edmentum™

Designing Score Reports to Maximize Validity and Instructional Utility

Karen Barton & Audra Kosh

Validity is about trust and utility.

It is a balance between purpose, defensibility, and the decisions to be made.

Validity in the Balance

Why measure anything?

"We assess for two reasons:

(1) to gather evidence to inform instructional decisions and

(2) to encourage students to try to learn" (Stiggins, 2008, p.3)

From "purpose" and intention to decisions and consequences:

- What are the instructional decisions to be made?
- Who will be making those decisions?
- What information will help them make good decisions?
- *What are the consequences?*

Thomas Kuhn's

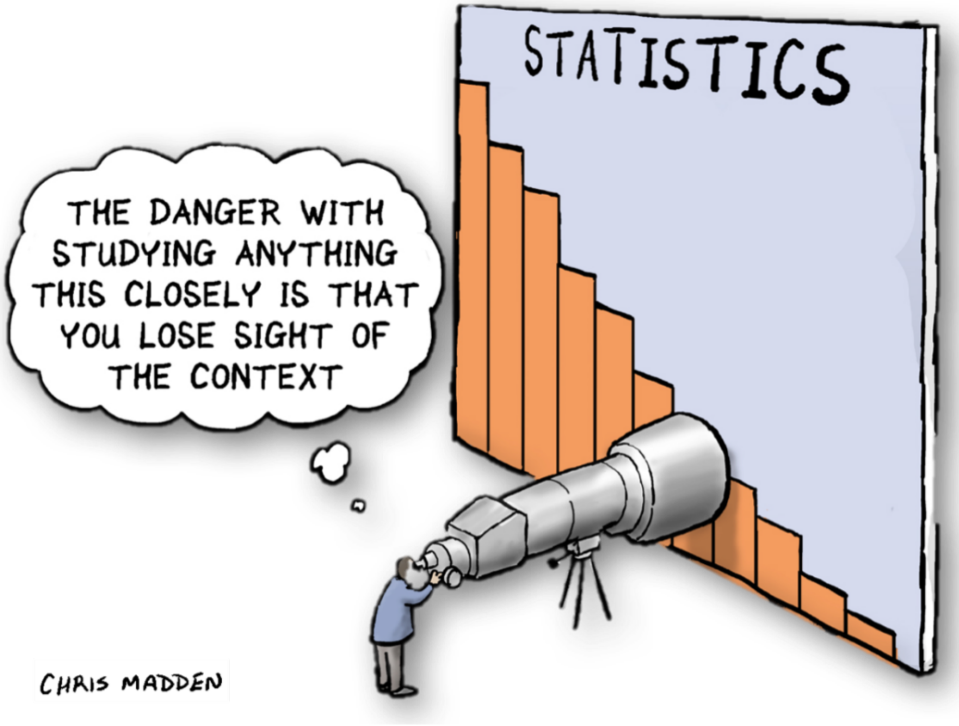
Theory Laden Perspective

And the Impact on

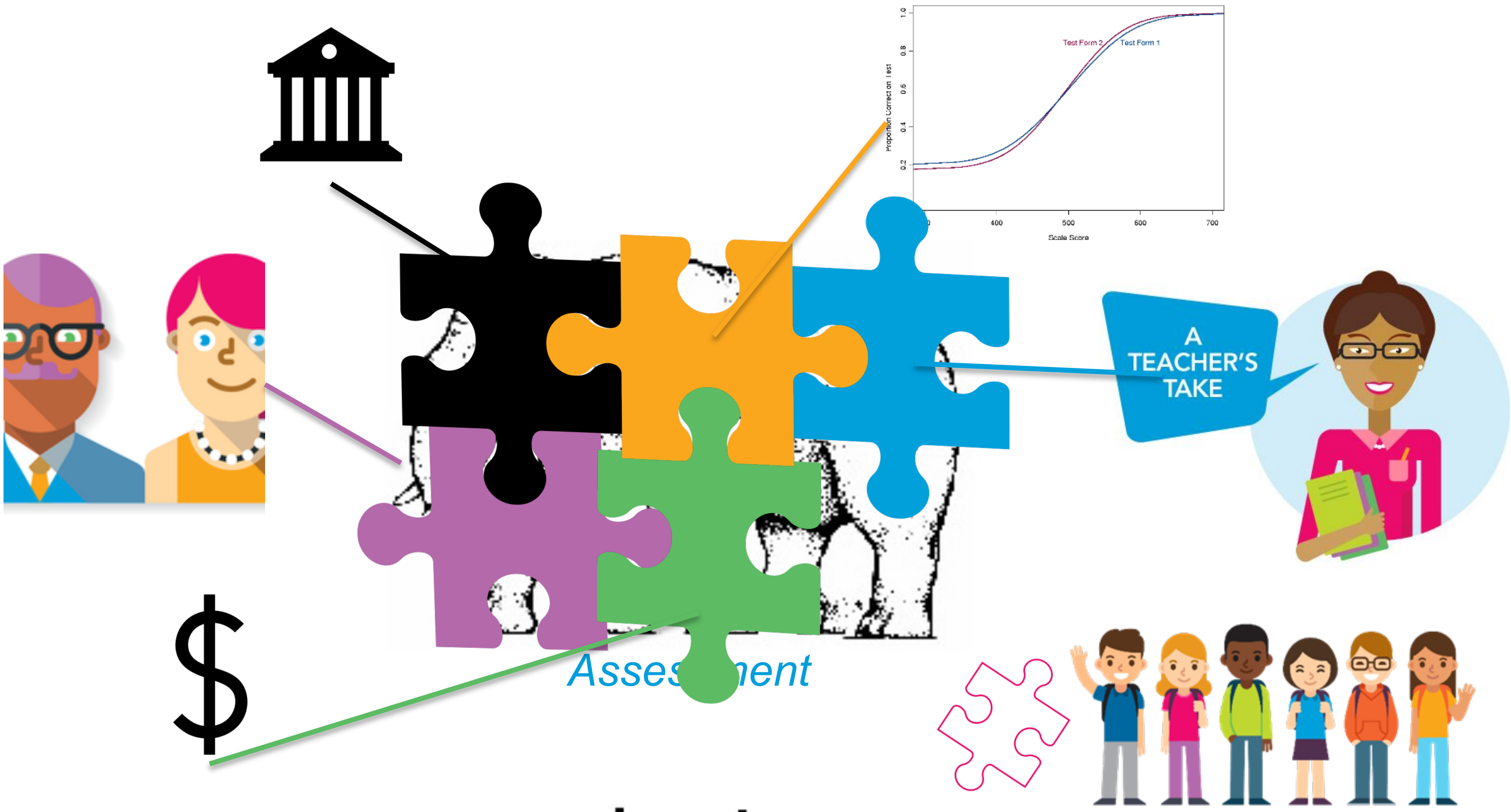
2-Way Communications

SummerPlace
MPLS.
MN.

How close are we?


STATISTICS
THE DANGER WITH STUDYING ANYTHING THIS CLOSELY IS THAT YOU LOSE SIGHT OF THE CONTEXT
CHRIS MADDEN

Paradoxical Perspectives



Validity in the Balance

According to the Standards (2014):

Test score reporting is a developer responsibility.

providing the information

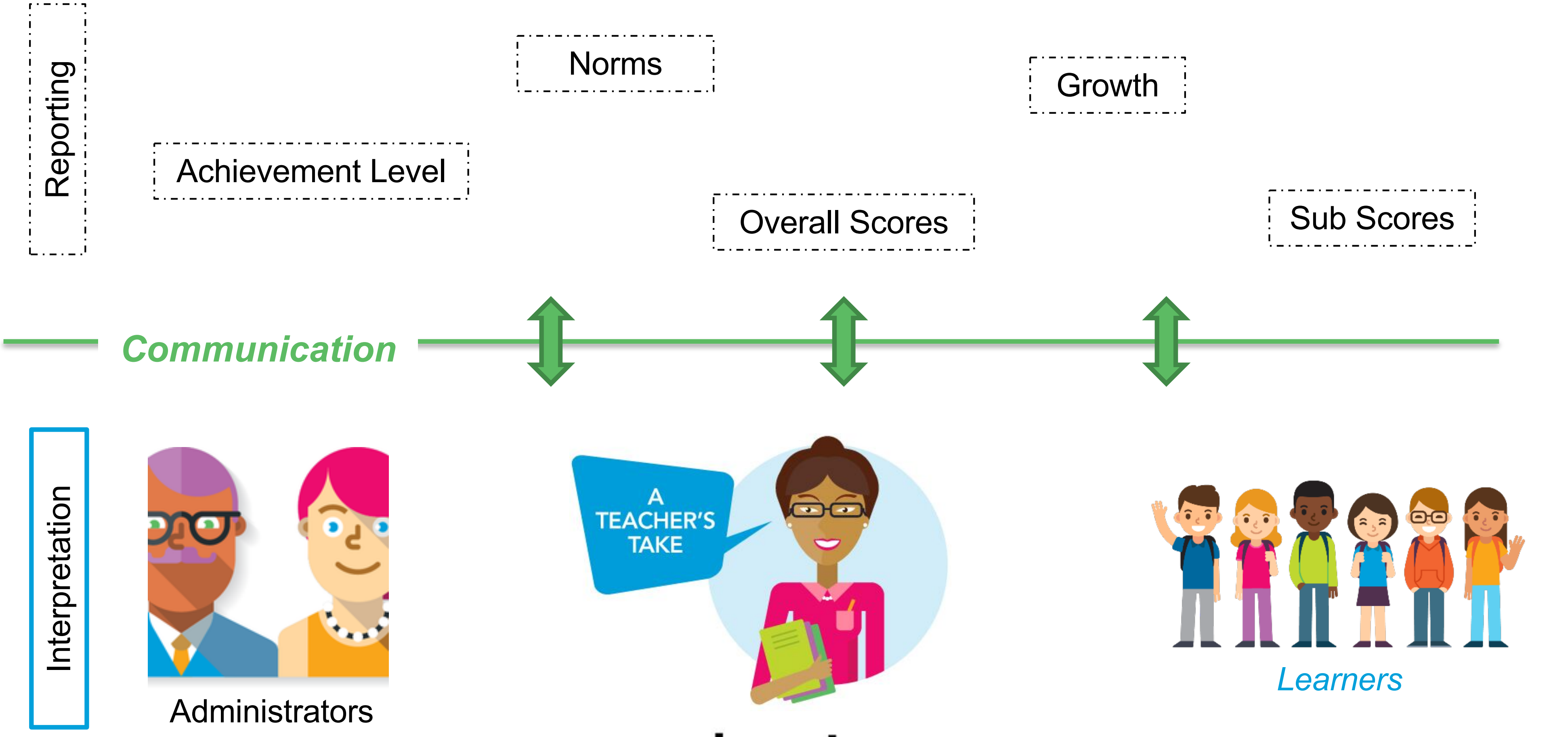
↓↑

Interpretation is the test user responsibility.

understanding, communicating and making decisions

To increase validity of reporting requires attending to information and how it is communicated, as well as greater awareness of context, decisions, and consequences.

The Role of Communication in Validity



Reporting

Norms

Growth

Achievement Level

Overall Scores

Sub Scores

Communication

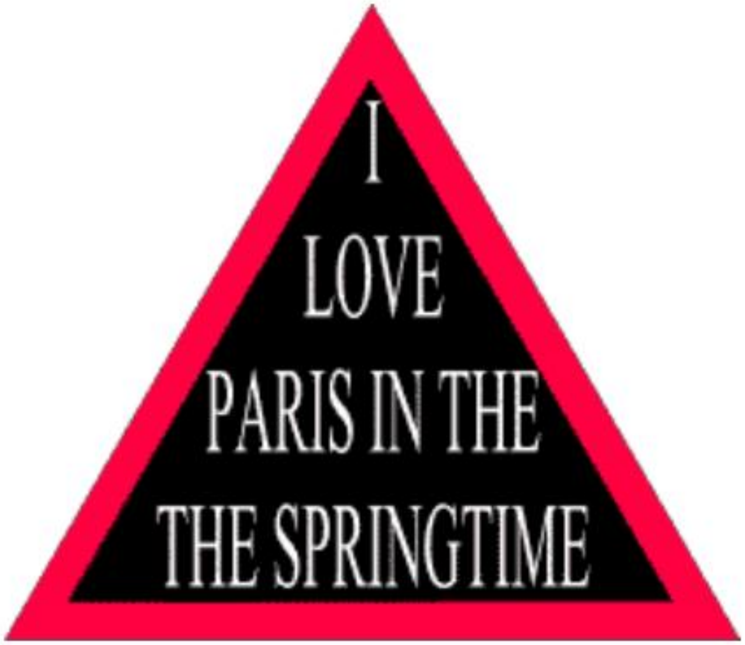
Interpretation

Administrators

Learners

edmentum

Perspectives and Communication



Educator First Approach

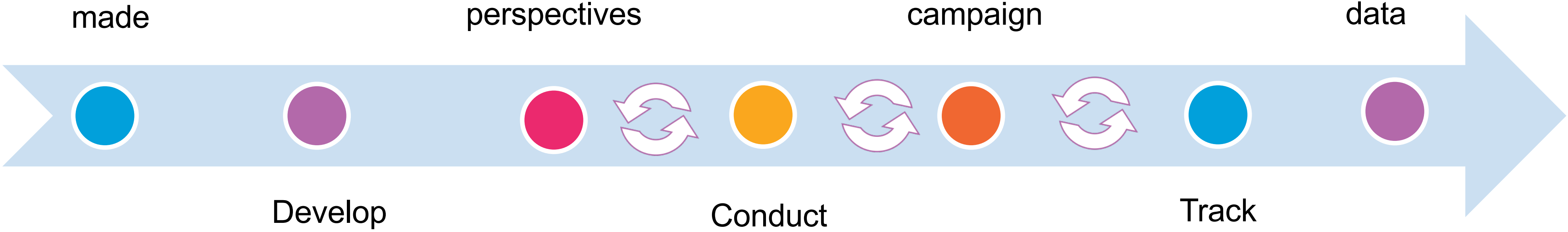
Prior research and guidance

- Principled approach (Lewis, 2019)
- Design recommendations (Zenisky & Hambleton, 2012)

Key elements

- Validity of the design, not just the information
- Reliability or consistency of the interpretations
- Literacy of the information conveyed
- Transparency
- Ease of use
- Actionable

Educator First Workflow Example



Educator interviews to determine key decisions to be made

Develop Research Based Theory of Action relative to key decisions

Launch design and development with multiple perspectives

Conduct UX and Validity testing with Educators

Develop multi-modal literacy campaign

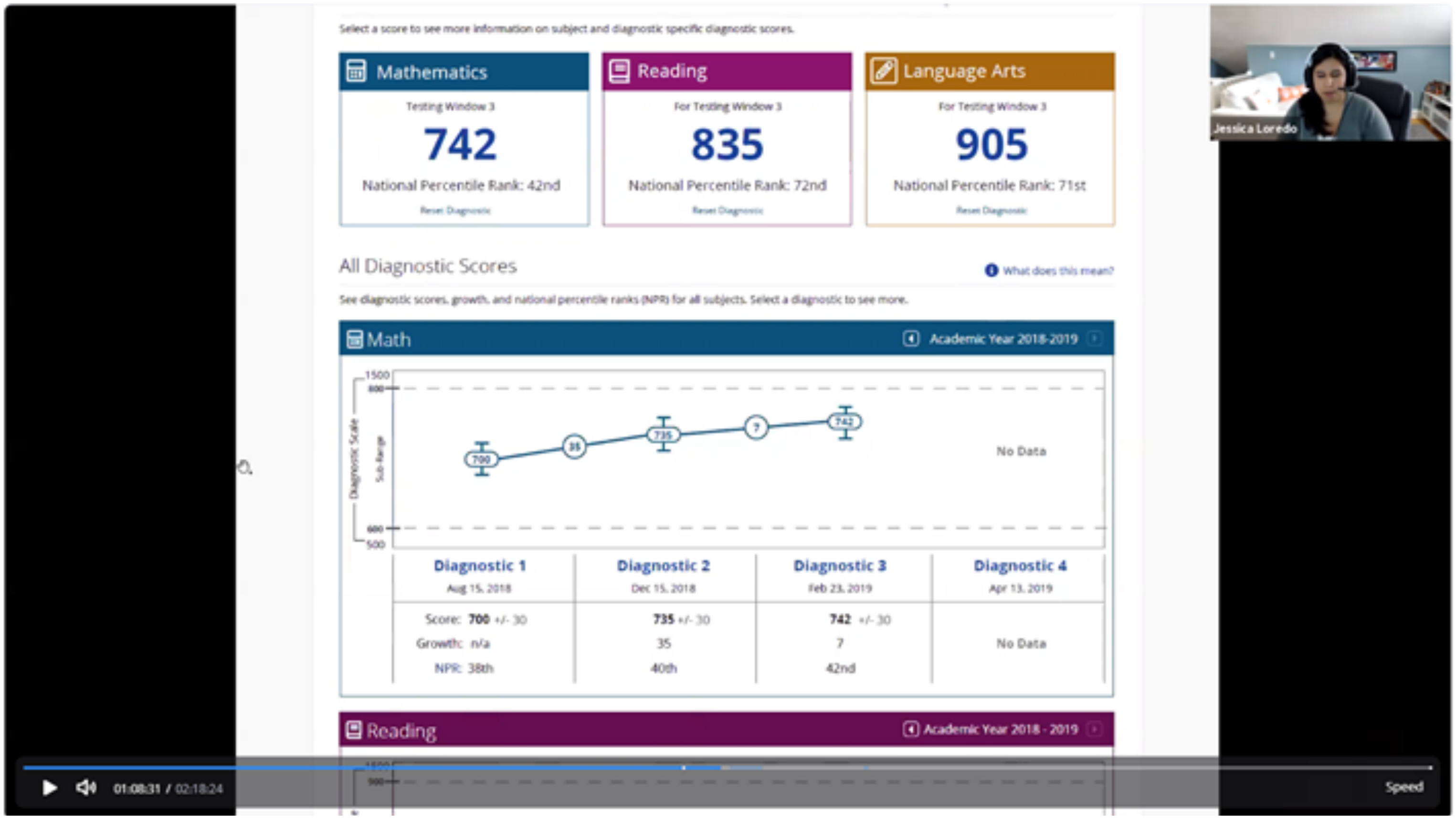
Track feedback in ongoing surveys and interviews

Capture clickstream and usage data

Usability vs Validity Testing

	Usability	Validity
Driver	UI/UX	Research
Format	Focus Group	Individual Educators
Tasks	Open feedback, prompting questions	Locate information, true/false questions
Interaction	Highly conversational	Mostly listening

Usability Testing



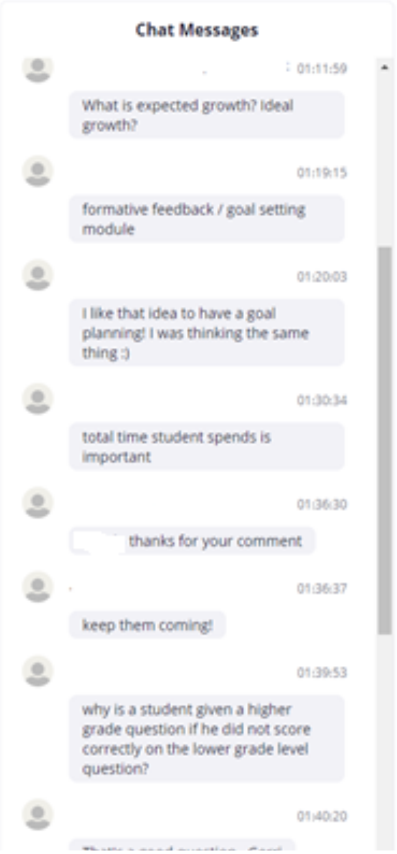
Select a score to see more information on subject and diagnostic specific diagnostic scores.

Mathematics	Reading	Language Arts
Testing Window 3	For Testing Window 3	For Testing Window 3
742	835	905
National Percentile Rank: 42nd	National Percentile Rank: 72nd	National Percentile Rank: 71st
Reset Diagnostic	Reset Diagnostic	Reset Diagnostic

All Diagnostic Scores

See diagnostic scores, growth, and national percentile ranks (NPR) for all subjects. Select a diagnostic to see more.

	Diagnostic 1 Aug 15, 2018	Diagnostic 2 Dec 15, 2018	Diagnostic 3 Feb 23, 2019	Diagnostic 4 Apr 13, 2019
Score:	700 +/- 30	735 +/- 30	742 +/- 30	No Data
Growth:	n/a	35	7	
NPR:	38th	40th	42nd	



Chat Messages

01:11:59
What is expected growth? ideal growth?

01:19:15
formative feedback / goal setting module

01:20:03
I like that idea to have a goal planning! I was thinking the same thing :)

01:30:34
total time student spends is important

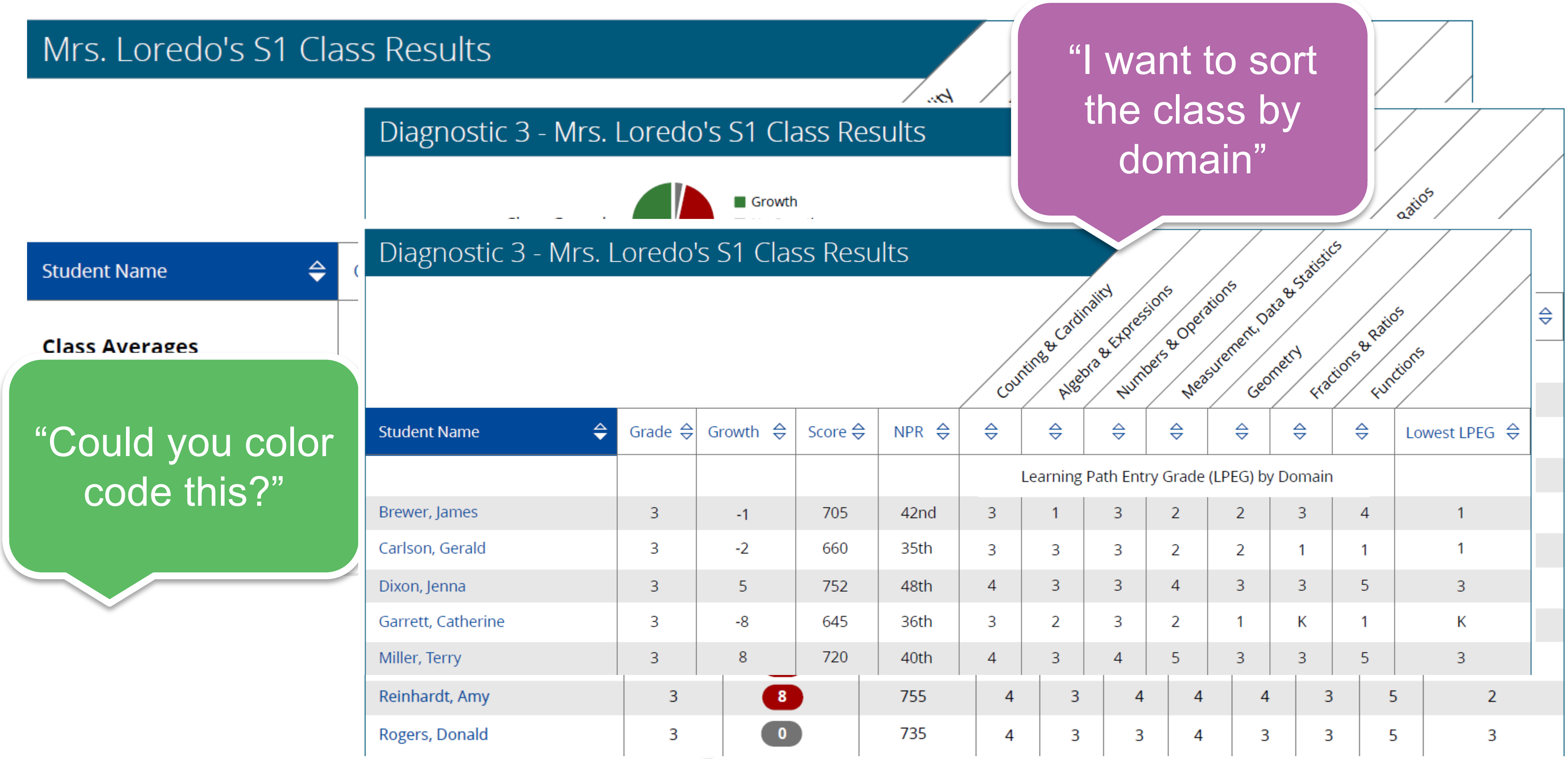
01:36:30
thanks for your comment

01:36:37
keep them coming!

01:39:53
why is a student given a higher grade question if he did not score correctly on the lower grade level question?

01:40:20

Applying Feedback



Mrs. Loredo's S1 Class Results

Diagnostic 3 - Mrs. Loredo's S1 Class Results

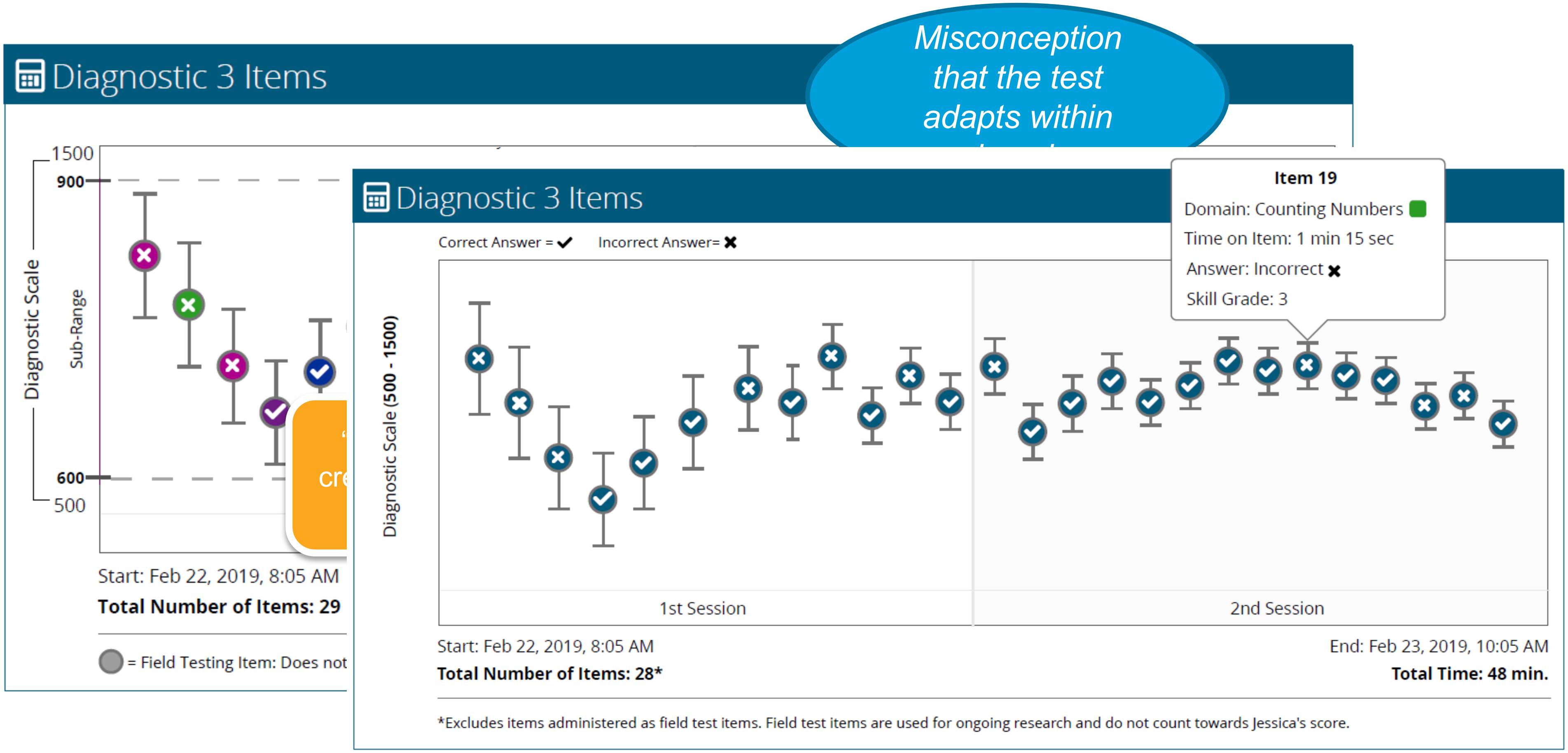
Diagnostic 3 - Mrs. Loredo's S1 Class Results

Student Name	Grade	Growth	Score	NPR	Counting & Cardinality	Algebra & Expressions	Numbers & Operations	Measurement, Data & Statistics	Geometry	Fractions & Ratios	Functions	Lowest LPEG
					Learning Path Entry Grade (LPEG) by Domain							
Brewer, James	3	-1	705	42nd	3	1	3	2	2	3	4	1
Carlson, Gerald	3	-2	660	35th	3	3	3	2	2	1	1	1
Dixon, Jenna	3	5	752	48th	4	3	3	4	3	3	5	3
Garrett, Catherine	3	-8	645	36th	3	2	3	2	1	K	1	K
Miller, Terry	3	8	720	40th	4	3	4	5	3	3	5	3
Reinhardt, Amy	3	8	755		4	3	4	4	4	3	5	2
Rogers, Donald	3	0	735		4	3	3	4	3	3	5	3

“I want to sort the class by domain”

“Could you color code this?”

Applying Feedback



Misconception that the test adapts within

Start: Feb 22, 2019, 8:05 AM

Total Number of Items: 29

● = Field Testing Item: Does not

Item 19

Domain: Counting Numbers 🟩

Time on Item: 1 min 15 sec

Answer: Incorrect ✖

Skill Grade: 3

Start: Feb 22, 2019, 8:05 AM

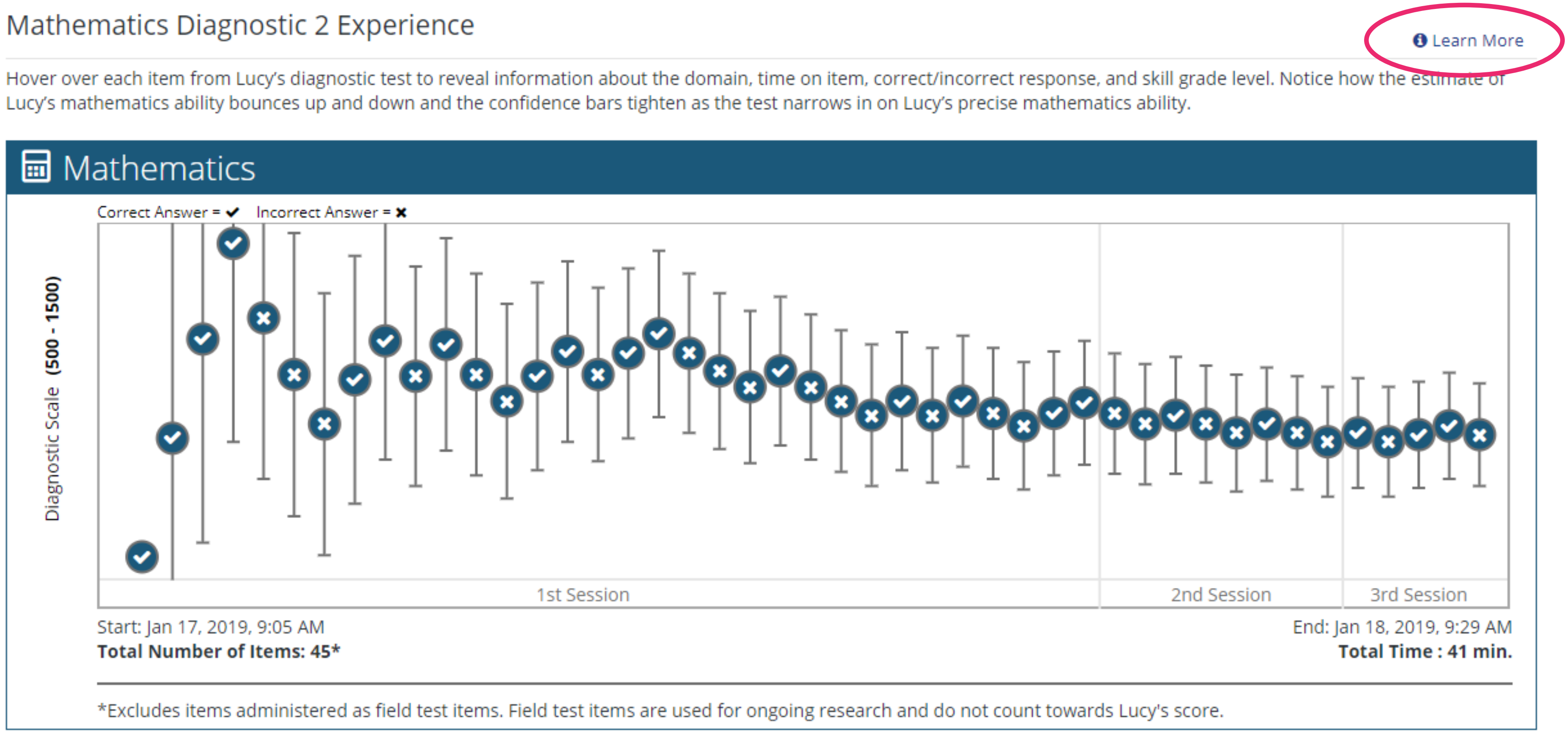
Total Number of Items: 28*

End: Feb 23, 2019, 10:05 AM

Total Time: 48 min.

*Excludes items administered as field test items. Field test items are used for ongoing research and do not count towards Jessica's score.

The Final Design



Mathematics Diagnostic 2 Experience

ℹ Learn More

Hover over each item from Lucy's diagnostic test to reveal information about the domain, time on item, correct/incorrect response, and skill grade level. Notice how the estimate of Lucy's mathematics ability bounces up and down and the confidence bars tighten as the test narrows in on Lucy's precise mathematics ability.

Mathematics

Start: Jan 17, 2019, 9:05 AM

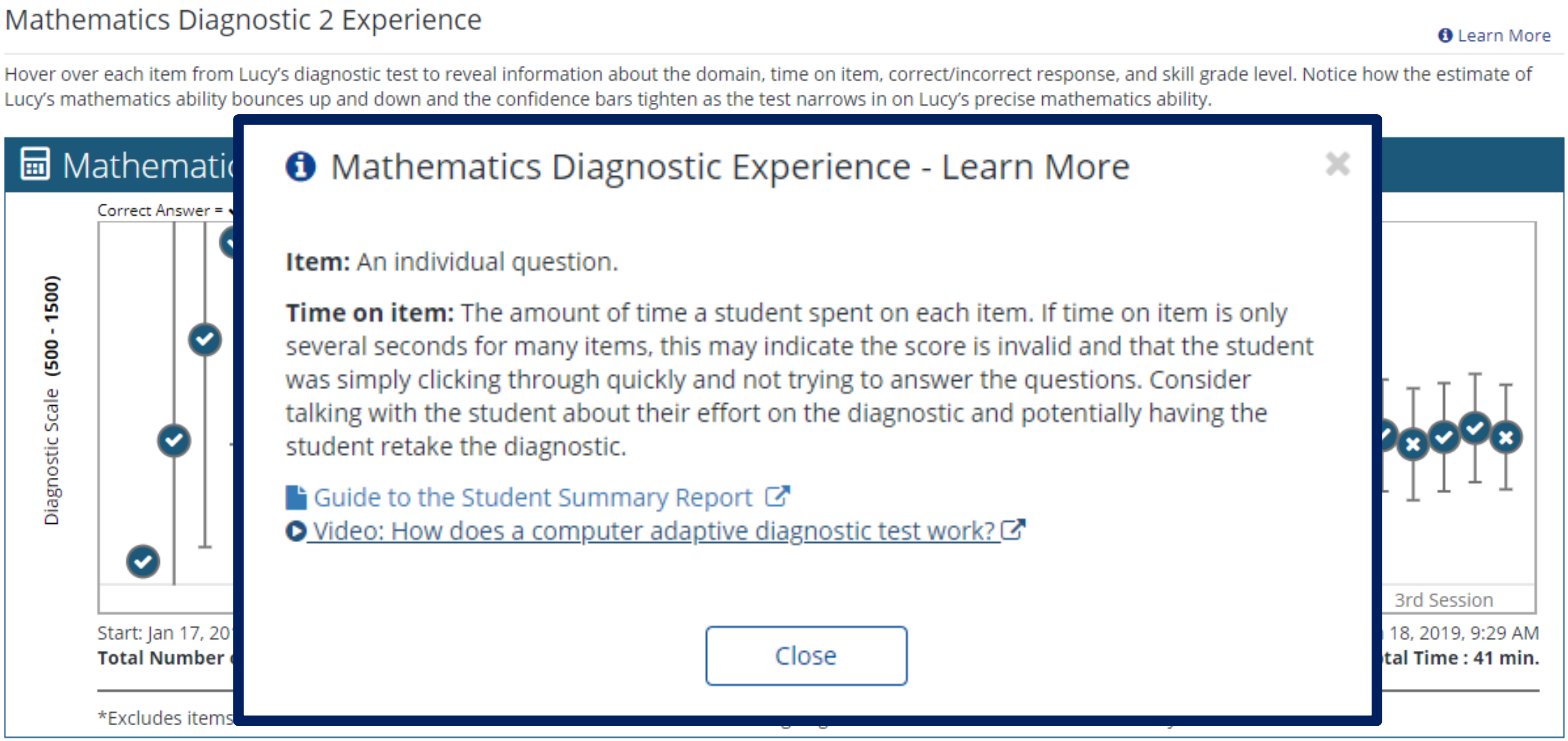
Total Number of Items: 45*

End: Jan 18, 2019, 9:29 AM

Total Time : 41 min.

*Excludes items administered as field test items. Field test items are used for ongoing research and do not count towards Lucy's score.

In-Product Support



Mathematics Diagnostic 2 Experience

Learn More

Hover over each item from Lucy's diagnostic test to reveal information about the domain, time on item, correct/incorrect response, and skill grade level. Notice how the estimate of Lucy's mathematics ability bounces up and down and the confidence bars tighten as the test narrows in on Lucy's precise mathematics ability.

Mathematics Diagnostic Experience - Learn More

Item: An individual question.

Time on item: The amount of time a student spent on each item. If time on item is only several seconds for many items, this may indicate the score is invalid and that the student was simply clicking through quickly and not trying to answer the questions. Consider talking with the student about their effort on the diagnostic and potentially having the student retake the diagnostic.

Guide to the Student Summary Report

Video: How does a computer adaptive diagnostic test work?

Validity Testing Results

Correct interpretation with ease	Correct interpretation with struggle	Incorrect interpretation
✓	—	✘

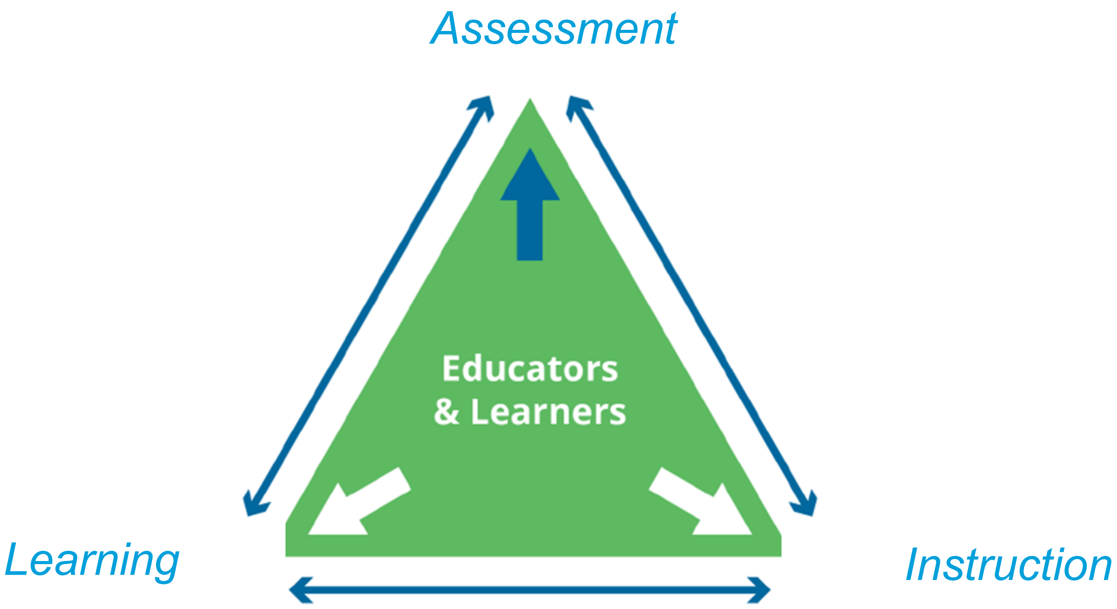
Report	Concept Assessed	Educator A	Educator B	Educator C	Educator D	Educator E
Student Report	National Percentile Rank	✓	✘	✘	✓	✓
	CAT Visual right/wrong indicators	✓	✓	—	✓	✓
	Growth (to get to certain NPR)	—	✘	—	✓	—
	Zoomed-in view of scale		✘	✓	✓	✓
	SEM	✘	✘	✘	✘	✓
	Skill-level raw score information	✓		✓	✘	✓
	Growth (gain score)	—	✘	✓	✘	✘
	CAT Visual (domain color coding)	—	✓	✓	✓	
	CAT Visual (number of test sessions)	✓	✓	✓	✓	✓
	CAT Visual (adaptive nature)	—	✘	✓	✘	✓
	CAT Visual (SEM)	✓	✘	—	✘	✘
Class Report	High/low overall students	✓	✓	✘	✓	—
	Learning Path Entry Grade by domain	—	✘	✓	✓	✓
	Growth	✓	✓	✓	—	✓
	Learning Path Entry Grade Overall	✓	✓	—	—	—
	Scale Score Standard Deviation	—	✘	✘	—	✓
	National percentile rank	✘	✘	✘	✘	✓

Educating and Strengthening Receivers

- Technical Documents
- Blogs, Marketing, Messaging
- Videos
- Training Professional Development
- In Product Support

Triangulation and Context in Reporting

Measures of Growth, beyond a single metric.



Who is the ultimate stakeholder?

How might reports go even further by encouraging student agency and building communications between students and teachers?

Purpose to Impact

- Consider purpose in context
- Establish trust and transparency
- Increasing literacy and impact of actions during testing
- Don’t overestimate utility – *ask*
- Don’t underestimate responsibility – *go beyond*

“You can have brilliant ideas

(or psychometrics and assessment designs),

but if you can't get them across,

your ideas won't get you anywhere.”

~Lee Iacocca

Be Valid – Be Useful