

Assessment 101

Kevin Cappaert – Pearson

kevin.cappaert@pearson.com

Yu-Feng Chang - Minnesota Department of Education

Yu-Feng.Chang@state.mn.us

Formative and Summative Assessment

- Two types of assessment are often discussed
 - Formative assessment (assessment FOR learning) – the process of *gathering evidence* of student learning, *providing feedback* to students, and *adjusting instructional strategies* to enhance achievement (McMillan, 2014).
 - During instruction
 - Example: End of topic quiz, thumbs-up & thumbs-down, or even non-verbal cues
 - Summative assessment (assessment OF learning) – a way to document (at the conclusion of a period of instruction) what students know, understand, and can do which results in adjustments to future learning and decisions
 - Can be used to evaluate the effectiveness of educational programs
 - Have students learned what they were expected to learn?
 - Example 1: In-class end of unit test
 - Example 2: Accountability test such as MCA

Formative and Summative Assessment (cont.)

- Role of Assessment

- Formative -> Immediate in order to provide instructional changes to improve student learning
- Summative -> To measure proficiency following a given instructional period by comparing against a standard or benchmark to evaluate the effectiveness of educational programs

- Level of Use

- Formative -> Specific, micro level, used for the individual student
- Summative -> General, macro level, used for a group of students

- Structure

- Formative -> Flexible, ever changing, informal
- Summative -> Rigid, structured, formal
 - The Minnesota Accountability tests are standardized meaning all students are assessed on the same content, in the same manner, and the scoring is conducted in a standard way.

Theta, Ability, Scale Score, & Proficiency Level

- **Ability:** In item response theory (IRT) ability refers to the amount of the variable being measured
 - Example: What degree of knowledge in third grade mathematics does a student have?
- **Theta:** The actual estimate of ability in IRT
 - Theta range for Minnesota Assessments (-3 to 3)
 - Ability and theta can be thought of as interchangeable terms
- **Scale Score:** The theta/ability estimate is transformed into the scale score via transformation
 - For MCA III scale scores range from 01-99 with the grade as a prefix
 - (e.g. 301-399 for 3rd grade)
- **Proficiency Level:** Defines the level of student achievement
 - Does not meet standards, Partially Meets Standards, Meets Standards, Exceeds Standards
 - Interpreted based on student scale scores
 - Standard settings are conducted to define the performance levels

Validity and Reliability/Precision

- **Validity** – the degree to which evidence and theory support the interpretations of test scores for proposed uses of the tests (AERA, APA, & NCME; 2014).
 - Concerned with the accuracy of measurement
 - Does a test measure what it purports to measure
- **Reliability** – Refers to the consistency, stability, and dependability of the scores and inferences made by those scores.
 - A necessary but not sufficient condition for validity
 - In assessment, evidenced through error

Validity Evidence

- Test Content
 - Does the test adequate sample the domain it purports to cover?
 - Alignment of assessment content to standards
- Item Functioning Equivalency by Subgroup
 - Bias and Sensitivity Review conducted by Minnesota educators
- Predictive Validity
 - Career and College Readiness linked to nationally recognized college entrance exam
- Validity is not absolute, rather the degree of validity is of concern

Reliability

- Consistency of testing scores over time
- Observed Score = True Score + Error
 - The lower the error in measurement the greater the consistency (reliability)
- In IRT reliability is directly related to the item information and number of items on a given assessment
- Reliability is not absolute, rather the degree of reliability is of concern

Question

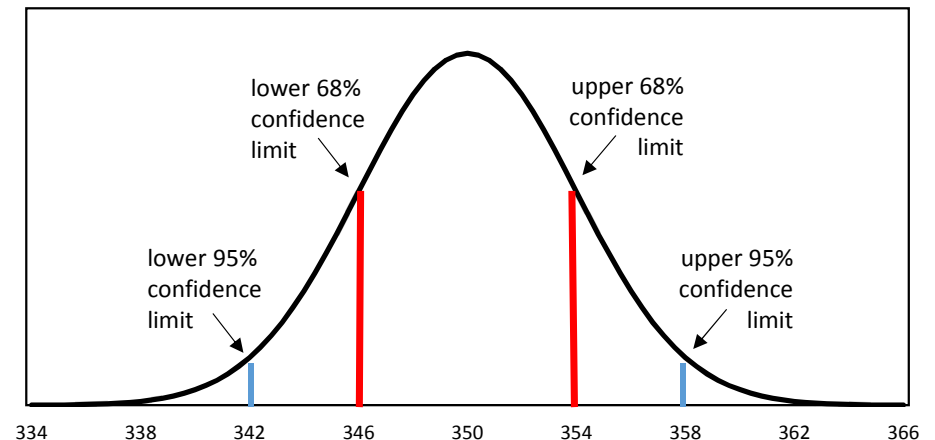
- Consider a student who received a 821 on the first OLPA and a 819 on the second. His strand scores on the first OLPA were Number & Operation=1, Algebra=2, Geometry & Measurement=1, and Data Analysis=2. The strand scores for his second test were 2,3,2,3, respectively.
- Why did the scale score decrease but the strand scores increase?

Standard Error of Measurement (SEM)

- Because student's scale score estimates may vary (differ) across testing instances, a single test may not produce an exact estimate of one's true proficiency.
- The standard error of measurement is able to represent a lack of score consistency for the population of students.
 - The standard error of measurement (SEM) expresses score inconsistency (unreliability) in terms of the reported score metric. Because Minnesota students are only tested at one point during the testing window each academic year, it is not possible to estimate the standard error through multiple measures.
- SEM is used to quantify the precision of a test in the metric on which scores will be reported.
- The SEM can be helpful for quantifying the extent of errors occurring on a test. A standard error of measurement band placed around the student's scale score would result in a range of values most likely to contain a student's observed score upon replication.

Interpreting the Standard Error of Measurement

- Take a student scale score of 350 on a given assessment. If the standard error of measurement is 4 then confidence bands can be put around that value.
- A 68% confidence band can be placed by adding and subtracting the SEM to the student estimate.
 - $350 \pm 4 = [346, 354]$
- A 95% confidence band (approximately) can be computed by adding and subtracting 2 times the SEM to the student estimate.
 - $350 \pm (2 \times 4) = [342, 358]$
- With 95% confidence, the student's true scale score will fall between 342, and 358.
 - There is still some uncertainty (in this case 5%) that the student has a true score outside of that range.



Item Response Theory(IRT)

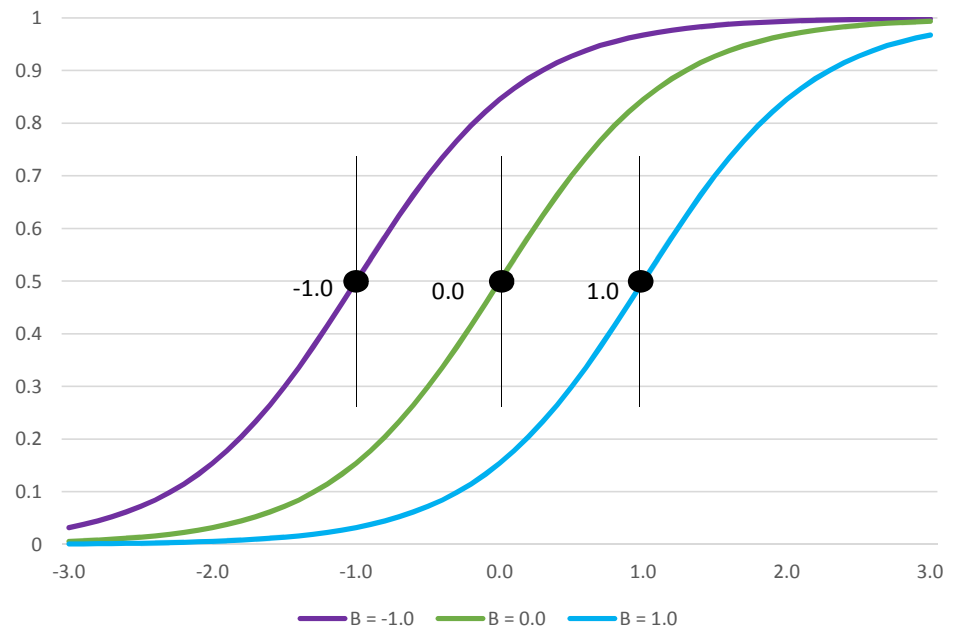
- Two common approaches to measurement
 - Classical Test Theory (CTT)
 - Item Response Theory (IRT)
- IRT - The probability of a correct response to item i depends on both the amount of the person's trait ability level (θ) and a set of item parameters (in this case A, B, C).
- IRT Benefits
 - The error of measurement in IRT depends on the true ability of the student and the response pattern of the items administered.
 - Shorter tests can be more reliable than longer tests
 - Unbiased estimates of item properties may be obtained even if the sample is less than optimal
 - Person trait ability estimates can be applied to the context of items to apply meaning
 - A single trait ability can be estimated efficiently with mixed item formats

Item Characteristic Curve (ICC)

The Difficulty Parameter

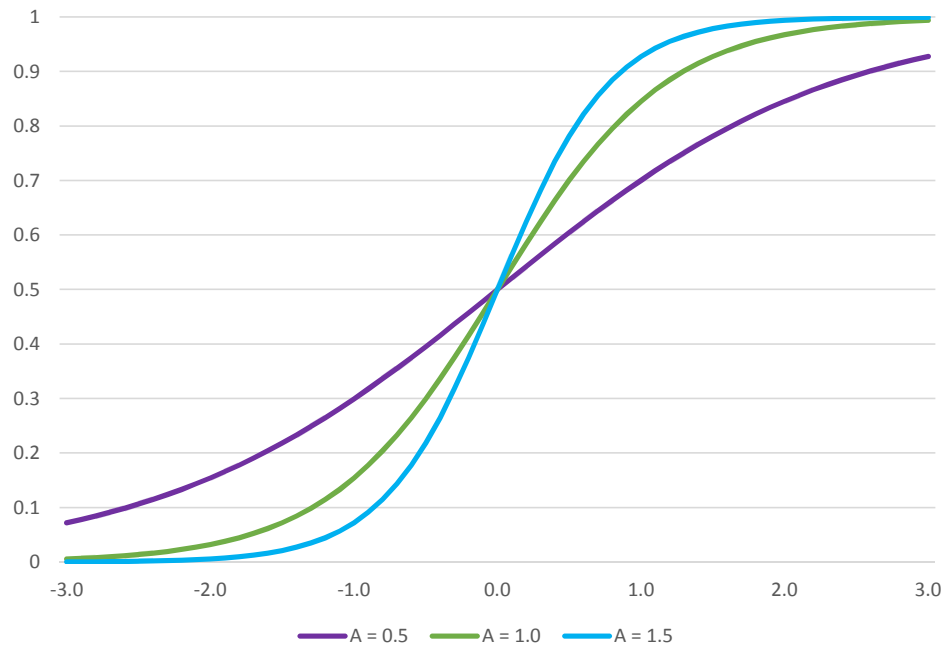
- IRT models the probability of a correct response which depends on the ability level of the student, from -3.0 (low ability) to 3.0 (high ability) in the subjects tested (in the MN tests), and a set of item parameters.
- In its most basic form, IRT uses the difficulty parameter (B) to estimate student ability.
- The probability of a correct response to item i is expected to increase as the latent trait (θ , or ability) increases.
- The item difficulty parameter in the 1 Parameter Logistic Model in IRT occurs where students have a 50% chance of getting a correct answer.
 - If a student has greater than a 50% chance to get the item correct then they will, more times than not, get the item correct if presented with an item measuring at that position in the theta distribution.
 - If a student has less than a 50% chance to get the item correct then they will, more times than not, get the item incorrect if presented with an item measuring at that position in the theta distribution.
 - If we were to know the exact person ability then an item with a difficulty parameter equal to that ability would result in equal chances to get the item correct or incorrect.

1 Parameter Logistic Model (1PL)



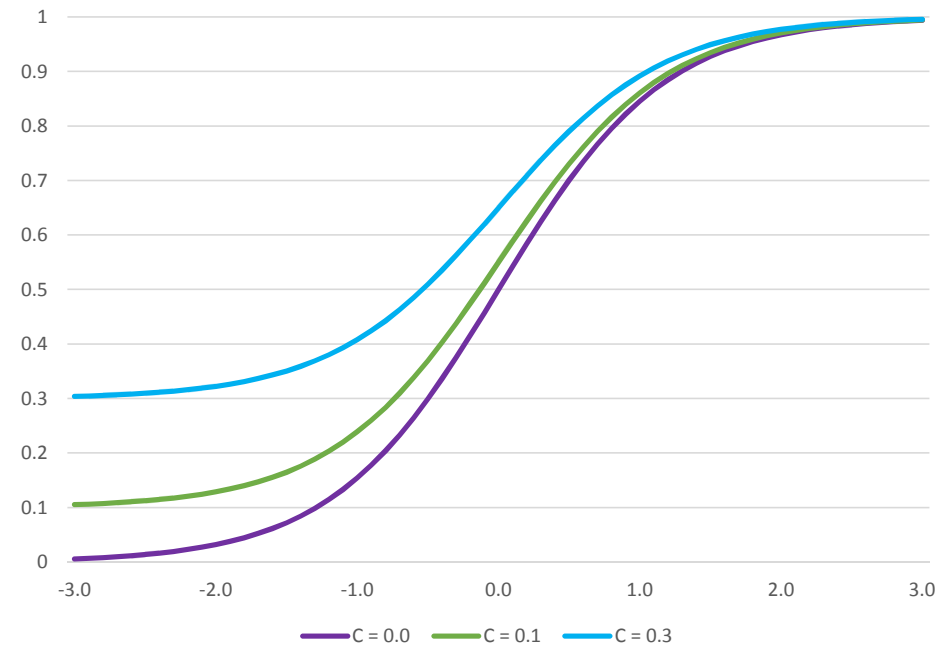
Item Characteristic Curve (ICC)

2 Parameter Logistic Model (2PL) The Discrimination Parameter



Can be thought of as the weight of the item

3 Parameter Logistic Model (3PL) The Pseudo-Guessing Parameter



Pattern Scoring

Discrimination	Difficulty	Pseudo-Guessing
1.2	-1	0.2
2.0	0	0.2
0.7	2	0.2
1.0	-1	0.2
1.4	1	0.2

The above represents a short test containing 5 items of varying item discrimination and difficulties

Response Pattern	Theta Estimate	Scale Score
11100	0.0	350
11010	0.4	356
10011	-0.3	346
10101	-1.1	334

Where 1=correct and 0=incorrect, it can be seen that even though all 4 above students answered 3 of the 5 items correctly, their respective scale scores are different.

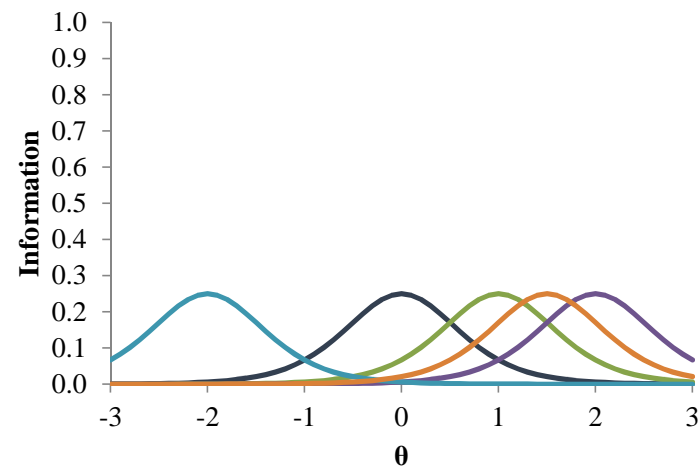
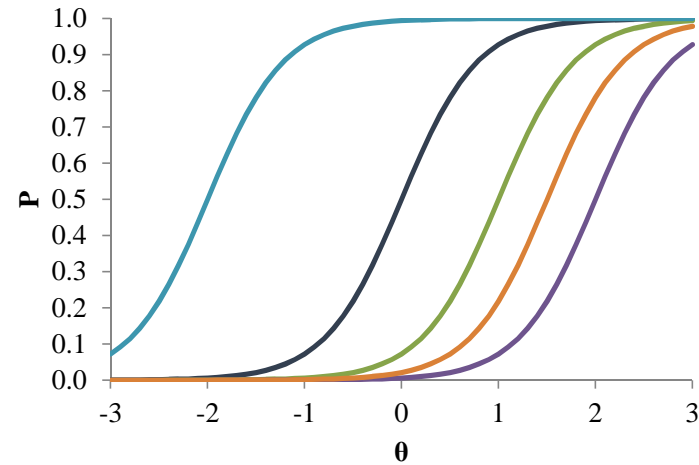
This occurs because the theta estimate depends on the response pattern of the student. Because the item parameters have varying discrimination and difficulty parameters, the resulting ability estimates differ.

The resulting scale score estimate for each student is simply a transformation of the ability estimate.

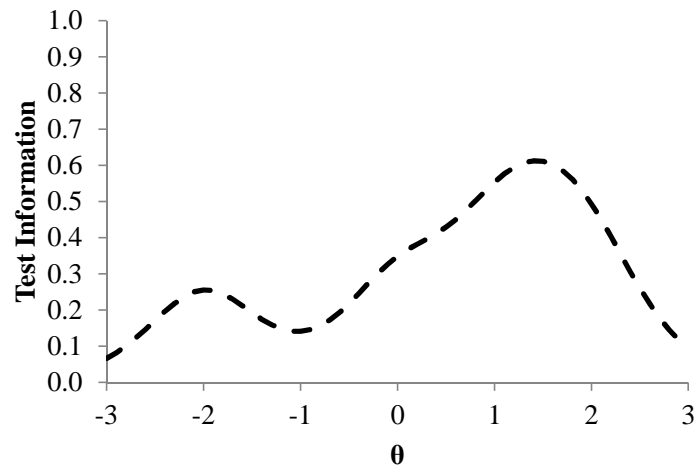
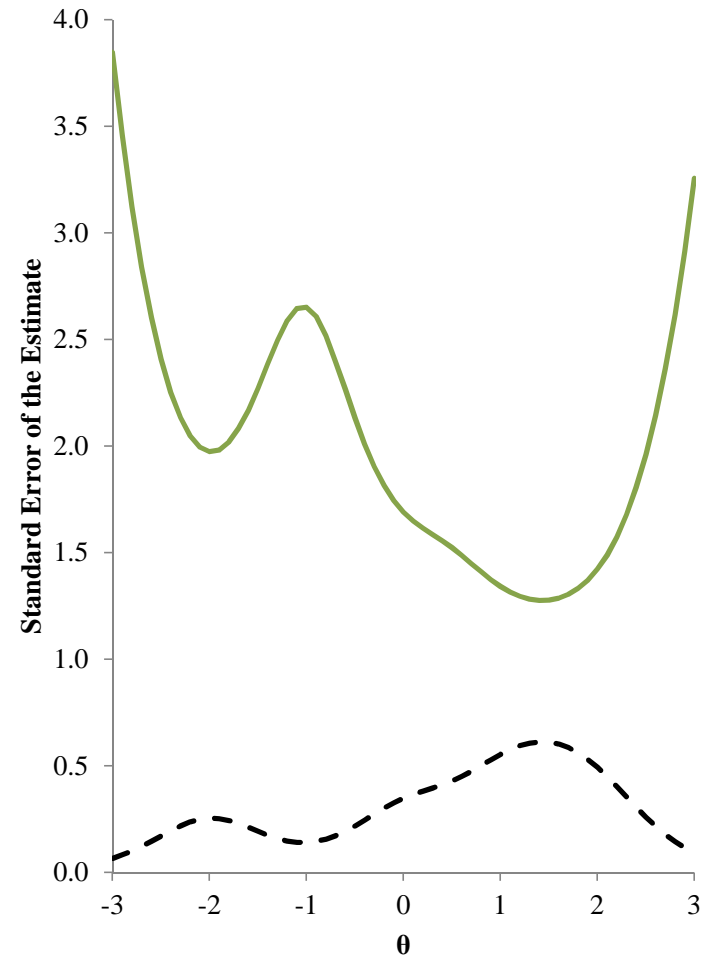
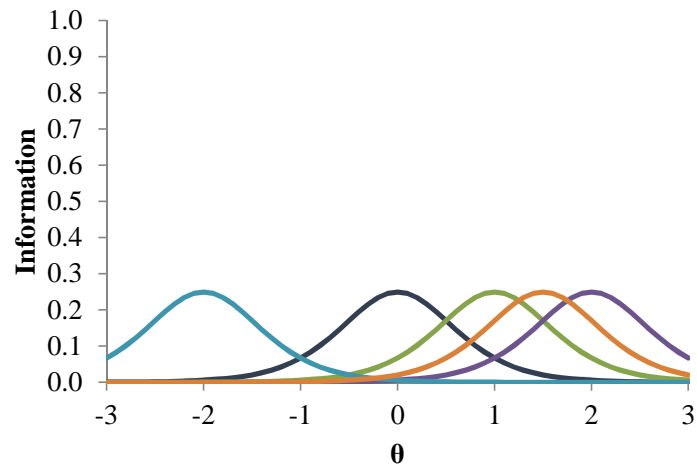
Test Information

- For a set of items
 - $A = 1.5$ (for all items)
 - $C = 0$ (for all items)

- $B_1 = -2$
- $B_2 = 0$
- $B_3 = 1$
- $B_4 = 1.5$
- $B_5 = 2$

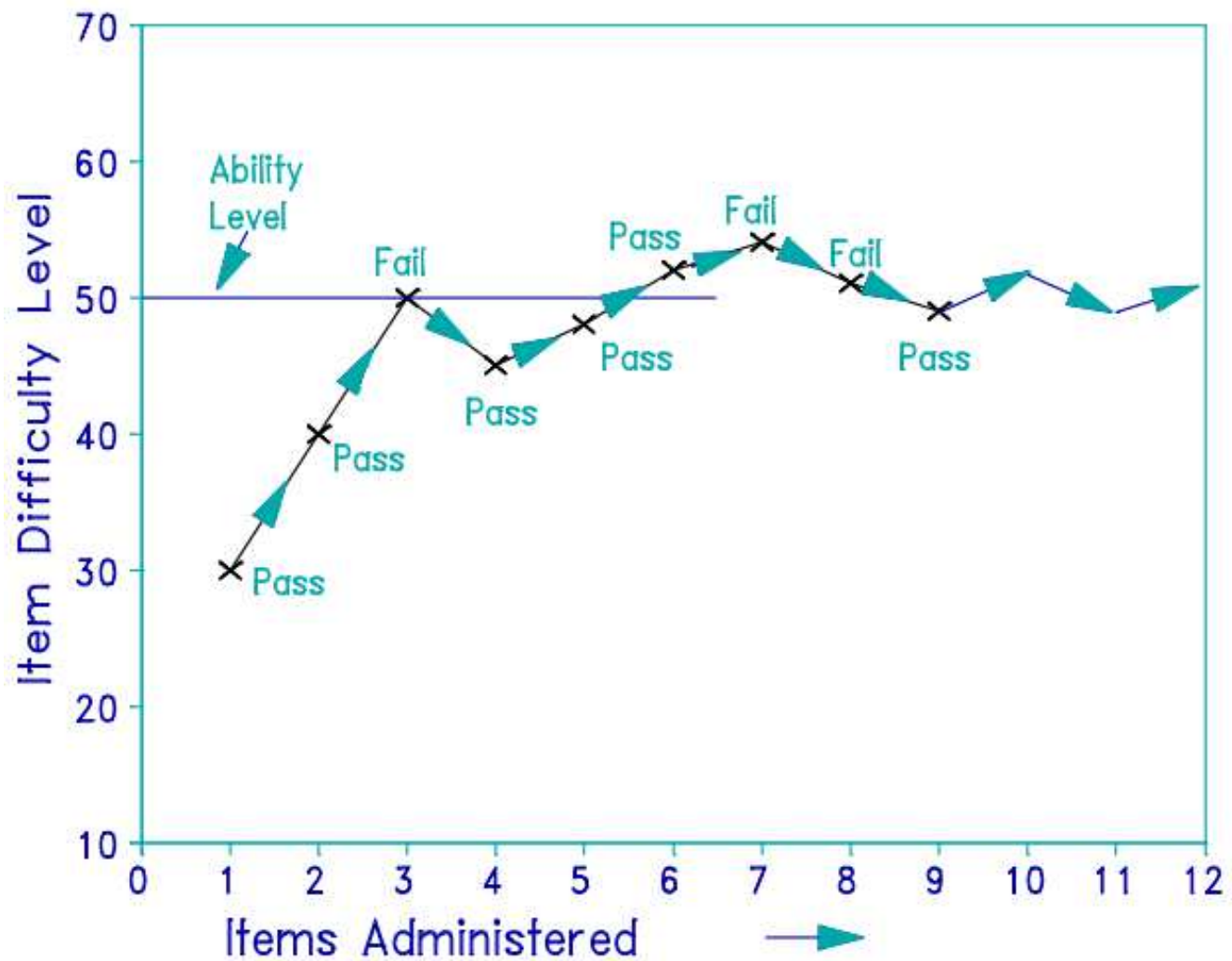


Standard Error of Measurement in IRT



Computer Adaptive Testing (CAT)

Linear (Fixed-Form) Testing	Adaptive
Each item is chosen prior to administration. Students receive the same items or equivalent forms	Each item is iteratively chosen based on prior performance on the assessment
Uniformly designed to measure across the ability distribution	Adaptive to the student taking the test. Tailored to each student's estimated theta
Requires longer tests because items need to measure across the entire ability distribution	Test length can often be reduced because items are more efficiently tailored to the student near a given theta location
Items on each form are chosen to meet test specifications	The CAT algorithm iteratively selects items by weighting the item information in relation to one's estimate of ability as well as strict content constraints to ensure each test meets test specifications
Less secure because all students see the same items. Cheating is easier	More secure because a CAT test requires a very large item bank of items. Cheating is much more difficult



From Linacre (2000)

Common Questions

Question 1 (General):

How does MDE administrate Minnesota Comprehensive Assessment (MCA) for 2015-2016 and 2016-2017 and Optional Local Purpose Assessment (OLPA) for 2015-2016 and 2016-2017?

Grade and Test	Computerized Adaptive Testing (CAT)	Linear forms	May include off-grade items
G3-8 Reading MCA	✓		✓
G10 Reading MCA	✓		
G3-8 Math MCA	✓		✓
G11 Math MCA	✓		
G5 & G8 & HS Science MCA		✓	
G3-8 & G10 Reading OLPA		✓	
G11 Math OLPA		✓	
G3-8 Math OLPA	✓		

Question 2.1 (CAT):

What is the difference between the Reading MCA CAT and the Math MCA CAT?

- An adaptive test constructs a test form unique to each student which is targeted to the student's level of ability.
- Each item administered in the adaptive tests for MCA III Math Grades 3-8 & 11 are administered according to the students' responses and item parameters of all of the previous questions the student has been administered.
- The sets of items (passage or passages) administered in the adaptive tests for MCA III Reading Grades 3-8 & 10 are administered according to the student's responses to the previous passage(s). This form of testing is referred to as computerized multistage testing.

Question 2.2 (CAT):

If a student answers the first X number of questions wrong for MCA III Math, is it then impossible to get Meets or Exceeds the Standards?

- No, the items are administered according to students' responses to the previous questions for the adaptive tests for MCA III Math . In general, if an item is answered incorrectly, an easier item is expected next and if an item is answered correctly, a more difficult item is expected.
- The CAT algorithm does not limit student's score range based on their performance in the first few items. Rather, it continues to be adaptive until the very end of the test. In addition, at the conclusion of the test, the scoring system will compute a final ability estimation based on the students' responses to the items on the entire test.

Question 3.1 (Off-grade):

What grade level items are given for the off-grade items?

- Any off-grade level items will be no more than two grade levels above or below a student's grade.
 - EX: Fifth grade students may see items from any of grades 3, 4, 5, 6, 7.
- The exception to this are grades 3 and 8 because there are no items below grade 3 and no items above grade 8.
 - EX: Third grade students may see on-grade items or above-grade items (4 or 5 grade items) and eighth grade students may see on-grade items or below-grade items (6 or 7 grade items).

Question 3.2 (Off-grade):

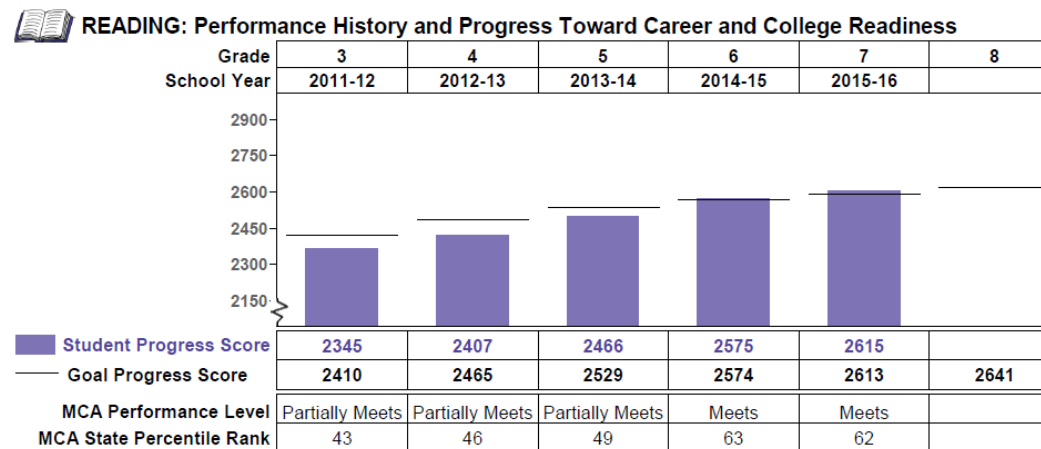
Which off-grade items are used for scoring?

- Only on-grade items will be used for AYP and MMR accountability calculations as laid out in the federal guidelines.
- Both the on-grade and off-grade items for G3-8 Reading & Math MCA will be used to calculate the student progress score.

Question 3.3 (Off-grade):

How do I interpret the progress score results for MCA Reading & Math G3-8?

- A student whose Progress Score is at or above the Goal Progress Score is considered on track to demonstrate career and college readiness at the end of Grade 11. Put another way this means that student scores indicate whether students are on track to pass a college entrance exam by the end of grade 11 in Reading and Mathematics.
 - If the progress score is at or above the goal progress scores at each grade, it is projected to be on track for next grade's coursework.
 - If the progress score is below the goal progress scores at each grade, it is not projected to be on track for next grade's coursework.



Question 4.1 (Scale scores):

How the items are "weighted"? How is the scale score determined?

- Each item has item parameters: the difficulty parameter, the discrimination parameter, and the guess parameter.
- The pattern scoring method is used to estimate scores - the item parameters and the student's answer pattern are used to calculate the theta score. The theta score with a mean=0, standard deviation=1, minimum=-3, maximum=+3, as well as the standard error of measurement (SEM) for the theta are calculated.
 - Since CAT tests contain a relatively unique combination of items for each student, the items and response pattern are unique to the student which most often result in differing theta and SEM estimates for students who answer the same number of items correctly.
- After the theta score is calculated, it is then transformed to the scale score to be used for reporting purposes.

Question 4.2 (Scale scores):

How is the scoring method for CAT different from previous linear forms?



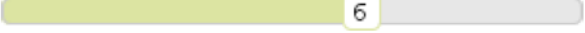

- The scoring method is the same for the computer adaptive testing (CAT) and linear forms.
- The pattern scoring method is used for:
 - MCA III Reading and Mathematics adaptive forms
 - OLPA Mathematics adaptive forms for grades 3-8
 - OLPA Mathematics linear form for grade 11
 - OLPA Reading linear forms

Question 5.1 (Strand scores):

How are the strand/substrand scale scores calculated?

- The pattern scoring method is used to calculate the strand level theta score and SEM of strand level theta score.
- The strand level theta score and SEM of strand level theta score are then transformed to a strand scale score with range 1-9 and the SEM of the strand scale score=1.

Performance Details

	Strand or Sub-Strand Scale Score	Max Score
1. Nature of Science and Engineering	 5	9
2. Physical Science	 7	9
3. Earth and Space Science	 6	9
4. Life Science	 8	9

Question 5.2 (Strand scores):

Since the range for strand scores is from 1 to 9 is the scale a Stanine Scale? Can I say that 4 % MN students with strand scores=1?

- The Stanine Scale and the strand/substrand scores for MCA are different.
- The MCA III strand scale scores are solely transformed from the strand theta score without doing any kind of normalization from administration to administration.
 - A norming process is required to place students' scores on the Stanine Scale.

Question 5.3 (Strand scores):

One student received a 821 on the first OLPA and a 819 on the second. His strand scores on the first OLPA were Number & Operation=1, Algebra=2, Geometry & Measurement=1, and Data Analysis=2. The strand scores for his second test were 2,3,2,3, respectively. Why did the scale score decrease but the strand scores increase?

- Firstly, students with a lower scale score did not get many items correct, so the level of uncertainty for the scale score is high. As a general rule there is more information near the center of the score distribution than the ends of the score distribution.
- Secondly, we can expect that these students did not get many items correct for each strand (or sub-strand), so the level of uncertainty for the sub-scale score is also high.
- Since there is a relationship between SEM and the level of uncertainty (increased uncertainty equals increased SEM), the SEM can have an influence on the scores for that individual near the low and high range of strand and scale scores.



Question 5.4 (Strand scores):

How are the strand/substrand performance levels calculated?

- The strand/substrand performance levels are listed on the individual student report (ISR) 2015-2016.
- The pattern scoring method is used to calculate the strand theta score and SEM of strand theta score.

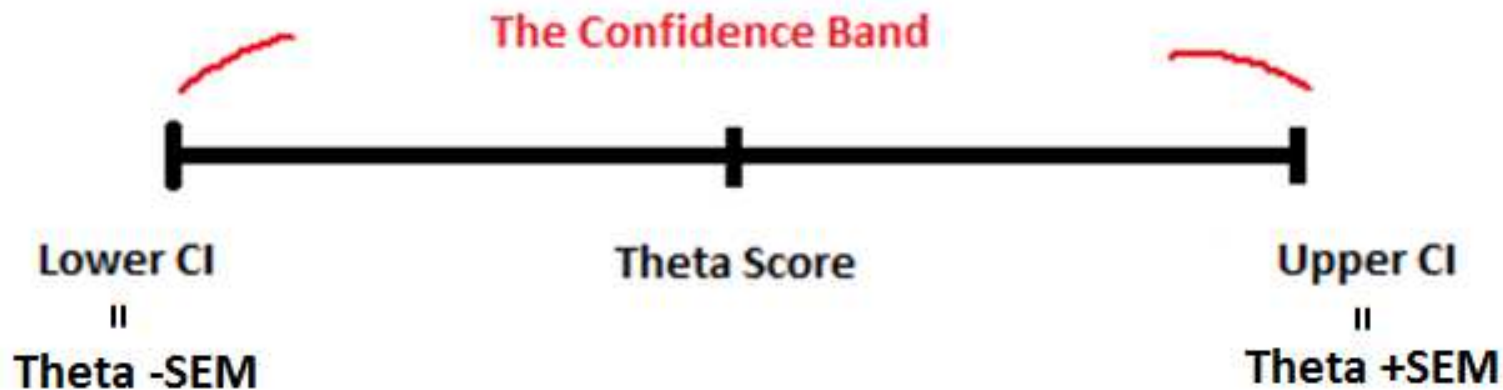


READING: Performance Details

READING AREA	DESCRIPTION	PERFORMANCE
Literature:	Use strategies to analyze, interpret, and evaluate fiction (such as short stories, fables, poetry, and drama).	 Below Expectations
Informational Text:	Use strategies to analyze, interpret, and evaluate nonfiction (such as expository and persuasive text, and literary nonfiction).	 At or Near Expectations

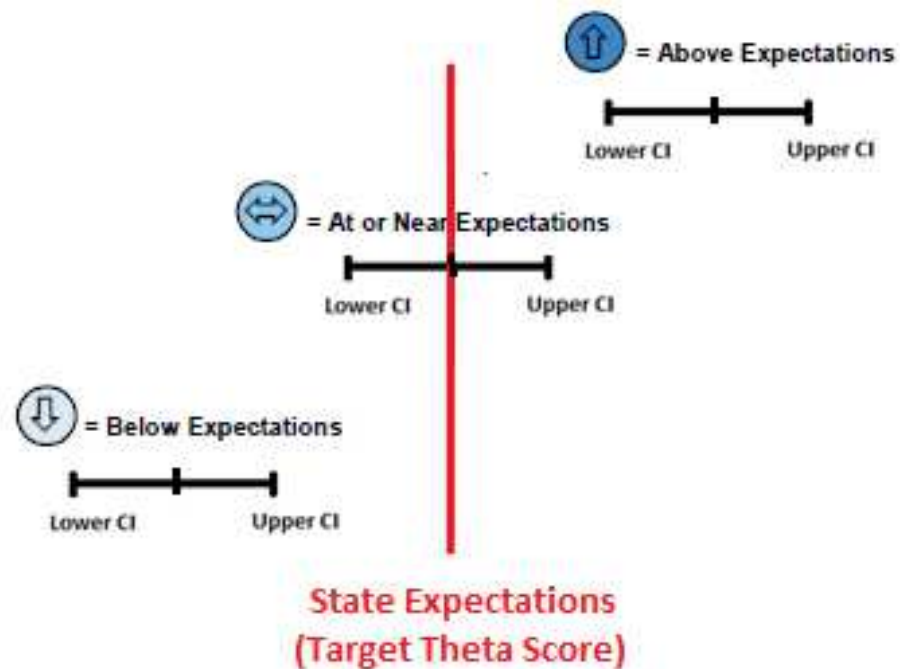
Determine the Strand Performance Level

- The confidence band (lower & upper confidence limit) using the strand theta score and SEM of strand theta score is calculated to determine the strand/substrand performance level.



Determine the strand performance level:

- Above Expectations: Lower CI > Target Theta Score
- At or Near Expectations: Lower CI ≤ Target Theta Score ≤ Upper CI
- Below Expectations: Upper CI < Target Theta Score



References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Joint Technical Committee. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chae, U. Kang, E. Jeon & J. M. Linacre (Eds.), *Development of computerized middle school achievement test (in Korean)*. Seoul, South Korea: Komesa Press.
- McMillan, J. H. (2014). *Classroom assessment: Principles and practice for effective standards-based instruction*. USA: Pearson.