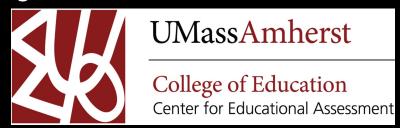
Aligning Curriculum, Instruction, & Assessment to Promote Learning: Technological Advances

Stephen G. Sireci

University of Massachusetts Amherst



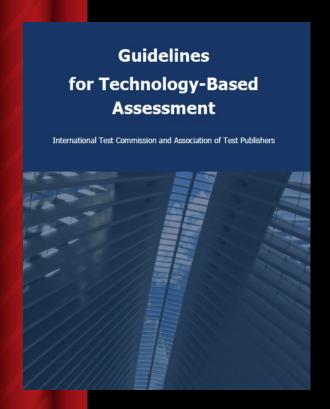
Presentation delivered at the Association of Language Testing-Europe's 59th Meeting and Conference, November 10, 2023, Rome

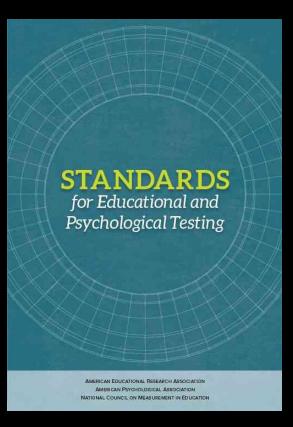
© Copyright, Stephen G. Sireci, 2023. All rights reserved.

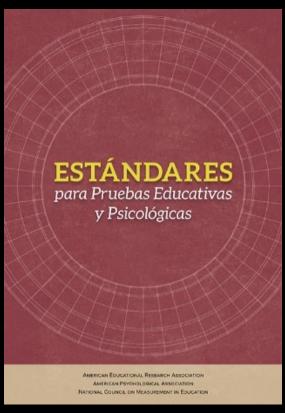
Topics for today

- What is the E η glish 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









ALTE Principles of Good Practice

(Some) Articles on which this talk is based



Psicothema (2023) 35(3) 217-226 **Psicothema**

https://www.psicothema.com • ISSN 0214-9915

Colegio Oficial de Psicólogos del Principado de Asturias





Psicothema 2014, Vol. 26, No. 1, 100-107 doi: 10.7334/psicothema2013.256

ISSN 0214 - 9915 CODEN PSOTEG Copyright @ 2014 Psicotherna www.psicothema.com

Article

Evidence for Test Validation: A Guide for Practitioners

Stephen Sireci¹ and Isabel Benítez^{2,3}

1 University of Massachusetts Amherst, USA ² University of Granada, Spain 3 Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain

ARTICLE INFO

ABSTRACT

Received: November 16, 2022

Background: Validity is a core topic in educational and psychological assessment. Although there are many available resources describing the concept of validity, sources of validity evidence, and suggestions about how

Validity evidence based on test content

Stephen Sireci and Molly Faulkner-Bond University of Massachusetts Amherst (USA)

Abstract

Background: Validity evidence based on test content is one of the five forms of validity evidence stipulated in the Standards for Educational and Psychological Testing developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. In this paper, we describe the logic and

Article 3

Resumen

Evidencia de validez basada en el contenido del test. Antecedentes: la evidencia de validez basada en el contenido del test es una de las cinco formas de evidencias de validez estipuladas en los Standards for Educational and Psychological Testing de la American Educational Research Association. En este artículo describimos la lógica y teoría que subyace a tal fuente de

Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 1 | Issue 1

2020

De-"Constructing" Test Validation

Stephen G. Sireci

Review of Educational Research December 2009, Vol. 79, No. 4, pp. 1332-1361 DOI: 10.3102/0034654309341375 © 2009 AERA. http://rer.aera.net

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-assessmentinstruction cycle facilitates student learning. Additional potential benefits of alignment research include valuable professional development for teachers and better understanding of the results from standardized assessments.

KEYWORDS: assessment, test theory and development, test validity and reliability, teacher education and development, psychometrics.

Eηglish 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

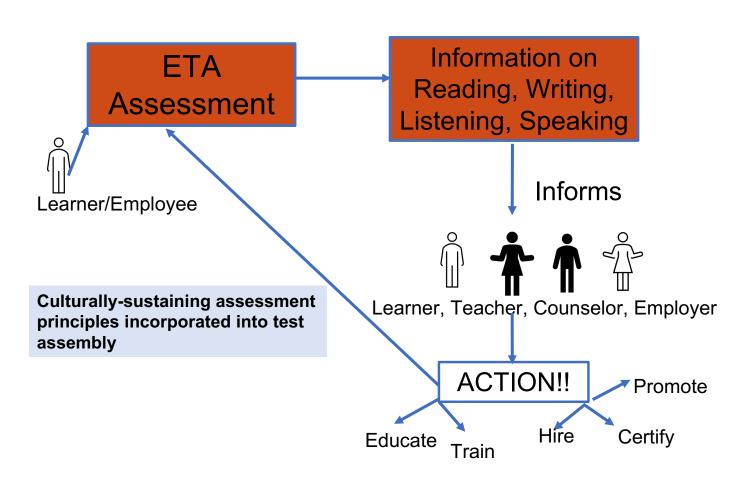
ETA Theory of Action

 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:

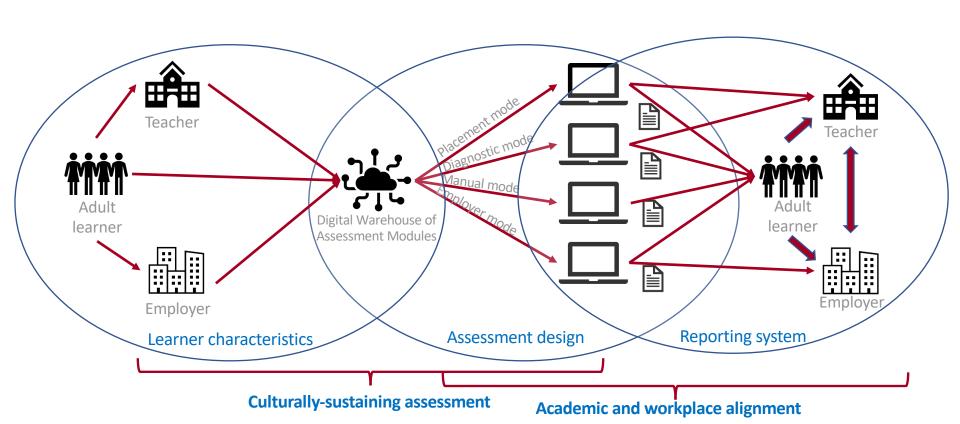
- Are accessible to all learners
- Value adult learners' funds of knowledge
- Provide scaffolds, when needed
- Provide actionable information

ETA Theory of Action



Theory of Action





Proposed Reading Test Specifications Table

Modality	Standard	Percent of Test by EFL					
		1	2	3	4	5	6
Interpretive	Focus on Meaning	25-35	25-35	25-35	15-25	15-25	15-25
	Organization and Style	5-10	5-10	5-10	15-25	15-25	15-25
	Components of English	35-45	35-45	35-45	10-20	10-20	10-20
	Total Interpretive	65-75	65-75	65-75	55-65	55-65	55-65
Interactive	Focus on Meaning	10-25	10-25	10-25	15-30	15-30	15-25
	Organization and Style	0-10	0-10	5-10	5-15	5-15	10-20
	Components of English	5-15	5-15	10-20	5-10	5-10	10-15
	Total Interactive	25-35	25-35	25-35	35-45	35-45	35-45

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)



Standards for Educational & Psychohological Testing (1999, 2014)

Five "Sources of Validity Evidence"

- 1. Test content
- 2. Response processes
- 3. Internal structure
- 4. Relations to other variables
- 5. Consequences of Testing



AMERICAN ECUCATIONAL RESEARCH ARTOCISTY

AMERICAN PSYCHOLOGICAL ASSOCIATION

NATIONAL COLNICIL ON MCASUREMENT IN BOUCK

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

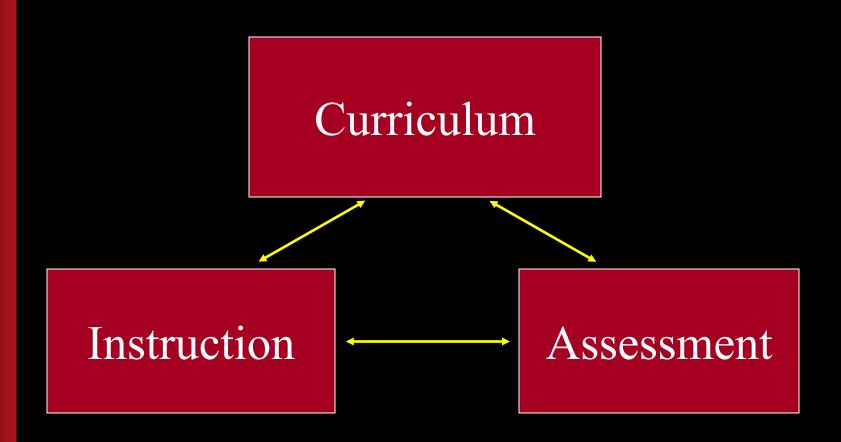
Webb (1997): Alignment is the

"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).

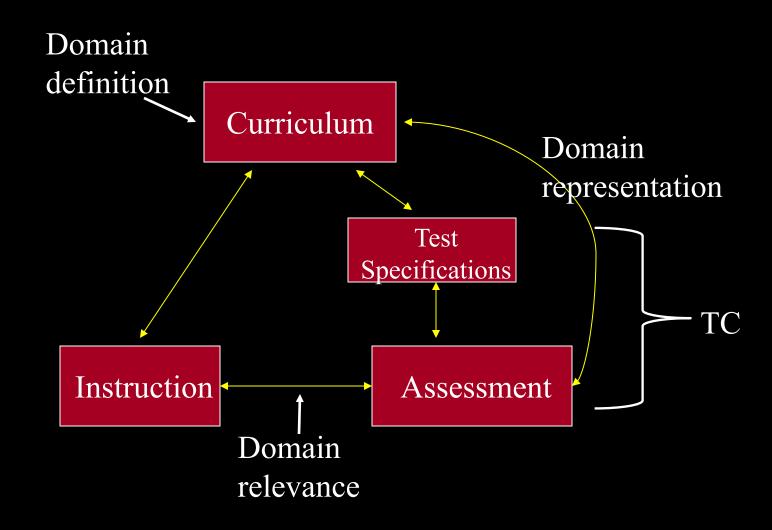
Alignment and Validity

- 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



Example Massachusetts ELP Standard

Standard: States in broad terms what learners are able to do; applies to all six levels and thus extends across all columns.

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

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

ESOL Level 1 (NRS Beg. Lit.) ESOL Level 2 (NRS High Beg.) ESOL Level 4 (NRS Low Int.) ESOL Level 5 (NRS High Int.)

A. Use pre-reading strategies.

R4A.1a. R4A.2a. Preview Preview key Previ

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.

Prave v t

Previous the title, key vocabulary, an section

eadings.

R4A.4a. R4A.5a. Preview key sections of the

text (e.g., heading(s), first sentences

author of paragraphs), biography).

Source documents:

dyanced)

QL Level 6+

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

Benchmark notation:

R = the Reading strand

4 = the Standard to which the benchmark belongs

B = the Thread (here: Use pre-reading strategies)

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"

ELP Standard 1	Level 1	Level 2	Level 3	Level 4	Level 5
An ELL can construct meaning from oral presentations and literary and informational text through level- appropriate listening, reading, and viewing.	Level 1 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 communicatio ns and simple spoken and written texts.	By the end of English language proficiency level 2, an ELL can use an emerging	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	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	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
			explain how the theme is developed by specific details in texts summarize part of a text.	ideas cite specific details and evidence from texts to support the analysis summarize a text.	 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: <u>Matching</u> or <u>Rating</u>
- (a) "Match each test item to the objective (area) you believe it measures..."
- (b) "Read objective (area) and <u>rate</u> the degree to which each item measures it."

(a) Please <u>match</u> each item to 1 of the 3 ELP domains:

```
Item Reading Listening Speaking123
```

(b) <u>Rate</u> the congruence of each item to the objective where 1=high congruence, 0=medium congruence, and -1=no congruence:

Objective: context.	Infer word meaning from
<u>Item</u>	
1	
2	
3	

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."

Item	Reading	Writing	Speaking
1			
2			
3			
4			

SME data can be summarized using descriptive statistics

- 1. Hambleton's (1984) item—objective congruence index.
- 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
 - Others: see Crocker et al. (1989), Osterlind (1989), Sireci (1998), textbook.

Content Validity Results: ELP Test

Summary of SMEs' Content Validity Ratings by Domain

Domain	# Itoma	%	%	% Not	
Domain	# Items	Congruent	Unanimous	Congruent	
Reading	89	94%	55%	6%	
Writing	85	87%	54%	13%	
Listening	102	80%	50%	20%	
Speaking	76	53%	22%	47%	
Total	352	80%	46%	20%	

"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?

1 2 3 4 5 6 7 8 9 10 Very Similar Very

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.

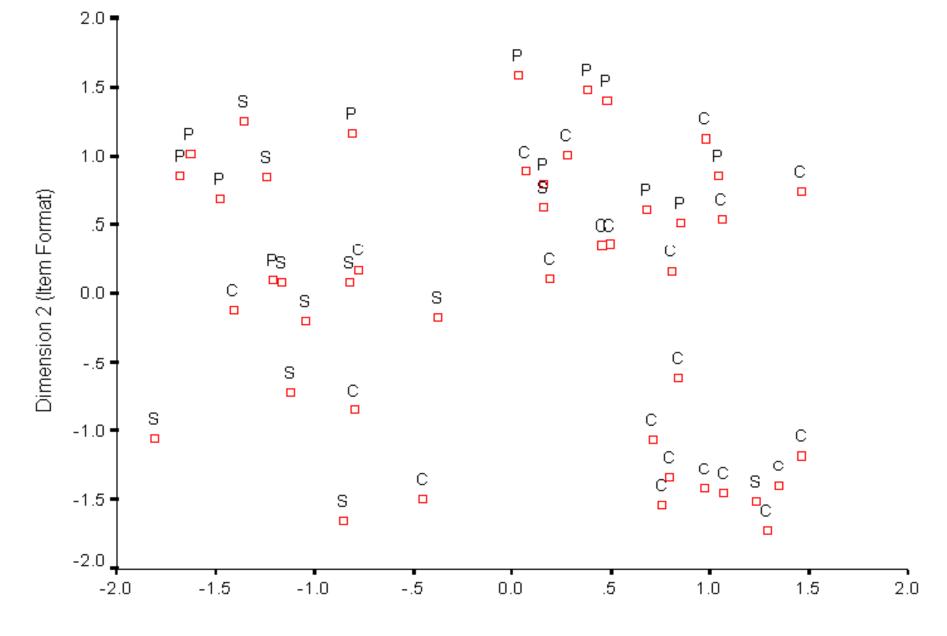
How to analyze SME similarity data: Multidimensional Scaling (MDS)

$$d_{ij} = \sqrt{\sum_{a=1}^{r} (x_{ia} - x_{ja})^2}$$

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* or *j*) on dimension *a*.

Sireci, Robin, Meara, Rogers, & Swaminathan (2000)

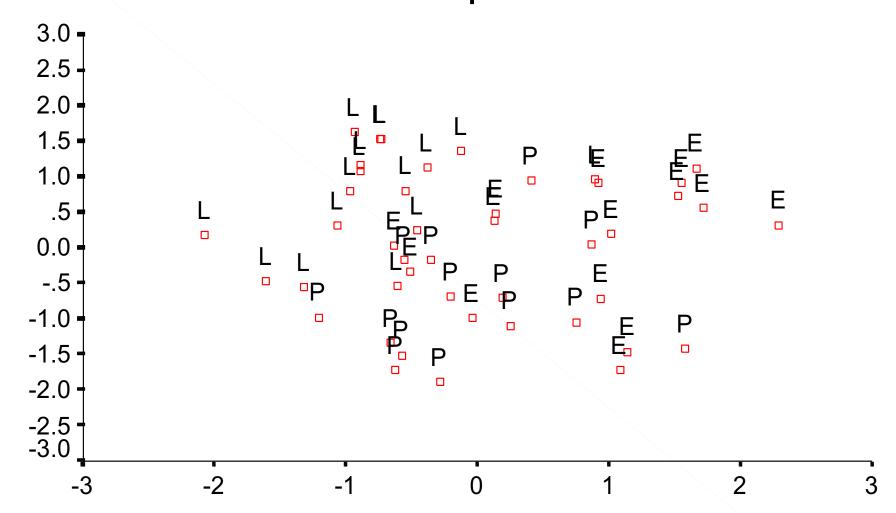
- Science teachers rating NAEP science test items
- Paired comparisons (similarity ratings)



Dimension 1 (Conceptual Understanding)

C=Concept. Understand., P=Pract. Reason., S=Sci. Investig.

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)

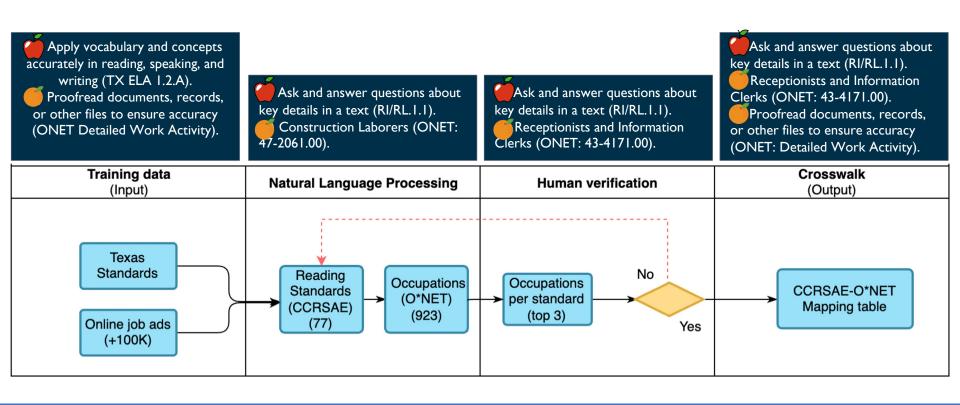
Webb Methodology: SMEs' tasks

- 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



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.

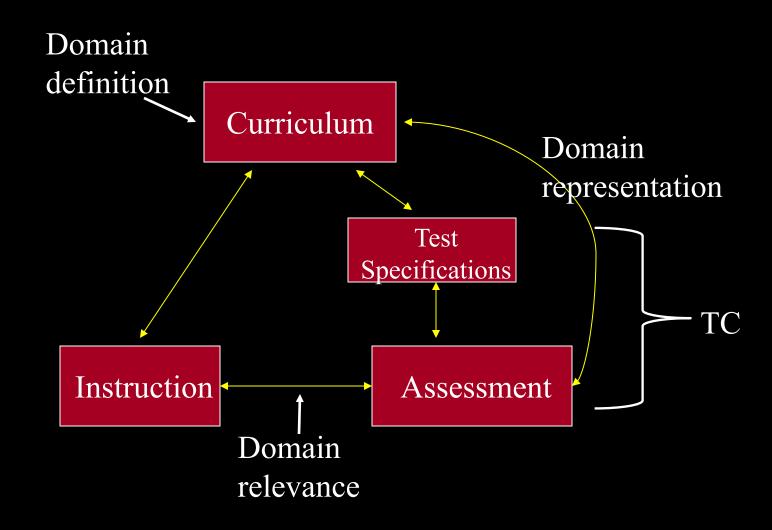
Content Validity Questions

- 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



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

Educational Measurement



Educational Measurement: Issues and Practice July 2020, Vol. 0, No. 0, pp. 1–6

Standardization and UNDERSTANDardization in Educational Assessment

Stephen G. Sireci, University of Massachusetts Amherst, Amherst

Abstract: Educational tests are standardized so that all examinees are tested on the same material, under the same testing conditions, and with the same scoring protocols. This uniformity is designed to provide a level "playing field" for all examinees so that the test is "the same for everyone. Thus, standardization is designed to promote fairness in testing. In practice, the material tested, the conditions under which a test is administered, and the scoring processes, are often too rigid to provide the intended level playing field. For example, standardized testing conditions may interact with personal characteristics of examinees that affect test performance, but are not construct-relevant. Thus, more flexibility in standardization is needed to account for the diversity of experiences, talents, and handicaps of the incredibly heterogeneous populations of examinees we currently assess. Traditional standardization procedures grew out of experimental psychology and psychophysics laboratories where keeping all conditions constant was crucial. Today, accounting for and measuring what is not constant across examinees is crucial to valid construct interpretations. To meet this need I introduce the concept of understandardization, which refers to ensuring sufficient flexibility in standardized testing conditions to yield the most accurate measurement of proficiency for each examinee.

Keywords: culturally responsive assessment, educational testing, scaling, standardization, test accommodations, validity

Educational Measurement



Educational Measurement: Issues and Practice March 2023, Vol. 0, No. 0, pp. 1–7

Personalizing Large-Scale Assessment in Practice

Heather M. Buzick, Jodi M. Casabianca, and Melissa L. Gholson, Educational Testing Service

Abstract: The article describes practical suggestions for measurement researchers and psychometricians to respond to calls for social responsibility in assessment. The underlying assumption is that personalizing large-scale assessment improves the chances that assessment and the use of test scores will contribute to equity in education. This article describes a spectrum of standardization and personalization in large-scale assessment. Informed by a review of existing theories, models, and frameworks in the context of current and developing technologies and with a social justice lens, we propose steps to take, as part of assessment research and development, to contribute to the science of personalizing large-scale assessment in technically defensible ways.

Keywords: equity, large-scale assessment, measurement, personalization

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



Click here Sports



Click here Food



Click here Something else



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

UMass Football



Click here UMass Women's Basketball



Click here Psychometricians playing ping pong



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



Generating items to ensure comparability across test variations

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.

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 (83), Grade Level 3.9

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

