Aligning Curriculum, Instruction, & Assessment to Promote Learning: Technological Advances

Stephen G. Sireci

University of Massachusetts Amherst

Presentation delivered at the Association of Language Testing-Europe’s 59th Meeting and Conference, November 10, 2023, Rome

© Copyright, Stephen G. Sireci, 2023. All rights reserved.
Topics for today

- What is the English Test for Adults?
  - We will use the ETA as an example of the importance of “alignment” validity evidence

- What is alignment, and how does it relate to “content validity?”

- How do we gather, evaluate, and summarize alignment data?

- Future research directions
  - And use of artificial intelligence
Standards and Guidelines

Guidelines for Technology-Based Assessment

International Test Commission and Association of Test Publishers

STANDARDS for Educational and Psychological Testing

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

ESTÁNDARES para Pruebas Educativas y Psicológicas

American Educational Research Association
American Psychological Association
National Council on Measurement in Education

ALTE Principles of Good Practice

Association of Language Testers in Europe
(Some) Articles on which this talk is based

Evidence for Test Validation: A Guide for Practitioners
Stephen Sireci1 and Isabel Benitez2,3
1 University of Massachusetts Amherst, USA
2 University of Granada, Spain
3 Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain

Validity evidence based on test content
Stephen Sireci and Molly Faulkner-Bond
University of Massachusetts Amherst (USA)

Chinese/English Journal of Educational Measurement and Evaluation

Evaluating Alignment Between Curriculum, Assessment, and Instruction
Andrea Martone
The College of Saint Rose
Stephen G. Sireci
University of Massachusetts Amherst

The authors (a) discuss the importance of alignment for facilitating proper assessment and instruction, (b) describe the three most common methods for evaluating the alignment between state content standards and assessments, (c) discuss the relative strengths and limitations of these methods, and (d) discuss examples of applications of each method. They conclude that choice of alignment method depends on the specific goals of a state or district and that alignment research is critical for ensuring the standards-assessment-instruction cycle facilitates student learning. Additional potential benefits of alignment research include valuable professional development for teachers and better understanding of the results from standardized assessment.

Keywords: assessment, test theory and development, test validity and reliability, teacher education and development, psychometrics.
English Test for Adults (ETA) Overview

- **Goal:** Develop high quality reading, writing, speaking, and listening tests proficiency assessments for adult English learners in Massachusetts

- **ETA will be**
  - Aligned with Massachusetts and National ESL curriculum standards
  - Tied to Federal achievement levels
  - Integrated with English instruction in MA adult education courses
Purposes of ETA

1. Measure adult EL’s knowledge and skills in reading, writing, listening, and speaking English
2. Measure adult EL’s gains …
3. Provide teachers with actionable information to improve instruction
4. Provide valid information that can be aggregated for state and Federal accountability purposes
ETA Theory of Action

- Providing information on adult learners’ skills will help them gain the proper education and training they need to accomplish their academic and career goals.

- ASAP assessments:
  - Are accessible to all learners
  - Value adult learners’ funds of knowledge
  - Provide scaffolds, when needed
  - Provide actionable information
ETA Theory of Action

ETT Assessment

Information on Reading, Writing, Listening, Speaking

Culturally-sustaining assessment principles incorporated into test assembly

ACTIONS!!

Learner, Teacher, Counselor, Employer

Educate, Train, Hire, Promote, Certify
Theory of Action

- Adult learner
- Employer
- Teacher

Digital Warehouse of Assessment Modules

- Assessment design
- Reporting system

- Placement mode
- Diagnostic mode
- Manual mode
- Employer mode

Learner characteristics

Culturally-sustaining assessment

Academic and workplace alignment
## Proposed Reading Test Specifications Table

<table>
<thead>
<tr>
<th>Modality</th>
<th>Standard</th>
<th>Percent of Test by EFL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Organization and Style</td>
<td>5-10</td>
</tr>
<tr>
<td></td>
<td>Components of English</td>
<td>35-45</td>
</tr>
<tr>
<td></td>
<td>Total Interpretive</td>
<td>65-75</td>
</tr>
<tr>
<td></td>
<td>Organization and Style</td>
<td>0-10</td>
</tr>
<tr>
<td></td>
<td>Components of English</td>
<td>5-15</td>
</tr>
</tbody>
</table>
Alignment is the “glue” that holds the ETA system together

- Alignment of ETA tasks to
  - English language proficiency curriculum standards (Massachusetts)
  - English language proficiency curriculum standards (Federal)
  - Workplace competencies

- Alignment of curriculum standards to one another
  - And to workplace competencies
Alignment data provide validity evidence!!
Validity and Alignment

What is validity?

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, APA, & NCME, 1999, p. 11)

Five “Sources of Validity Evidence”

1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables
5. Consequences of Testing
What is “content validity?”

The degree to which the content of a test is congruent with the purposes of the testing.

4 Elements of CV:
- Domain definition
- Domain relevance
- Domain representation
- Appropriate test construction procedures

Sireci (1998a,b); Sireci & Faulkner-Bond (2014)
Alignment data as validity evidence based on test content

Webb (1997): Alignment is the "Degree to which expectations and assessments are in agreement...and guide the system towards students learning what they are expected to know and do" (p. 4).
Alignment and Validity

- Alignment studies provide validity evidence based on test content.
  - Evidence regarding the degree to which the content of the test is congruent with the testing purpose.

- However, alignment related to instruction may also provide validity evidence based on testing consequences.
Aligning curriculum, assessment, and instruction
Aligning Curriculum, Instruction, & Assessment

- Domain definition
- Test Specifications
- Domain representation
- Domain relevance
Example Massachusetts ELP Standard

**Reading Strand**

**Reading 4 (Use of effective strategies):** Use a variety of reading strategies appropriate to the reading purpose and type of text. **Referring standards:** Language CCR 4; MA 3

<table>
<thead>
<tr>
<th>ESOL Level 1 (NRS Beg. Lit.)</th>
<th>ESOL Level 2 (NRS Low Beg.)</th>
<th>ESOL Level 3 (NRS High Beg.)</th>
<th>ESOL Level 4 (NRS Low Int.)</th>
<th>ESOL Level 5 (NRS High Int.)</th>
<th>ESOL Level 6+ (NRS Advanced)</th>
</tr>
</thead>
</table>

**A. Use pre-reading strategies.**

- **R4A.1a.** Preview pictures related to a text.
- **R4A.2a.** Preview key vocabulary.
- **R4A.3a.** Preview the title, key vocabulary, and section headings.
- **R4A.4a.** Preview key sections of the text (e.g., advanced organizers, author biography).
- **R4A.5a.** Preview key sections of the text (e.g., heading(s), first sentences of paragraphs).

**Benchmark:** Specific skills and knowledge learners need to develop and demonstrate at a particular level to meet the more broadly stated standard; describes exit-level performance.

**Benchmark notation:**

- **R** = the Reading strand
- **4** = the Standard to which the benchmark belongs
- **B** = the Thread (here: Use pre-reading strategies)

**Strand:** A category of knowledge within the study of a given discipline; here (as in the CCRSAE): Reading, Writing, and Listening/Speaking.

**Source documents:** Indicates one or more of the three standards documents integrated into the MA ELPS: 1) MA Framework, 2) CCRSAE, 3) ELP, 4) OR Standards.

**Thread:** A subcategory for organizing the benchmarks within a standard (here: Use pre-reading strategies.)
Alignment and “Standard Setting”

- In addition to being aligned to curriculum standards, students’ performance must be classified into Federally-established “Educational Functioning Levels.
  - Similar to CEFR levels

- Therefore, “alignment” to these performance levels is also needed.
  - In the USA, most often referred to as “Standard Setting.”
## Federal “Educational Functioning Levels”

<table>
<thead>
<tr>
<th>ELP Standard 1</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
</tr>
</thead>
</table>
| **An ELL can...**
  construct meaning from oral presentations and literary and informational text through level-appropriate listening, reading, and viewing. | **By the end of English language proficiency level 1, an ELL can...**
  use a very limited set of strategies to:
  - identify a few key words and phrases in oral communications and simple spoken and written texts. | **By the end of English language proficiency level 2, an ELL can...**
  use an emerging set of strategies to:
  - identify the main topic in oral presentations and simple spoken and written texts
  - retell a few key details. | **By the end of English language proficiency level 3, an ELL can...**
  use a developing set of strategies to:
  - determine a central idea or theme in oral presentations and spoken and written texts
  - retell key details
  - answer questions about key details
  - explain how the theme is developed by specific details in texts
  - summarize part of a text. | **By the end of English language proficiency level 4, an ELL can...**
  use an increasing range of strategies to:
  - determine a central idea or theme in oral presentations and spoken and written texts
  - analyze the development of the themes/ideas
  - cite specific details and evidence from texts to support the analysis
  - summarize a text. | **By the end of English language proficiency level 5, an ELL can...**
  use a wide range of strategies to:
  - determine central ideas or themes in oral presentations and spoken and written texts
  - analyze the development of the themes/ideas
  - cite specific details and evidence from texts to support the analysis
  - summarize a text. |
So, how do we do all this alignment research?

- And how do we set the performance level standards?
Methods for evaluating content validity/alignment involve:

- **Subject matter experts (SMEs)**
  - Reviewing test items

- **Gathering judgmental data**
  - From SME item review
  - A quality rating form for gathering judgments is critical

- **Summarizing the data**
  - Typically using descriptive statistics, but there are some statistical indices, too.
Evaluating the methods:

- How long does it take for SMEs to make their judgments?
- How complex are the judgments?
- Are there (response) biases associated with any methods?
- How much validity evidence do the methods provide?
  - And are the results easy to understand?
Types of CV rating tasks (1)

- Congruence ratings:
- Two variations: Matching or Rating

(a) “Match each test item to the objective (area) you believe it measures…”
(b) “Read objective (area) and rate the degree to which each item measures it.”
(a) Please **match** each item to 1 of the 3 ELP domains:

<table>
<thead>
<tr>
<th>Item</th>
<th>Reading</th>
<th>Listening</th>
<th>Speaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(b) Rate the congruence of each item to the objective where 1=high congruence, 0=medium congruence, and –1=no congruence:

**Objective:** Infer word meaning from context.

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparing Item—objective (area) congruence methods

- Advantage of MATCHING task to RATING task is SMEs are not informed of the content areas (objectives) each item is supposed to measure.

- Advantage of RATING to MATCHING is more information regarding degree of congruence.
Item relevance ratings:

- “Please rate the relevance of each test item to …”
  - the objective it is intended to measure
  - all objectives
Please rate each item with respect to its relevance for measuring each domain, where 1=not at all relevant and 9=very relevant.”

<table>
<thead>
<tr>
<th>Item</th>
<th>Reading</th>
<th>Writing</th>
<th>Speaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SME data can be summarized using descriptive statistics

2. Proportion of SMEs correctly classifying each item.
   - No rule of thumb but > 70% has been used.
3. Calculating the average proportion “correctly” identified over all items.
4. Mean relevance ratings, Aiken index

# Content Validity Results: ELP Test

Summary of SMEs’ Content Validity Ratings by Domain

<table>
<thead>
<tr>
<th>Domain</th>
<th># Items</th>
<th>% Congruent</th>
<th>% Unanimous</th>
<th>% Not Congruent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>89</td>
<td>94%</td>
<td>55%</td>
<td>6%</td>
</tr>
<tr>
<td>Writing</td>
<td>85</td>
<td>87%</td>
<td>54%</td>
<td>13%</td>
</tr>
<tr>
<td>Listening</td>
<td>102</td>
<td>80%</td>
<td>50%</td>
<td>20%</td>
</tr>
<tr>
<td>Speaking</td>
<td>76</td>
<td>53%</td>
<td>22%</td>
<td>47%</td>
</tr>
<tr>
<td>Total</td>
<td>352</td>
<td>80%</td>
<td>46%</td>
<td>20%</td>
</tr>
</tbody>
</table>

“Congruent” = items matched by 4-6 SMEs. “Unanimous” = items matched by all 6 SMEs. “Not Congruent” = items matched by less than 4 SMEs.
Advantages/Disadvantages:

- **Matching (congruence) ratings:**
  - quick and easy for SMEs (+)
  - simple calculations (+)
  - data are easy to understand (+)
  - no information regarding *how well* items measure objectives (-)
  - expectancy bias/social desirability (-)
  - no statistical index of quality
Advantages/Disadv. (cont.):

- **Item relevance ratings:**
  - provide information regarding how well items measure objectives (+)
  - Aiken and other indices can be evaluated for statistical significance (+)
  - more work for SMEs
  - harder to compute and explain (-)
  - expectancy bias/social desirability (-)
Another CV method: Item similarity ratings:

- Please rate the following item pairs with respect to the science knowledge and skills measured:

  1=very similar

  10= very different
Please rate the similarity of these two test items with respect to the knowledge and skills they measure.

1) What did the author mean by the word “force?”
2) Why did Javonte want to go to school on a Saturday?

Very Similar

Very different

1 2 3 4 5 6 7 8 9 10
Similarity ratings:

- **Logic:** Items that are designed to measure the same objectives will be perceived as more similar than items designed to measure different objectives.

- **Advantage:** no social desirability in responding.
How to analyze SME similarity data: Multidimensional Scaling (MDS)

This equation defines distances between points in Euclidean space, where \( a \) is a specific dimension in \( r \)-dimensional space, and \( x \) is the coordinate for stimuli \((i \text{ or } j)\) on dimension \( a \).

\[
d_{ij} = \sqrt{\sum_{a=1}^{r} (x_{ia} - x_{ja})^2}
\]
Sireci, Robin, Meara, Rogers, & Swaminathan (2000)

- Science teachers rating NAEP science test items
- Paired comparisons (similarity ratings)
2-D Item Subspace: D4 & D5

Dimension 4 (Life vs. Earth)

E=Earth Science, L=Life Science, P=Physical Science
Similarity ratings” Adv/Disadv:

- No expectancy bias/social desirability (+)
- Visual interpretation (+)
- Also assesses domain definition (+)
- Time consuming for SMEs (-)
- Can be difficult to interpret (-)
- No statistical index (-)
- Complex data analysis (-)
Evaluating Test Content via “Alignment”

- Many different “models” or methods
  - Webb
  - La Marca
  - Porter (Surveys of Enacted Curriculum)
  - Achieve
  - Hybrids

(see Bhola, Impara, & Buckendahl, 2003; Martone & Sireci, 2009)
Webb Methodology: SMEs’ tasks

- Categorical concurrence: match test items to benchmarks/objectives
- Depth-of-knowledge consistency:
  - rate cognitive complexity of objectives and of items measuring them
- Range-of-knowledge consistency:
  - # of benchmarks w/in standard measured by ≥ 1 item
- Balance of representation
  - how evenly distributed are items across objectives?
Can artificial intelligence help us evaluate, or create, alignment?

- Ovviamente!
  (of course)
Using AI to Link CCRSAE and O*NET

<table>
<thead>
<tr>
<th>Training data (Input)</th>
<th>Natural Language Processing</th>
<th>Human verification</th>
<th>Crosswalk (Output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texas Standards</td>
<td>Reading Standards (CCRSAE) (77)</td>
<td>Occupations (O*NET) (923)</td>
<td>Occupations per standard (top 3)</td>
</tr>
<tr>
<td>Online job ads (+100K)</td>
<td>Ask and answer questions about key details in a text (RI/RL.1.1).</td>
<td>Construction Laborers (ONET: 47-2061.00)</td>
<td>Receptionists and Information Clerks (ONET: 43-4171.00).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ask and answer questions about key details in a text (RI/RL.1.1).</td>
<td>Proofread documents, records, or other files to ensure accuracy (ONET: Detailed Work Activity).</td>
</tr>
</tbody>
</table>

**Application:**
- Apply vocabulary and concepts accurately in reading, speaking, and writing (TX ELA 1.2.A).
- Proofread documents, records, or other files to ensure accuracy (ONET Detailed Work Activity).

**Questions Answered:**
- Ask and answer questions about key details in a text (RI/RL.1.1).
Questions? What are the important validity questions about test content YOU think need to be answered?

What *evidence* is needed to justify use of the test for a specific purpose?
Content validity and alignment research

- Testing agencies, researchers, and educators have different reasons for evaluating alignment.
- Thus, the goals of an alignment study should be clearly specified in advance before deciding on alignment method.
Content Validity Questions

1. Do the *test specifications* represent the knowledge and skills specified in the appropriate *curriculum frameworks*?

2. Does the test content sufficiently represent the *test specifications*?

3. Does the content sufficiently represent the *curriculum framework*?

4. Are all items *relevant* to the curricular domain?

5. Are any items potentially biased against certain types of students?
Content Validity Questions

6. Is the content sufficient for providing the information desired, given the testing purposes?
Aligning Curriculum, Instruction, & Assessment

- Domain definition
- Domain representation
- Domain relevance

Curriculum

Test Specifications

Instruction

Assessment

TC
Additional Alignment Questions

7. Has the mandated curriculum had an effect on instruction?
   - Would need to be evaluated over time

8. Are teachers better trained or resourced due to mandated testing?

Provides validity evidence based on testing consequences
In conclusion (1)

- Alignment research can provide important information regarding:
  - The degree to which tests are fulfilling their purposes
  - The degree to which students’ performance can be interpreted with respect to a defined domain
  - How an assessment should be improved to better meet its goals
  - Students’ opportunity to learn

- Research should be INDEPENDENT of test developers
Conclusions (2)

- There are many ways to evaluate alignment.
- To select the best method, or piece of a method, identify your goals.
- Alignment research is an important part of quality language assessment.
Future directions

- **Artificial intelligence**
  - Can it be used to compute alignment indices?
  - Test-test alignment
  - Test-curricula alignment

- **Content validity indices for assessment systems**
  - No more test forms
  - Item banks
  - DIRTy assessment
21st-century Testing: “Personalized assessment”

- Goal is to develop best assessment for each individual person
- Consistent with UNDERSTANDardization
Thank you for taking this math test. Would you like to take the first item in English, Korean, or Spanish?

수학 시험에 응해주셔서 감사합니다. 첫 번째 문항을 영어와 한국어 중 어떤 언어로 푸시겠습니까?

Gracias por tomar este examen de matemáticas. ¿Le gustaría tomar el primer elemento en inglés o español?
MOVING FROM CHOICE IN LANGUAGE TO CHOICE IN ITEM CONTEXT

- Reading passages
- Writing prompts
- Contexts in other subject areas (e.g., licensure areas of specialization)
READING TEST

Determine the meaning of general academic and domain-specific words and phrases in a text relevant to a topic or subject area. (RI.3.4)
We are going to give you an article to read?
Would you like to read about…
(choose one)

Critical Race Theory  
Sports  
Food  
Something else

Click here  
Click here  
Click here  
Click here
You chose “sports.” Which sports article would you like to read? (choose one)

UMass Football

UMass Women’s Basketball

Psychometricians playing ping pong

Click here

Click here

Click here
The UMass football team lost another *tough* home game on Saturday. They moved the ball well both on the ground and in the air. They scored two rushing touchdowns and two passing touchdowns. However, the defense was not as good. Smith College scored 120 points.

What does the author mean by the word “*tough*” in this paragraph?

(a) sad  
(b) rough  
(c) mean  
(d) tender
The UMass women’s basketball team won a great game last night. They moved the ball well and played well on defense. They made nine three-point shots and made almost all of their free throws. The final score was 80 to 60. They improved their record to 9 and 5.

What does the author mean by the word “final” in this paragraph?
(a) end  
(b) grand  
(c) game  
(d) foremost
The UMass football team lost another tough home game on Saturday. They moved the ball well both on the ground and in the air. They scored two rushing touchdowns and two passing touchdowns. However, the defense was not as good. Smith College scored 120 points.

 FK ease (83), Grade Level 3.9

The UMass women’s basketball team won a great game last night. They moved the ball well and played well on defense. They made nine three-point shots and made almost all of their free throws. The final score was 80 to 60. They improved their record to 9 and 5.

 FK ease (88), Grade Level 3.4
Closing remarks

- Validity evidence based on test content is NECESSARY to justify use of a test for a particular purpose
  - But is not SUFFICIENT for such justification
- Much work to do, but good news is there are methods and research available to help us.
Thanks to ALTE for the invitation!
UMass Center for Educational Assessment
Sireci@umass.edu
See you in Granada, Spain!!
International Test Commission Biennial Conference