From Test Development to Test Use Consequences: What Roles does the CEFR Play in a Validity Argument?

Dorry M. Kenyon
Center for Applied Linguistics
Washington, DC, USA

2014 EALTA Conference  Plenary
May 31, 2014
University of Warwick, Coventry, England
Outline

1. Introduction
2. An Integrated Validation Argument Framework
3. Roles of the CEFR in the Framework
4. Summary
5. Resources
6. Discussion
Outline

1. Introduction
2. An Integrated Validation Argument Framework
3. Roles of the CEFR in the Framework
4. Summary
5. Resources
6. Discussion
Where am I Coming From?

- Intersection of three hats
  - Project Director
  - Applied Linguist
  - Psychometrician
My Personal Goals

Improve and professionalize the practice of language testing on whatever level or whatever scale it occurs—classroom, school, country, internationally

To encourage all language testing professionals to know why they are doing what they are doing

To provide tools for language testing professionals to think about the complex issues they face
Personal Experience with CEFR?

With CEFR and Test Development Directly: Minimal until January 2014

With CEFR as an “Observer”: Quite a lot, especially through a series of international meetings exploring how to align the Proficiency Guidelines of the American Council on the Teaching of Foreign Languages (ACTFL) and the CEFR
Experience with Verbally-defined Proficiency Level Descriptions?

With Test Development Directly:

ACTFL Speaking Proficiency Guidelines *since 1987*

U.S. Gov’t Interagency Language Roundtable (ILR) Performance Level Descriptions *since 1987*

Student Performance Levels (U.S. Adult Basic ESL) *since 1988*

National Reporting System (U.S. Adult Basic Education) *since 2002*

WIDA Performance Level Descriptors (Kindergarten to Grade 12 Academic English Learners) *since 2003*

Dissertation Research *(1995)* *Linking Multiple-Choice Test Scores to Verbally-Defined Proficiency Levels: An Application to Chinese Reading Proficiency* (attempt to use an empirical approach through the LLTM)

Since *2002*, have conducted over 25 “standard setting” studies to link student performances on language assessments to verbally-defined proficiency level descriptions
Setting the Scene for Today’s Talk

Parsonage
First Parish Church
Grover’s Corners, New Hampshire

Jane Crofut
The Crofut Farm
Grover’s Corners
Sutton County
New Hampshire
United States of America
Continent of North America
Western Hemisphere
The Earth
The Solar System
The Universe
The Mind of God
Outline

1. Introduction

2. An Integrated Validation Argument Framework

3. Roles of the CEFR in the Framework

4. Summary

5. Resources

6. Discussion
Test Validation as an Organizing Principle

Samuel Messick (1989)

“Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.” (p. 13)

1. Consequences
   - A Beneficial

2. Decisions
   - A Values Sensitive
   - B Equitable

3. Interpretations
   - A Meaningful
   - B Impartial
   - C Generalizable
   - D Relevant
   - E Sufficient

4. Assessment Records
   - A Consistent

5. Assessment Performance
   - I Assessment Implementation
   - II Assessment Delivery

6. Design
   - I Conceptual Assessment Framework
     - IA Student Model
     - IB Evidence Model
     - IC Task Model
     - ID Assembly Model
   - II Domain Analysis
   - III Domain Modeling

7. Plan
   - State Anticipated Decisions/Consequences

Bachman and Palmer: Assessment Use Argument

Mislevy et al: Evidence Centered Design

Building a Validation Framework
© Center for Applied Linguistics 2014
1. Consequences
2. Decisions
3. Interpretations
4. Assessment Records
5. Assessment Performance
6. Design
7. Plan

I. Conceptual Assessment Framework
   IA. Student Model
   IB. Evidence Model
   IC. Task Model
   ID. Assembly Model

II. Domain Analysis
III. Domain Modeling

Bachman and Palmer: Assessment Use Argument
Mislevy et al: Evidence Centered Design
1. Consequences
   - Beneficial

2. Decisions
   - Values Sensitive
   - Equitable

3. Interpretations
   - Meaningful
   - Impartial
   - Generalizable
   - Relevant
   - Sufficient

4. Assessment Records
   - Consistent

5. Assessment Performance
   - Assessment Implementation
   - Assessment Delivery

6. Design
   - Conceptual Assessment Framework
     - Student Model
     - Evidence Model
     - Task Model
     - Assembly Model

7. Plan
   - State Anticipated Decisions/Consequences
Outline

1. Introduction
2. An Integrated Validation Argument Framework
3. Roles of the CEFR in the Framework
4. Summary
5. Resources
6. Discussion
Our Path

Examine each level of the validation framework

Highlight some of the most critical roles CEFR plays (if any) at each level

Comments are meant to be only *illustrative* and not exhaustive

My goal is to point out how a larger framework can be useful
1. Consequences

2. Decisions

3. Interpretations

4. Assessment Records

5. Assessment Performance

6. Design

7. Plan

Bachman and Palmer: Assessment Use Argument

Mislevy et al: Evidence Centered Design
What decisions vis-à-vis CEFR level(s) are anticipated to be made?
- CEFR level only, or
- **CEFR PLUS** (e.g., graduation?, entrance?, job?, qualification?)

What evidence will be needed to support decisions?
- About CEFR level only
- About **CEFR PLUS** (e.g., graduation?, entrance?, job?, qualification?)

What are the anticipated (positive) consequences?
- From interpretations based on CEFR levels
- From interpretations about **CEFR PLUS**

What potential negative consequences need to be mitigated?
Evidence Centered Design (ECD)

Years of solid thinking by Robert Mislevy and colleagues (originally at ETS)

Seminal explanations began appearing around 1999 (see references)
Layers in ECD

1. Domain Analysis
2. Domain Modeling
3. Conceptual Assessment Framework
   3.1 Student Model
   3.2 Evidence Model
   3.3 Task Model
   3.4 Assembly Model
4. Assessment Implementation
5. Assessment Delivery
What is “Domain Analysis?”

What substantive information about the domain needs to be collected?

Such as:

- Content
- Concepts
- Terminology
- Tools
- Representational forms
- Nature and type of knowledge in the domain
- Environment in which knowledge is gained
- Environment in which knowledge is demonstrated
7 (II) Domain Analysis and CEFR

While verbally-defined proficiency level descriptions seem to be a good place to start, they are not sufficient. Need to understand/interpret CEFR levels in relationship to the domain of what is being actually tested and the actual test use; for example…

How old are the students — 7, 17, 27?

How have the students acquired language; what have they been exposed to – homogeneity (e.g., national curriculum)? Great diversity?

How will the students be using language vis-à-vis the actual decision being made (target language use situation) – program evaluation? university entrance? job promotion?

What other understandings of second language acquisition (implicitly) underlie the language testing situation?
What is “Domain Modeling?”

How will the argument connecting observations of students’ actions to inferences about them be articulated?

- Based on Toulmin’s structure of arguments
Toulmin’s (1958) Argument Structure

- **Claim**
- **Warrant**
  - on account of
  - since
  - unless
- **Backing**
  - so
- **Data**
  - supports/refutes
- **Alternative explanation**
- **Rebuttal Data**
What evidence will support each claim in the chain of the validation framework?
1. Consequences
- A Beneficial
- A Values Sensitive
- A Consistent

2. Decisions
- A Values Sensitive
- B Equitable
- B Impartial

3. Interpretations
- A Meaningful
- B Impartial
- C Generalizable
- D Relevant
- E Sufficient

4. Assessment Records
- A Consistent

5. Assessment Performance

6. Design

7. Plan
- State Anticipated Decisions/Consequences

II Assessment Delivery
- I Assessment Implementation

III Domain Modeling
- II Domain Analysis

Mislevy et al: Evidence Centered Design

Bachman and Palmer: Assessment Use Argument

Building a Validation Framework
© Center for Applied Linguistics 2014
Going through this step at the beginning

(a) will help guide following design decisions, nature of research studies, and so on

(b) will help clarify what decision will actually be made on the basis of the assessment -- only for a CEFR level, or for **CEFR PLUS**
“Conceptual Assessment Framework?”

- How are the nuts and bolts of the operational assessment connected?
- That is, what coordinates the substantive, statistical, and operational aspects of an assessment?
Conceptual Assessment Framework

Mislevy and Riconscente, 2005, p. 17
Student Model

Mislevy and Riconscente, 2005, p. 17
What is the “Student Model?”

*What is it that the student “knows and can do” that will be assessed?*

(e.g., in terms of…
- knowledge,
- skills, or
- other attributes)
The “can-do” aspect of CEFR appears to provide a solid foundation to the “Student Model,” however generalizations can be deceptive and the “Student Model” needs to be worked out in accord with the domain analysis, particularly in the case of assessing **CEFR PLUS**.
Evidence Model

Mislevy and Riconscente, 2005, p. 17
What is the “Evidence Model?”

How is ‘what the student knows and can do’ going to be assessed, both

• in terms **observations** of student behavior, and

• in terms of a **measurement (statistical) model**?
Assuming a solid psychometric measurement model is chosen, a serious challenge remains in assessing reading and listening, particularly through selected response formats. We need evidence to support the link between *observations* of choosing correct or incorrect responses and the “can-do’s” of the CEFR (e.g., what does it mean “to understand?”). To what extent are student’s using test-dependent strategies that are not aligned with understanding aural language or written text outside the testing situation? What will count as an observation?

If *CEFR PLUS* is what actually is being assessed, we need ask ourselves the same questions vis-à-vis the actual domain of interest to establish a link.
Task Model

Mislevy and Riconscente, 2005, p. 17
What is the “Task Model?”

What **tasks** or **situations** should elicit the behaviors we need as evidence?
Our domain analysis and domain model should lead to tight task specifications that will provide a foundation for the types of student behavior we need to observe and thus of the types of tasks that *might* elicit such behavior. Several reminders:

(a) Start with the domain analysis and domain model, not with the task!

(b) Conduct a literature review to find out as much as you can about possible task types, but don’t just adopt one because it’s been widely used before!

(c) Tasks for assessing *CEFR PLUS*, where a target language use situation is well defined, are often easier than for CEFR level only.
Assembly Model

Mislevy and Riconscente, 2005, p. 17
What is the “Assembly Model?”

How much of observed student behavior do we need to reach our measurement goals?

How do we manage the interplay between the student, task and evidence models?
The amount of observed student behavior (i.e., evidence) needed may differ depending on whether the decision is CEFR level, or **CEFR PLUS**.

The amount of evidence needed will differ depending on the stakes of the decisions that will be made on the basis of the assessment, with higher stakes assessments generally requiring more evidence (and thus more confidence in making a correct decision) both from a content perspective and a psychometric perspective.
Assessment Implementation/Delivery

Mislevy and Riconscente, 2005, p. 17
Need to ensure that nothing in the assessment implementation or its delivery negatively detracts from the link between observations of student behavior and assessment records (next level up). For example, when is responding to a written prompt a measure of typing skills rather than of a (or in addition to) CEFR level?
1. Consequences

2. Decisions

3. Interpretations

4. Assessment Records

5. Assessment Performance

6. Design

7. Plan

Bachman and Palmer: Assessment Use Argument

Mislevy et al: Evidence Centered Design

Building a Validation Framework

© Center for Applied Linguistics 2014
Model of Rating Components

- Task
- Performance
- Test Taker
- Rubric
- Rater
- Rating

Scoring Conditions

Interpretations about Test Taker

Testing Conditions
Does everything in the assessment implementation/delivery and the production of assessment records align (e.g., task and rubric) in order to ensure

(a) that students have the maximum opportunity of providing behavior that provides evidence to support the decision to be made?

(b) that raters/scorers have the maximum opportunity to recognize and reward behaviors in the student performance that support a correct interpretation?

(c) that all students are scored consistently and fairly?
In light of the decisions to be made, what evidence supports the claim that the interpretations are …

*meaningful?*

*impartial to all examinees?*

*generalizable to the target language use domain?*

*relevant?*

*sufficient?*
Using “standard setting” methodologies to align student performances to a CEFR level may be best seen as collecting evidence to support the interpretation of assessment performances in terms of a CEFR level. In this case, it belongs to this level of the validation framework, perhaps under “meaningfulness.”

If the lower levels of the framework have been done properly (and documented!), materials needed for the standard setting panelists are pretty straightforward to assemble.

Evidence for other aspects of the interpretation should not be forgotten!
In the case of assessing **CEFR PLUS** (e.g., when the assessment is also being used as a criterion for another decision such as graduation, entrance to study, job, or a qualification), for higher stake decisions to be supported by evidence (e.g., that they are value sensitive and equitable), a more traditional standard setting focusing on language requirements as operationalized in the specific assessment may be required to provide evidence to support such decisions. If this decision step was built into the entire plan, then this should also be straightforward. Note that…

(a) stakeholders in this decision may not necessarily be language specialists who could best comprise the panel for the alignment study.

(b) without evidence from a traditional standard setting study, alternative strong evidence will be needed to support decisions based on the assessment results.
1 Consequences and CEFR

Is there evidence that the consequences of aligning test performances to the CEFR have in fact been beneficial? In what area? For whom?

Is there evidence that aligning test performances to the CEFR have had undesirable consequences? In what area? For whom? How would these be lessened?
Outline

1. Introduction
2. An Integrated Validation Argument Framework
3. Roles of the CEFR in the Framework
4. Summary
5. Resources
6. Discussion
(1) The use of a framework like this doesn’t necessarily add anything new, but it may be able to help us think through some important issues (e.g., CEFR level versus CEFR PLUS)
Most Important Take-Aways

(2) My B2 is not (quite) your B2, but that’s okay because…

my context is different

my students are different (e.g., age)

decisions made on the basis of the assessment are different (e.g., CEFR level versus **CEFR PLUS**)

and I have evidence that supports the use of the assessment and its claims to assessing in the B2 level
Most Important Take-Aways

(3) “Standard-setting” for alignment to verbally-defined proficiency levels is not quite standard-setting (e.g., WIDA case)

(a) How do scores and proficiency levels align? (Interpretation)

(b) When does a child have enough academic English language proficiency to exit from funded support services? (Decision)
Outline

1. Introduction
2. An Integrated Validation Argument Framework
3. Roles of the CEFR in the Framework
4. Summary
5. Resources
6. Discussion
On the Assessment Use Argument (AUA)


On Evidence Centered Design (ECD)

www.education.umd.edu/EDMS/mislevy/papers/ECD_overview.html


(4) Resources from the PADI project [padi.sri.com/](http://padi.sri.com/) , especially:

(padi.sri.com/downloads/TR9_ECD.pdf)


Outline

1. Introduction
2. An Integrated Validation Argument Framework
3. Roles of the CEFR in the Framework
4. Summary
5. Resources
6. Discussion
Questions/Discussion

Building a Validation Framework
© Center for Applied Linguistics 2014

1. Consequences
   A Beneficial
   Bachman and Palmer: Assessment Use Argument

2. Decisions
   A Values Sensitive
   B Equitable

3. Interpretations
   A Meaningful
   B Impartial
   C Generalizable
   D Relevant
   E Sufficient

4. Assessment Records
   A Consistent

5. Assessment Performance
   Assessment Implementation
   Assessment Delivery

6. Design

7. Plan
   State Anticipated Decisions/Consequences

CAL CENTER FOR APPLIED LINGUISTICS
Thank You Very Much for the Invitation!

Jane Crofut
The Crofut Farm
Grover’s Corners
Sutton County
New Hampshire
United States of America
Continent of North America
Western Hemisphere
The Earth
The Solar System
The Universe
The Mind of God