Evaluating a Large-Scale Educational Assessment

Psychometric Evidence

Questions from Committee Members

- What is test validity? Is SBAC valid, if so why and if not why not?
- What is test reliability? Is SBAC reliable, if so why and if not why not?
- Each year the students are tested on that grade's material or concepts. Can students be compared as to whether they made progress based on that one test since the test the year before was different?
- If a student scores in the below standard category in third grade, for example, isn't it possible that he might still score in the below standard category in fourth grade even if he made a year's worth of growth?

Theoretical Framework for Tests

- Educational assessments are designed to measure students' overall proficiency in domains (e.g., subject areas) of interest
- Scores on the assessments are one piece of information used to make decisions about students, teachers, and schools/districts
- The two most important properties of a test score are validity and reliability

Test Score Validity

- Validity refers to the degree to which **interpretations** of test scores are supported by theory and evidence (AERA/APA/NCME Test Standards)
- Instead of asking, "Is the test valid?", the appropriate question is, "***Is this a valid use or interpretation of the test scores?***"

Test Score Validity

- Validity is not an all-or-none property
- Validity evidence is gathered from a variety of sources to support the use of test scores for a particular purpose
- An argument is made that the test scores can reasonably be interpreted in the intended manner
- Establishing that a particular use of test scores is valid requires clear articulation of the claims that are being made about the scores

Validity Evidence

- The primary types of validity evidence are
  - Evidence based on test content
  - Evidence based on response processes
  - Evidence based on internal structure
  - Evidence based on relations to other variables

Validity Evidence for SBAC

(see detailed technical report)

- Evidence based on test content
  - Alignment studies were performed to show how the content of the assessment matches the Common Core standards
- Evidence based on response processes
  - Think-aloud protocols were used during pilot testing to assess whether items measured the intended cognitive skills

Validity Evidence for SBAC

- Evidence based on internal structure
  - Dimensionality analyses were performed to confirm that the assumption of a unidimensional construct within and across grades is reasonable

Validity Evidence for SBAC

- Evidence based on relations with other variables
  - The correlation between scores on the CMT (2013) and the SBAC operational test (2015) for students in Grades 3 through 6 shows that the SBAC test scores correlate almost as highly with CMT scores as CMT scores two years apart

ELA

Grade	CMT+2	SBAC
CMT G3	0.82	0.78
CMT G4	0.84	0.75
CMT G5	0.83	0.76
CMT G6		0.76

MATH

Grade	CMT+2	SBAC
CMT G3	0.80	0.78
CMT G4	0.85	0.80
CMT G5	0.83	0.81
CMT G6		0.82

Validity Evidence for SBAC

- Other evidence that supports a validity argument:
  - Rigorous and well-documented test construction procedures
  - Adequate measurement precision
  - Appropriate test administration procedures
  - Appropriate scoring procedures
  - Appropriate scaling and equating procedures
  - Appropriate standard setting procedures
  - Adequate investigation of fairness to different subgroups
  - Adequate test security

Measurement Error in Test Scores

- The objective of measurement is to measure what we want to measure (the “true” value) appropriately and with minimum error
- The **Standard Error of Measurement** quantifies the amount of error in a test score

Measurement Error in Test Scores

Standard Error of Measurement (SEM)

- Indicates the amount of error to be expected in using the test score as an estimate of a student's "true" proficiency
- Provides us with an error band for a student's true proficiency

  For example, we can be 95% confident that a student's true score is in the range

  Observed Score $\pm$ 2 SEM

Measurement Error in Test Scores

Standard Error of Measurement (SEM)

- Depends on the score scale, and is therefore difficult to compare across tests that use different score scales
- Can be re-expressed in terms of RELIABILITY, which is between zero and one regardless of the score scale

Test Score Reliability

- The reliability Index, ρ, is defined as the correlation between the observed score and the true score or between scores on parallel forms of a test (i.e., tests that are equivalent in all respects)
- Reliability refers to the test score, not to the test itself
- Reliability of 0 denotes totally unreliable test scores; reliability of 1 denotes perfectly reliable test scores
- We can express SEM in terms of Reliability and vice versa: if test scores are perfectly reliable, $\rho = 1$ and SEM = 0
- Reliability is necessary but not sufficient for validity

Test Score Reliability

1. Administering the same test to everyone does not guarantee score reliability
2. Low performing students will not be able to answer difficult items and for these students their true score will be poorly estimated, i.e., the error of measurement will be large; similarly, true scores for high performing students will be poorly estimated if the items are too easy for them
3. Tests made up of suitable items for each group of students will provide more reliable scores

Problems with Classical Test Framework

- Reliability and hence Standard Error of Measurement are defined in terms of parallel tests, which are almost impossible to realize in practice
- Item statistics (e.g., difficulty) based on classical test theory are group dependent, i.e., they change as the groups change
- Proficiency scores are test dependent: we cannot compare the test scores of individuals who have taken different sets of test items

Problems with Classical Test Framework

It would be better if…

- SEM is not defined in terms of parallel forms of the test, but defined for individual test takers
- Item indices did not depend on the characteristics of the individuals on whom the item data were obtained
- Student proficiency measures did not depend on the characteristics of the items that were administered

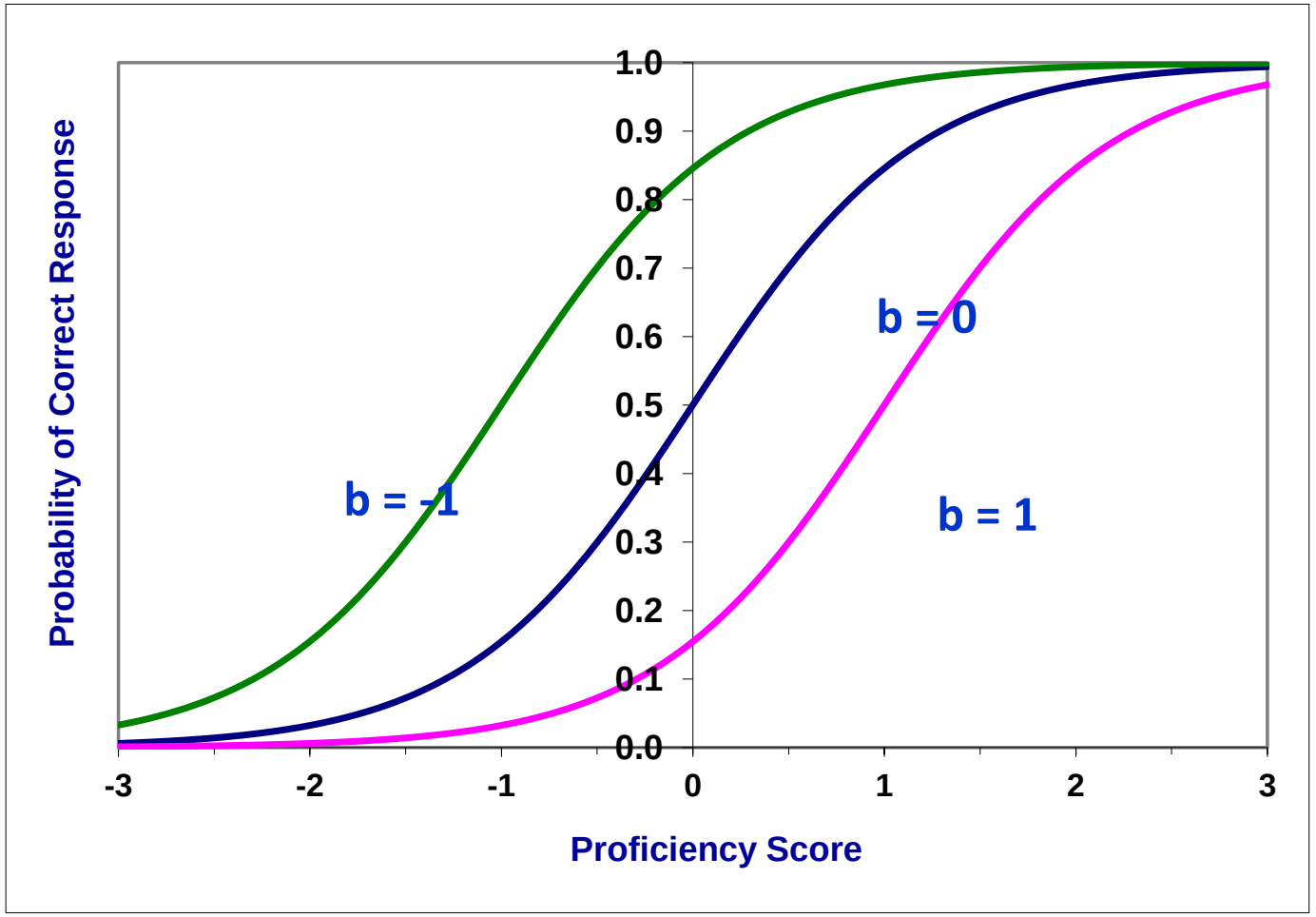
Item Response Theory Framework

- Item response theory is based on the postulate that the probability of a correct response to an item depends on the proficiency value of the student and the characteristics of the item
- An examine with a high proficiency value will have a high probability of answering the item correctly; a student with a low proficiency value has a low probability of answering the item correctly
- The relationship between the probability of a correct response depends on the characteristics of the items and the proficiency value of the student

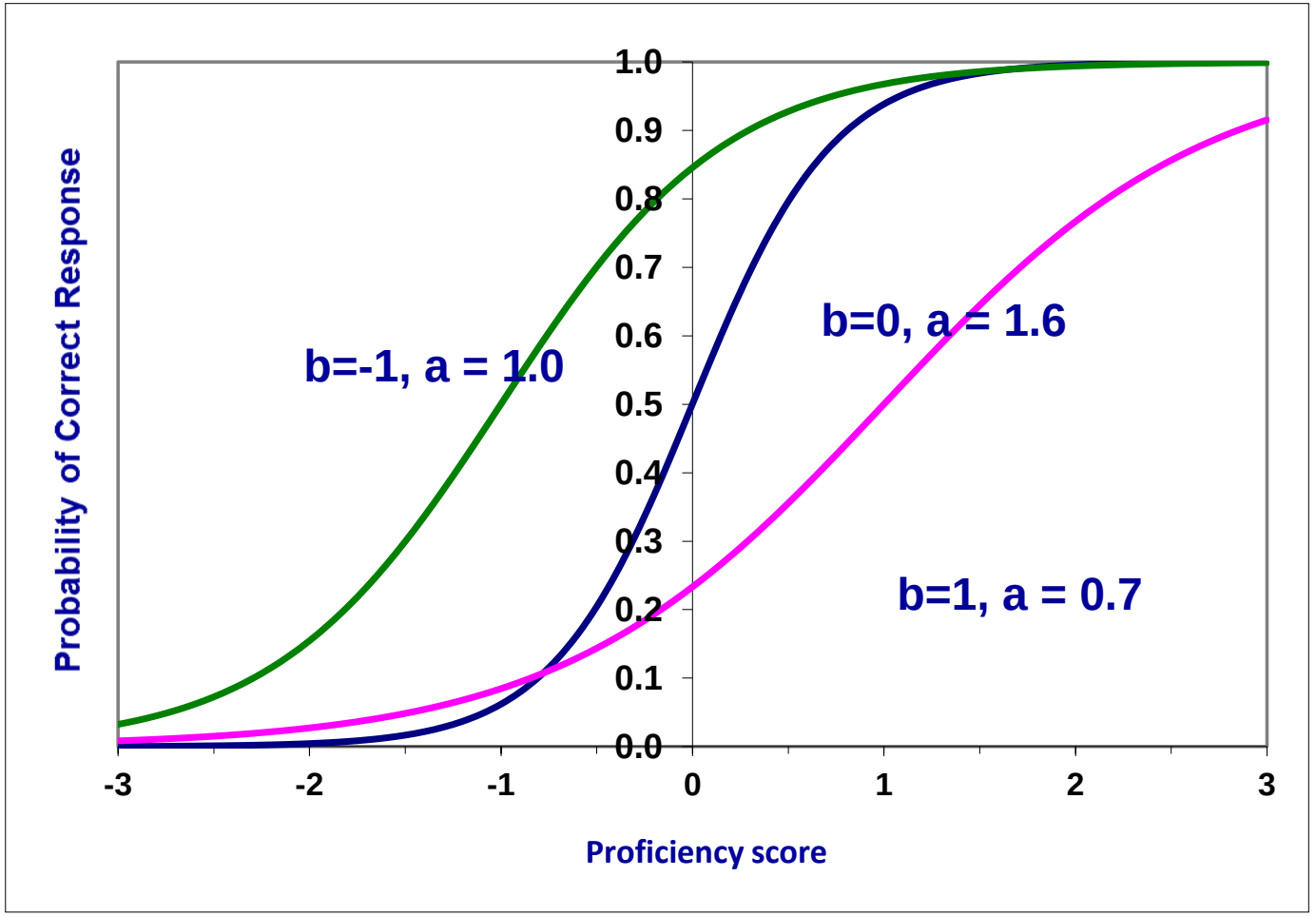
The Item Response Model

- The mathematical relationship between the probability of a response, the proficiency value of the student, and the characteristics of the item is specified by the ITEM RESPONSE MODEL
- Most common models:
  - One-parameter or Rasch model
  - Two-parameter model
  - Three-parameter model

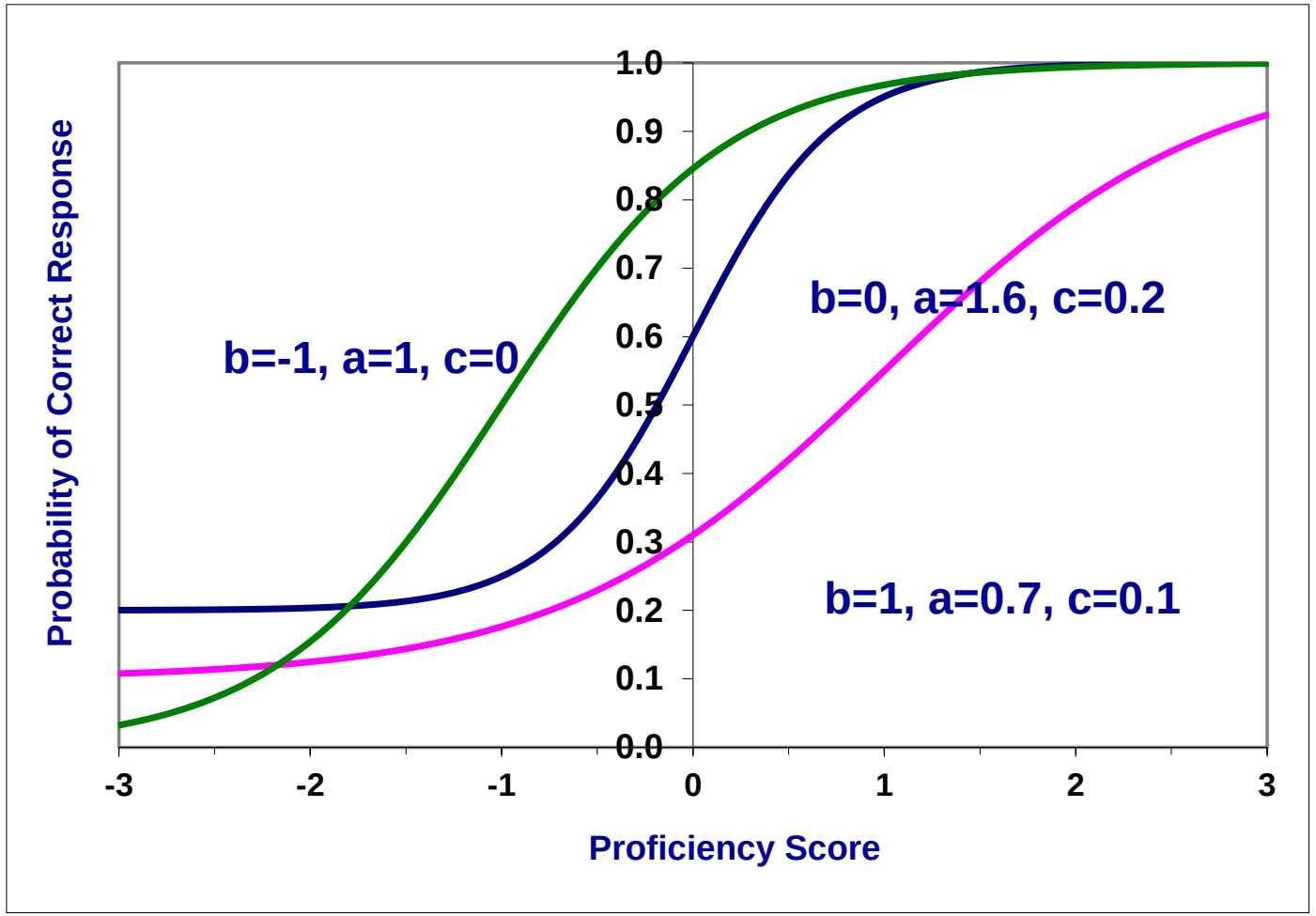
One-Parameter Model Item Response Functions



Two-Parameter Model Item Response Functions



Three-Parameter Model Item Response Functions



Advantages of Item Response Theory

- The proficiency score of a student is not tied to the specific items we administer
- We CAN compare the proficiency scores of students who have taken different sets of test items
- We can therefore match items to a student's proficiency and measure proficiency value more precisely with shorter tests

Advantages of Item Response Theory

- We can create a bank of items by administering different items to different groups of students at different times (Vertical Scales for Growth Assessment)
- This will allow us to administer comparable tests or individually tailored tests to students (Computer Adaptive Testing)
- By administering different items to different individuals or groups we can improve test security and minimize cheating

How Is IRT Used In Practice?

- Test construction
- Equating of test forms
- Vertical scaling (for growth assessment)
- Detection of differential item functioning
- Adaptive testing

Test Construction

- Items can be selected to maximize precision of measurement, i.e., small SEM (high reliability) in desired regions of the proficiency continuum (such as at cut scores)
- By selecting items that have optimal properties, we can create a shorter test that has the same degree of precision as a longer test
- We can tailor (customize) a test that is appropriate for a student
- By tailoring the test and choosing the appropriate items to administer, we minimize testing time and estimate a student's proficiency value efficiently

Computerized Adaptive Testing (CAT)

- Adaptive testing is the process of tailoring the test items to match the best current estimate of a student's proficiency value
- Items are most informative when their difficulty is close to the student's proficiency value
- Different students take different tests
- Only through IRT can items be appropriately selected, proficiency values estimated after each item or a set of items is administered, and the resulting test scores compared

Advantages of CAT

- Testing time can be shortened
- Students' trait values can be estimated with a desired degree of precision
- Scoring and reporting can be immediate
- Scoring errors and loss of data are reduced
- Test security is preserved (in theory)
- Paper use is eliminated
- Need for supervision is reduced

Reliability Evidence for SBAC Proficiency Scores

- Reliability of SBAC test scores compared with CMT scores

ELA

Grade	CMT Test Length (# of items)	CMT reliability	SBAC Test Length (# of items)	SBAC reliability
3	73	0.94	42-46	0.92
4	74	0.93		0.92
5	80	0.93		0.92
6	80	0.94		0.91
7	79	0.94		0.92
8	79	0.95		0.92

Reliability Evidence for SBAC Proficiency Scores

- Reliability of SBAC test scores compared with CMT scores

MATH

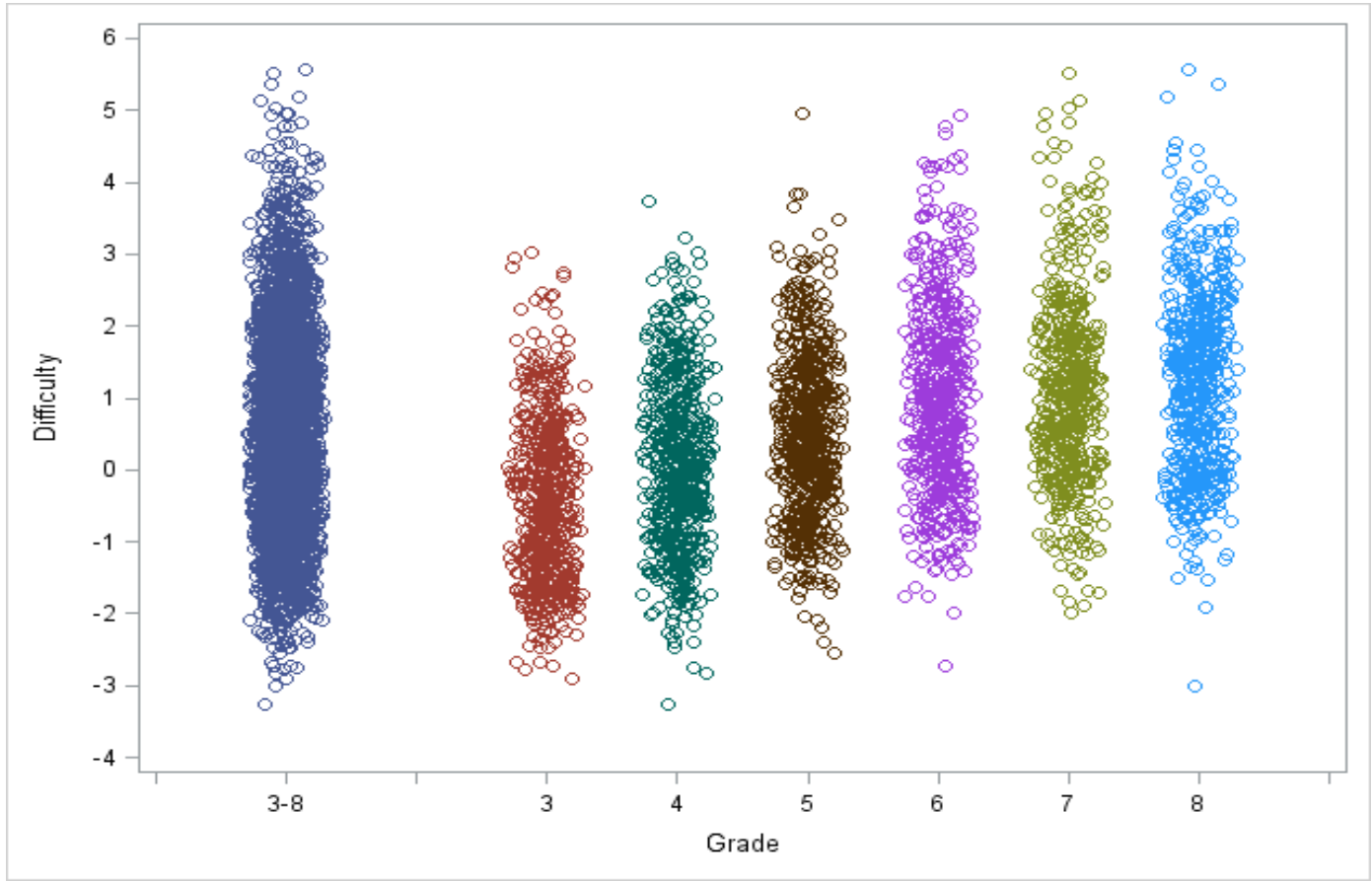
Grade	CMT Test Length (# of items)	CMT reliability	SBAC Test Length (# of items)	SBAC reliability
3	94	0.94	37-40	0.94
4	96	0.95		0.94
5	113	0.96		0.93
6	116	0.97		0.93
7	120	0.97		0.91
8	117	0.97		0.92

Growth Assessment and Vertical Scales

- In developing a vertical scale, sets of common items are administered to students in adjacent grades
- Through these common items, items in adjacent grades are placed on a common proficiency scale using IRT methods
- Items are designed so that there is sufficient overlap of items across grades

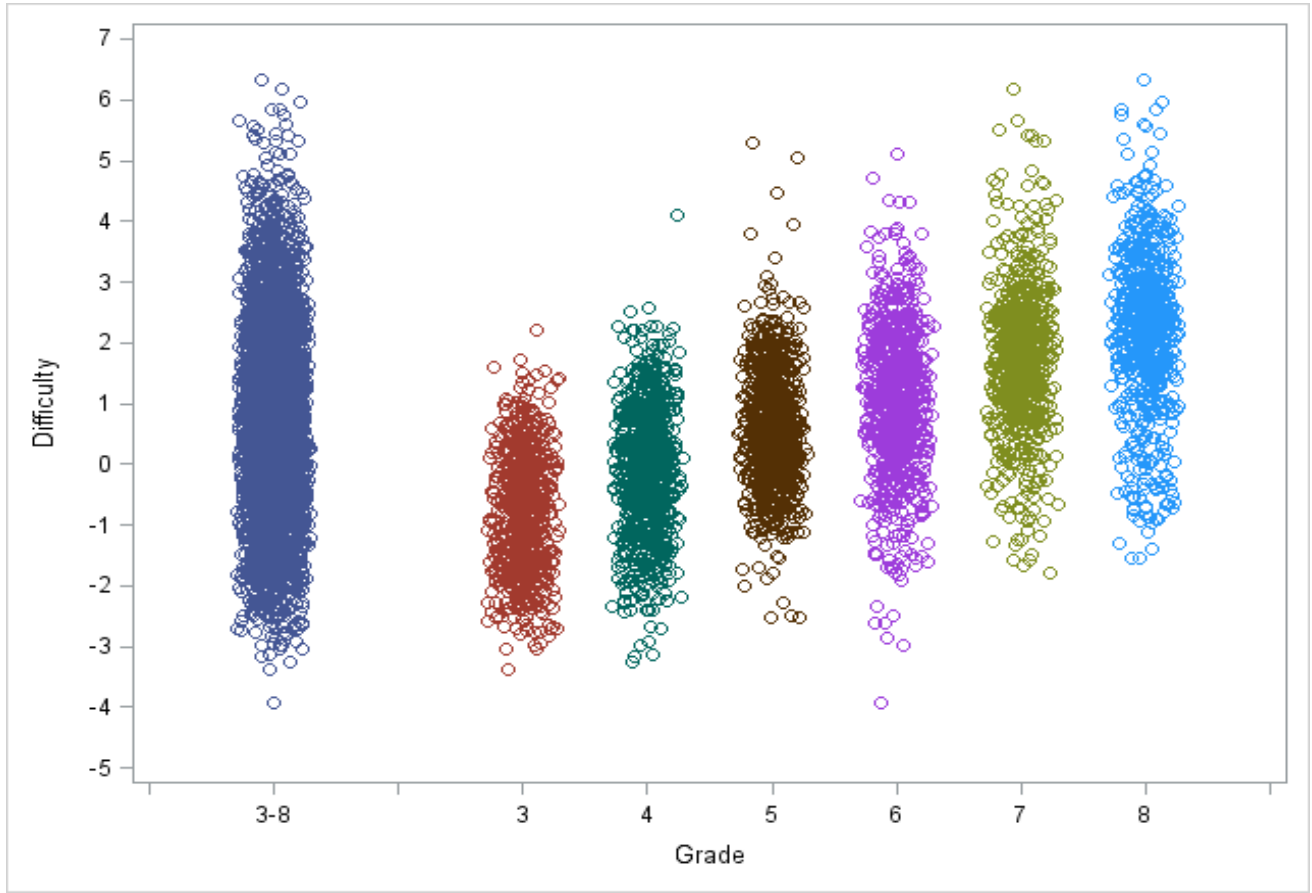
Growth Assessment and Vertical Scales

- Plot of SBAC item difficulty across grades: ELA



Growth Assessment and Vertical Scales

- Plot of SBAC item difficulty across grades: MATH



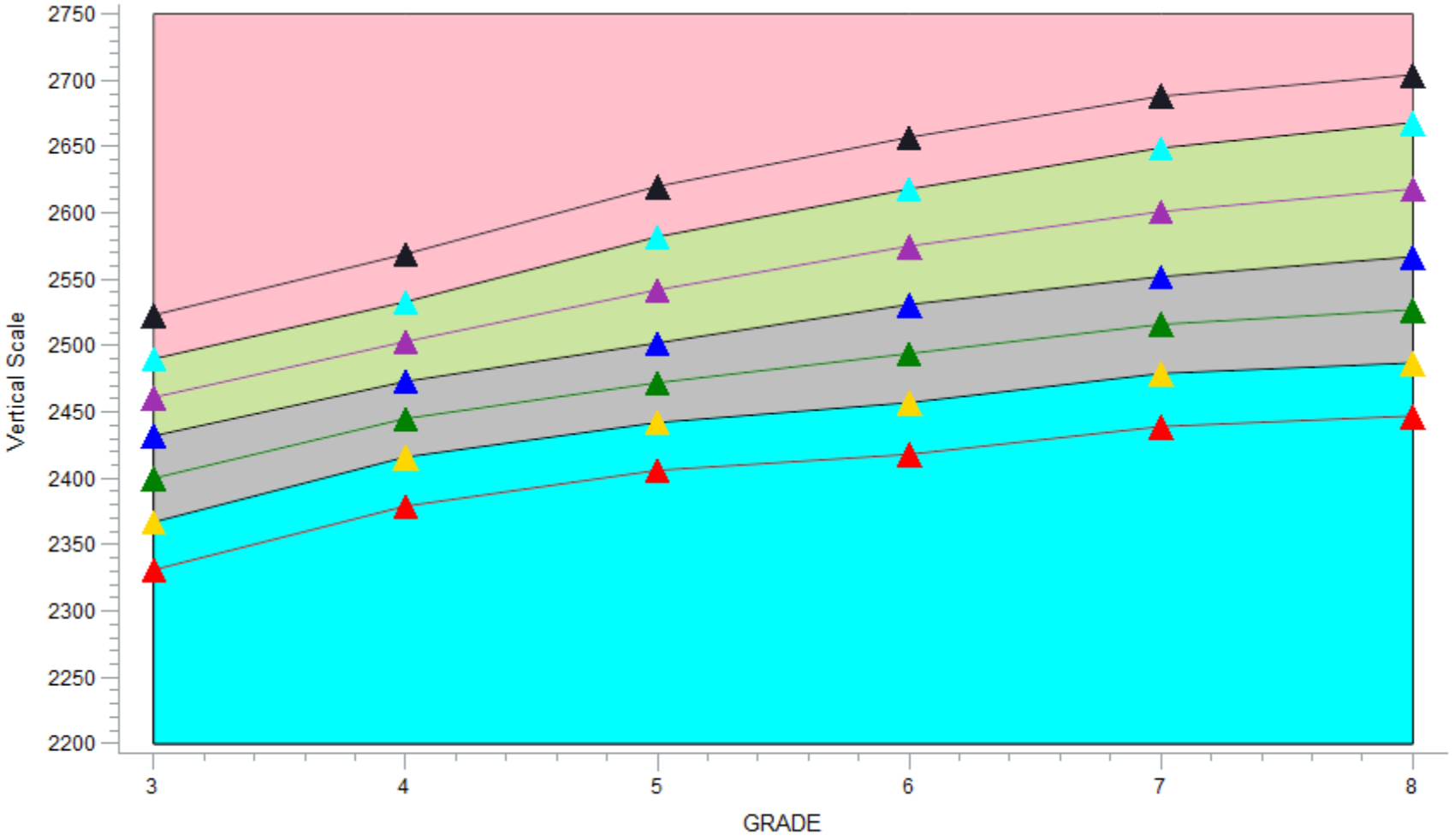
Growth Assessment and Vertical Proficiency Scale

- The proficiency score obtained is a general measure of proficiency
- Growth in proficiency can be assessed using the items that are on a common scale across grades
- The same items need not be administered since the IRT-based proficiency scores do not depend on the items administered
- The same procedure was used with the CMT vertical scale for assessing growth; the only difference is that the CMT was not adaptively administered

Growth Assessment and Achievement Level Categories

- It is possible that a student has grown, but remains in the same Achievement Level Category from one grade to the next
- This is one of the reasons states wanted a growth scale that is more sensitive to student growth
- The vertical scale provides a measure of growth on the IRT proficiency scale

Vertical Scale Growth Model: ELA



Summary

The test development procedures used by SBAC provide sufficient evidence that

- the test scores can validly be used as measures of overall proficiency
- the proficiency scores obtained through the CAT administration are reliable
- the IRT proficiency scores based on the vertical scale can be used to measure growth in proficiency across grades