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
The danger with studying anything this closely is that you lose sight of the context.
Paradoxical Perspectives

Assessment

A TEACHER’S TAKE

$
Validity in the Balance

According to the Standards (2014):

**Test score reporting** is a developer responsibility.

**Interpretation** is the test user responsibility.

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

- Norms
- Growth
- Overall Scores
- Sub Scores

Achievement Level

Reporting

Communication

Interpretation

Administrators

Learners

A TEACHER’S TAKE

edmentum

A TEACHER’S TAKE
Perspectives and Communication

I LOVE PARIS IN THE SPRINGTIME
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

1. Educator interviews to determine key decisions to be made
2. Launch design and development with multiple perspectives
3. Develop multi-modal literacy campaign
4. Capture clickstream and usage data

- Develop Research Based Theory of Action relative to key decisions
- Conduct UX and Validity testing with Educators
- Track feedback in ongoing surveys and interviews
# Usability vs Validity Testing

<table>
<thead>
<tr>
<th></th>
<th>Usability</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Driver</strong></td>
<td>UI/UX</td>
<td>Research</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>Focus Group</td>
<td>Individual Educators</td>
</tr>
<tr>
<td><strong>Tasks</strong></td>
<td>Open feedback, prompting</td>
<td>Locate information, true/false questions</td>
</tr>
<tr>
<td></td>
<td>questions</td>
<td></td>
</tr>
<tr>
<td><strong>Interaction</strong></td>
<td>Highly conversational</td>
<td>Mostly listening</td>
</tr>
</tbody>
</table>
Usability Testing
Applying Feedback

Mrs. Loredo's S1 Class Results

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

“I want to sort the class by domain”

“Could you color code this?”
Applying Feedback

“Showing field test items creates a conversation I just don’t want to have.”

Misconception that the test adapts within domain.
The Final Design

Mathematics Diagnostic 2 Experience

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

Correct Answer = ✓ Incorrect Answer = ✗

Diagnostic Scale (500 - 1500)

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

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?]

Close
## Validity Testing Results

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

Correct interpretation with ease: ✓
Correct interpretation with struggle: ✗
Incorrect interpretation: ✗
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