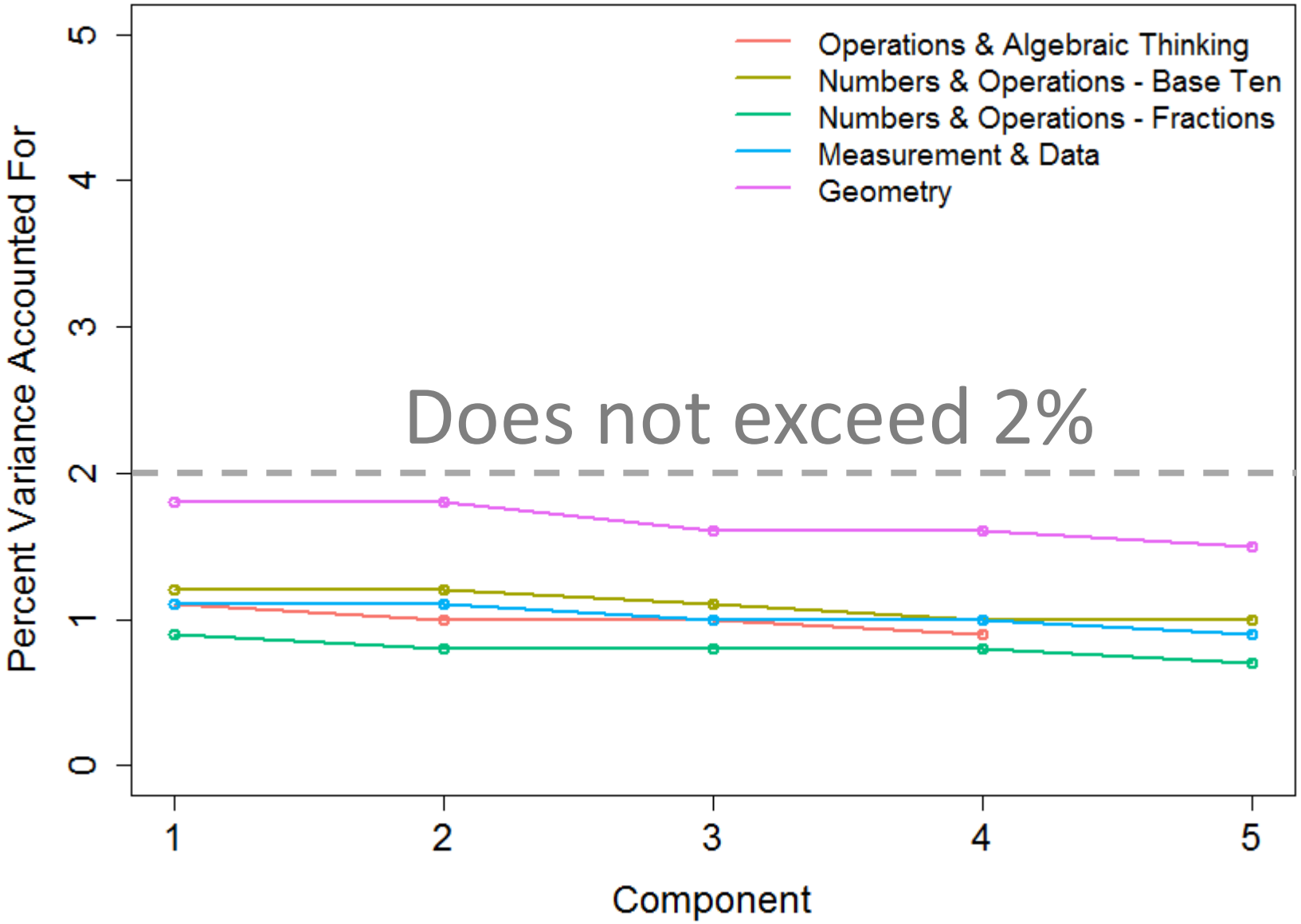
Domain Scaling Approach

- Create unidimensional scales for each CCSS Domain using the Rasch Model
- Use a pooled item response matrix (item responses from different time points and different administration patterns)
  - Best case for detecting multidimensionality

Domain Scaling Approach

- Examine results in terms of:
  - Unidimensionality via Principal Components Analysis of Item Residuals
  - Model Fit (Unweighted and Weighted Mean Squared Fit Statistics)

Results - PCA


Operations & Algebraic Thinking
Numbers & Operations - Base Ten
Numbers & Operations - Fractions
Measurement & Data
Geometry
Does not exceed 2%
Percent Variance Accounted For
0
1
2
3
4
5
1
2
3
4
5
Component

Results – Item Fit (Weighted MS)

	% <0.75	% > 1.33	# Items
Operations & Algebraic Thinking	0%	1%	72
Numbers & Operations - Base Ten	0%	0%	72
Numbers & Operations - Fractions	0%	0%	108
Measurement & Data	0%	2%	84
Geometry	3%	3%	36
Max	3%	3%	

Future Directions

- Additional Dimensionality Investigations
  - EFA
  - DIMTEST & DETECT
  - Comparison Data
- Modeling Approaches
  - Multigroup on time (e.g., month)
  - Selecting data that best matches recommended instructional sequences
  - Other models (e.g., treating the tests as attributes in a "system level DCM"; longitudinal Rasch model)

RQ2 | Creating Classifications

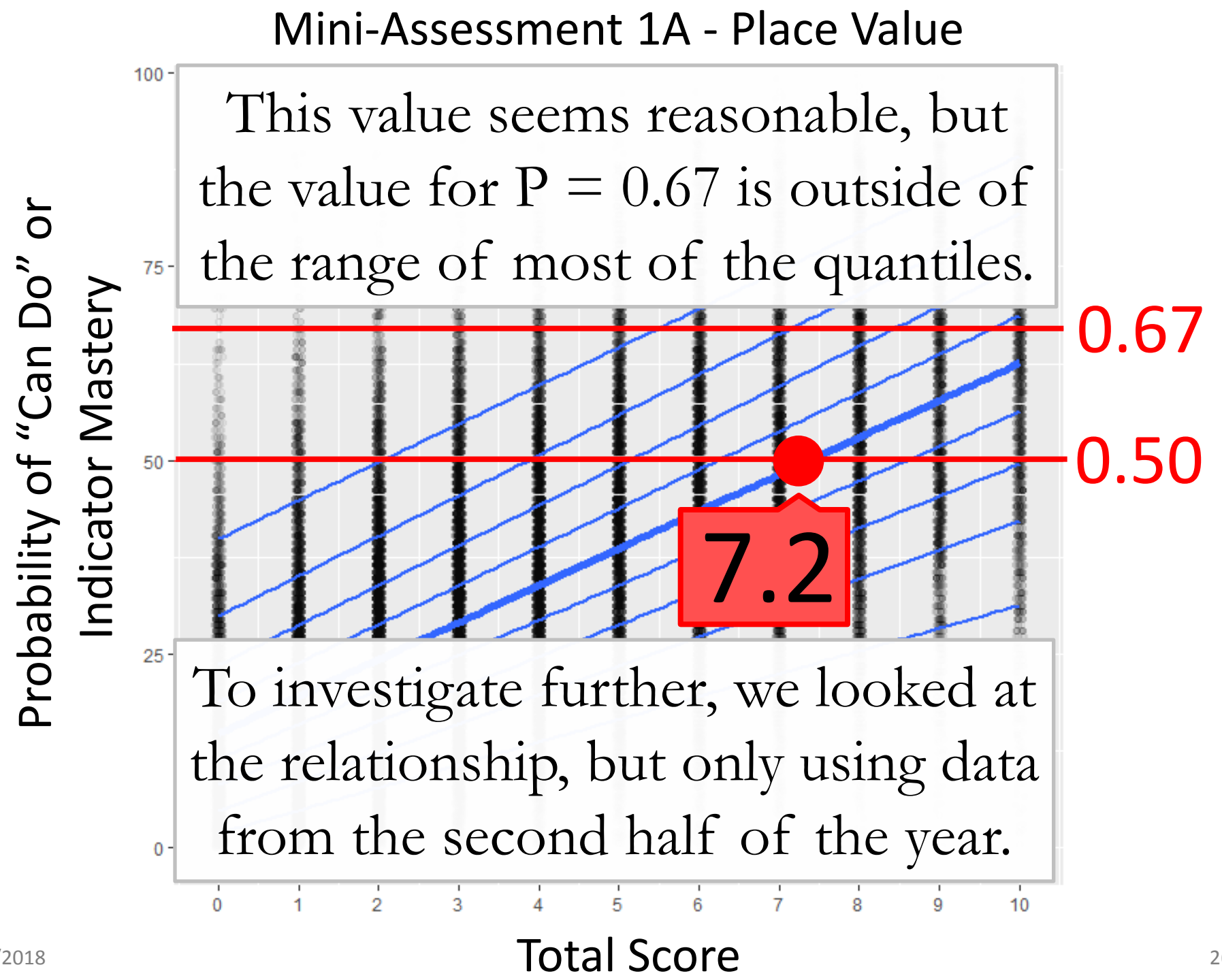
One Set of Possible Approaches

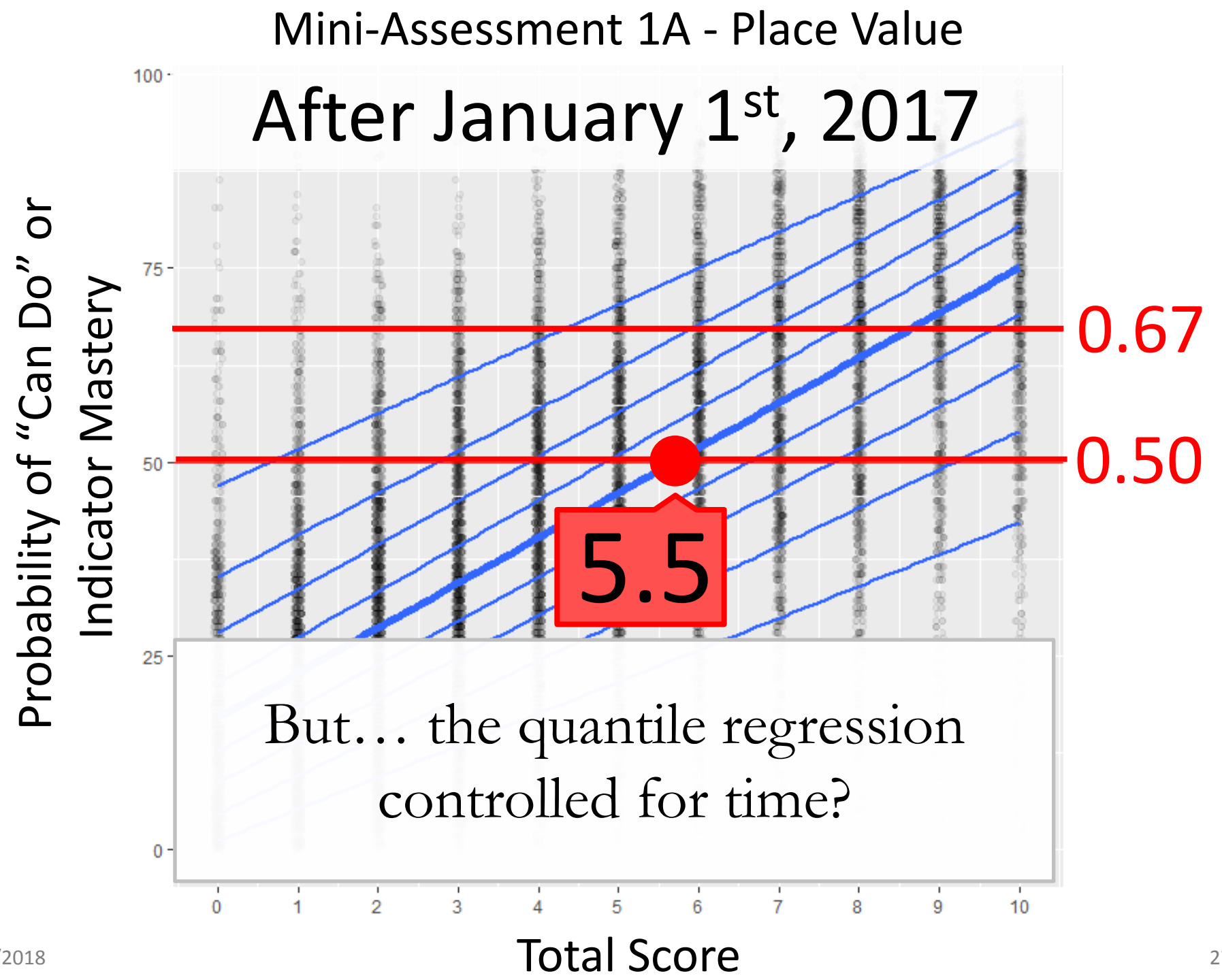
Create Preliminary Cut Scores, and thus Student Classifications based on:

- Cluster analysis (e.g., what DCMs devolve into with one attribute)
- Content Expert Judgments
- The relationship between each mini-assessment and the matching standard classification from the general assessment

The Prediction Approach

- Predict the probability of the "can do" classification from the general assessment using the raw scores from the mini-assessment.
- To do so, conduct quantile regression where
  - The dependent variable is the probability of classification from the closest general assessment to the student's mini-assessment administration
  - The independent variables are the mini-assessment raw score and the different between administrations (in days)
- Evaluate at multiple probabilities & quantiles

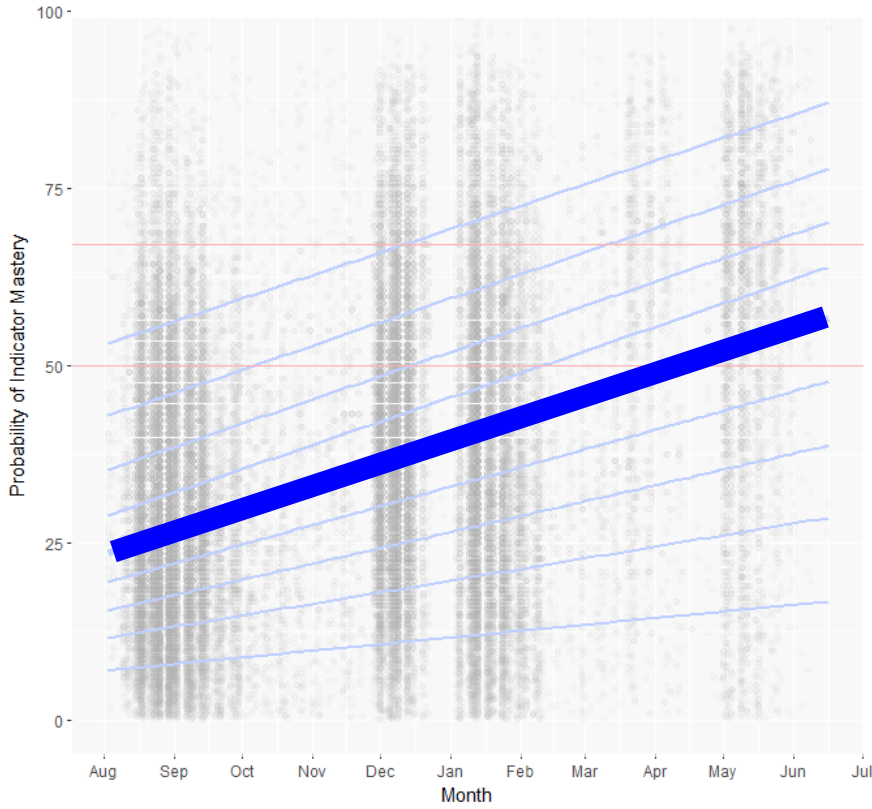

Mini-Assessment 1A - Place Value
This value seems reasonable, but the value for P = 0.67 is outside of the range of most of the quantiles.
0.67
0.50
7.2
To investigate further, we looked at the relationship, but only using data from the second half of the year.
Probability of "Can Do" or Indicator Mastery
100
75
50
25
0
0
1
2
3
4
5
6
7
8
9
10
Total Score


Mini-Assessment 1A - Place Value
After January 1st, 2017
Probability of "Can Do" or Indicator Mastery
100
75
50
25
0
0.67
0.50
5.5
But… the quantile regression controlled for time?
0
1
2
3
4
5
6
7
8
9
10
Total Score

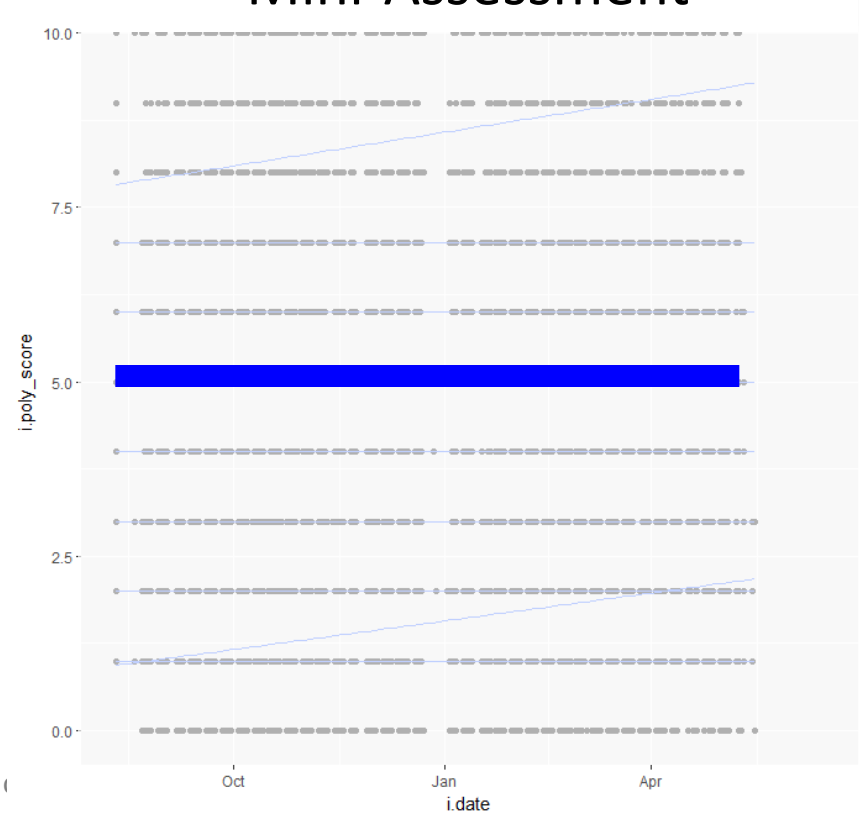
What's going on?

It comes down to the use case for each type of assessment.

General Assessment



Mini-Assessment



Future Directions

Further examine the time issue.

- Re-sample to have equal numbers of administrations by month?
- Look at changes in scores on the mini-assessments?