Test Validation for People with (Proficiency) Standards

Ray Clifford

ILR

15 February 2018

Language testing is a professional discipline.

- Language is the most complex of observable human behaviors.
- Testing is a complex science.
- Language testing is a professional discipline that requires scientific expertise.
- Language testing is also influenced by individuals’ philosophies.

Philosophies of Testing

- How many are there?

- Can you name them?

Glenn Fulcher

- Has described testers' radically different views in ontology (what they think they know) and epistemology (how they think they know it) as "fault lines" within the testing profession.
- He also provided a useful summary of the prevailing language testing philosophies.

"Philosophy and Language Testing," (pages 1-19) in *The Companion to Language Assessment*, First Edition. Antony John Kunnan, ed. John Wiley and Sons, 2014.

Fulcher's Summary of Testing Philosophies

1. **Realists**: The scientific method allows us to test everything.
   a. Extreme: We can measure language ability independently of any particular observer and testing method.
   b. Pragmatic: Theories must be testable, as simple as possible, coherent, and comprehensive.

Fulcher's Summary of Testing Philosophies

2. **Antirealists**: All tests influence behavior.
   - Constructivists: Our tests create the myth of a non-existent reality, and are used by those in power to subjugate the disadvantaged.
   - Instrumentalists*: If tests are useful, that is all that really matters.
3. ***Instrumentalists** are a.k.a. "nonrealists".
   - Instrumentalists don't have enough in common with constructivists to form a coalition.
   - Neither realists nor constructivists agree with instrumentalism.

Glenn Fulcher

- Argues that both of the extreme positions on the cline between the realists and the constructivists are untenable.
  - Naïve realists hold that the scientific method is equally applicable to physical things and to human beings.
  - Constructivists hold that, because of the transient nature of the social construction of meaning on an interaction-by-interaction basis, there is no such thing as general language proficiency.

Quiz: Who might make each statement?

- Constructs are not the same as traits.
- An instrument is valid if it is sensitive to the trait it claims to measure.
- Test developers are responsible for preventing any misuse of their tests.
- Language use is a social activity, so one’s language ability can’t be generalized.
- Only individuals with an enduring performable competence can engage in “co-construction” of meaning.

Who might have made each statement?

- Constructs are not the same as traits. Realists
- An instrument is valid if it is sensitive to the trait it claims to measure. Realists
- Test developers are responsible for preventing any misuse of their tests. Constructivists
- Language use is a social activity, so one's language ability can't be generalized. Constructivists
- Only individuals with an enduring performable competence can engage in "co-construction" of meaning. Realists

Given these philosophical fault lines, how can we best demonstrate the validity of ILR language proficiency tests?

Given these philosophical fault lines, how can we best demonstrate the validity of ILR language proficiency tests?

What does ILR testing have that none of these testing philosophies have?

Given these philosophical fault lines, how can we best demonstrate the validity of ILR language proficiency tests?

What does ILR testing have that none of these testing philosophies have?

Empirically-validated proficiency criteria!

ILR testers have empirically-validated criteria, and are more demanding than:

- **Constructivists**, because the ILR has defined the communication tasks, contexts, and accuracy expectations they want to measure.
- **Instrumentalists**, because ILR tests must do more than separate the best from the rest.
- **Realists**, because ILR tests must accurately assign specific ILR ratings rather than just being sensitive to increases in test takers' proficiency.

Freedom from these philosophical constraints opens the possibility of enlightened, “classical pragmatism”.

(Fulcher, p.2)

- ILR language testing is not a philosophical exercise that must infer an **unobservable** latent trait from a hypothesized nomological network of correlational relationships.
- ILR language testing is a scientific effort that describes an **observable** human trait by measuring an individual’s performance against defined Task, Conditions, and Accuracy expectations.

This Contrast is Pervasive

Philosophical Testing Research

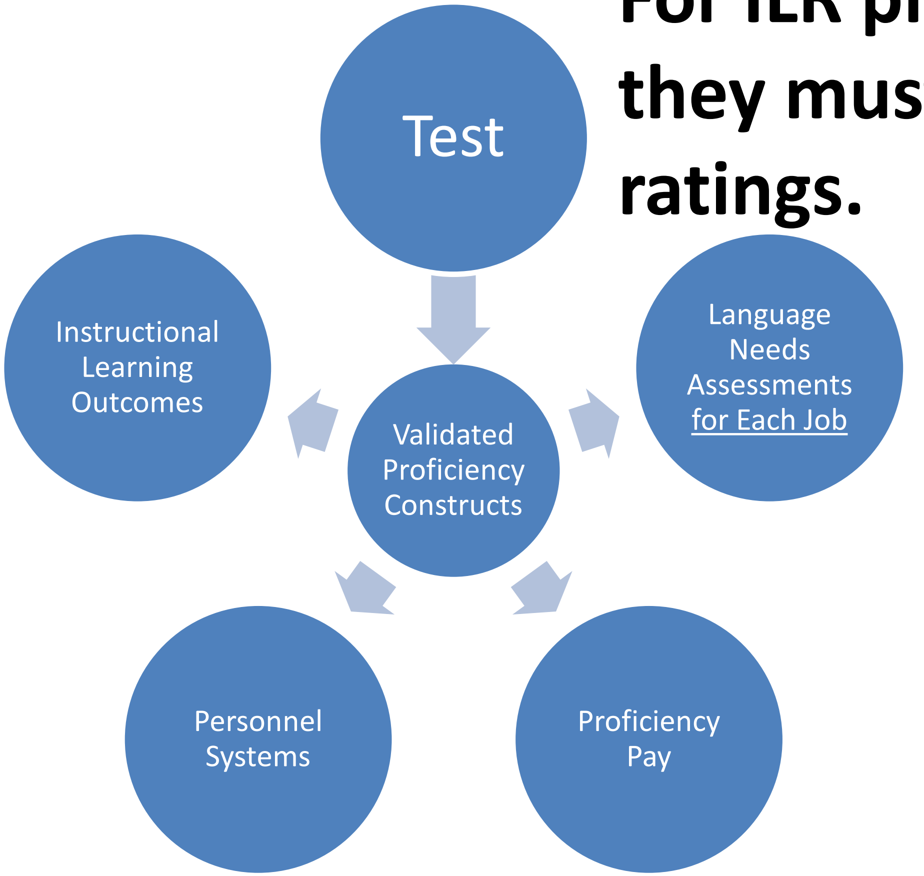
- Suggests underlying nomological networks.
- Hypothesizes correlations within that network.
- Infers an explanatory model.
- Develops arguments that defend the inferred explanation.

ILR Testing

- Measures observable phenomenon.
- Documents demonstrable relationships.
- Confirms the functioning of a validated ability model.
- Builds an evidence-based argument for that model.

This summary applies insights found in the article “The Concept of Validity” by Borsboom, Mellenbergh, and van Heerden in *Psychological Review,* 2004, Vol. 111, No. 4, 1061-1071.

ILR test developers link their test validation evidence directly to a validated proficiency model.


Test
Validated Proficiency Constructs
Instructional Learning Outcomes
Language Needs Assessments for Each Job
Personnel Systems
Proficiency Pay

For ILR proficiency tests to be valid, they must accurately assign ILR ratings.

It is not necessary to establish a validity argument for each of the possible uses for which proficiency ratings may be applied.

It is not necessary to validate an ILR proficiency test to a specific job or MOS. Technical job testing “owns” that performance link.

And in most cases, language proficiency is a necessary, but not a sufficient condition for job performance.

As early as 1962,

- Robert F. Mager was calling for performance objectives that described **“acceptable performance”** for observable behaviors.
  - *Preparing Instructional Objectives*, Palo Alta, CA. Feardon Publishers, 1962, p.44.
- In 1973, Hambleton and Novick pointed out that, **“Above all else, a criterion-referenced test must have content validity.”**
  - “Toward an integration of Theory and Method for Criterion-Referenced Tests,” *Journal of Educational Measurement*, Vol. 10, No. 3 (Fall 1973) p. 168.

DEPARTMENT OF STATE: MID-1950s (History provided by Pardee Lowe, Jr.)

- Needed to verify the foreign language skills of its employees.
- Surveyed possible approaches.
- Contacted Prof. John B. Carroll at Harvard.
- He suggested the first criterion-referenced (CR) test of observable foreign language ability.
- Based on Osgood's "semantic differential".
- Inherent in the scale were "yes, can do" or "no can't do" decisions.

In 1973,

- Hambleton and Novick pointed out that, **“Above all else, a criterion-referenced test must have content validity.”**
    - “Toward an integration of Theory and Method for Criterion-Referenced Tests,” *Journal of Educational Measurement*, Vol. 10, No. 3 (Fall 1973) p. 168.

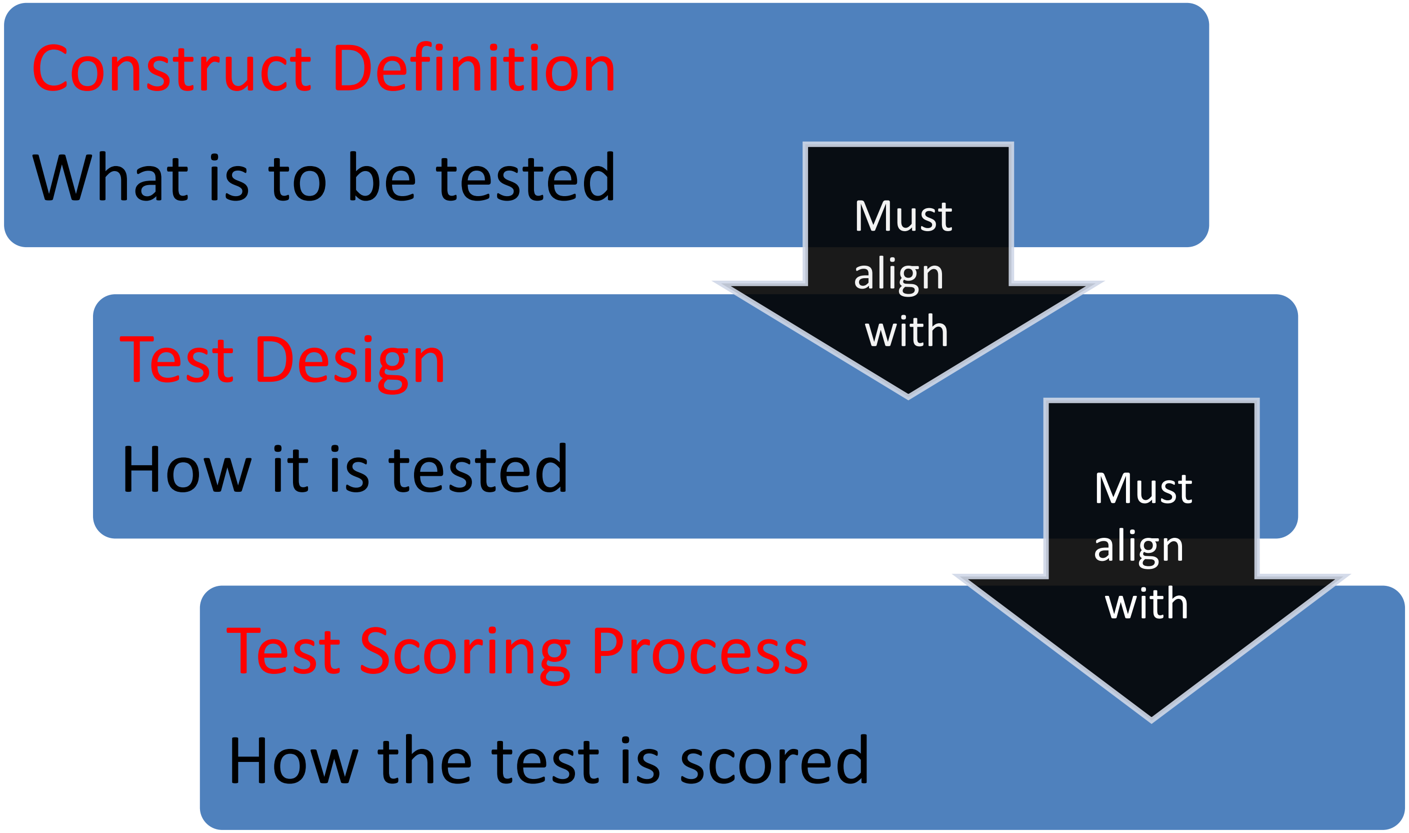
More recently, …

- Richard M. Luecht advised that linking constructs with test results
- **“is, fundamentally, a design problem that requires careful alignment of potentially incongruent models.”**
  - “Multistage Complexity in Language Proficiency Assessment: A Framework for Aligning Theoretical Perspectives, Test Development, and Psychometrics” *Foreign Language Annals*, Vol. 36, No. 4. (2013) p. 527.

Luecht's Alignment Establishes Provenance

- Observable testing constructs have Task, Context, and Accuracy (TCA) criteria.
- And these TCA elements must be aligned across all components of the test:
  - The construct.
  - The test design.
  - The scoring process.
- It is this alignment that establishes the **content validity** that Hambleton and Novick found essential for criterion-based tests.

With ILR testing, content validity requires full alignment with the TCA criteria of the ILR standard.



Construct Definition

What is to be tested

Must align with

Test Design

How it is tested

Must align with

Test Scoring Process

How the test is scored

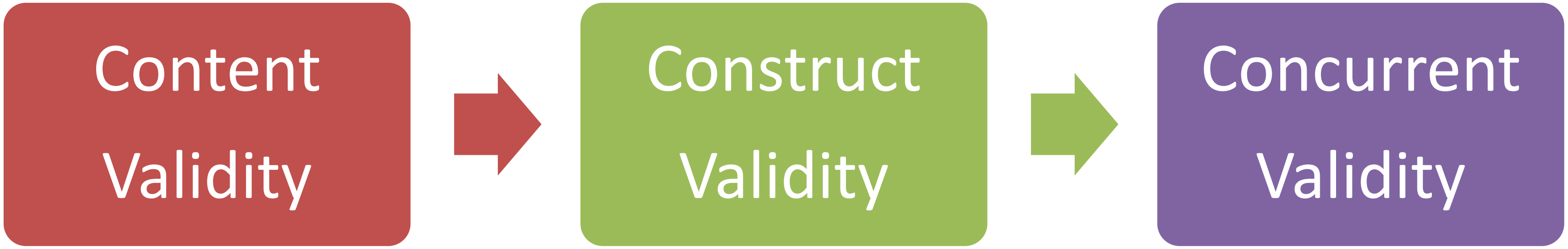
A Non-ILR Example of Non-Alignment

These quotes are from “Does an Argument-Based Approach to Validity Make a Difference?” by Chapelle, Enright, and Jamieson in *Educational Measurement: Issues and Practice*, Spring 2010, Vol. 29, No. 1, pp. 3-13.

- “Multiple perspectives were brought to bear … in the TOEFL project as the developers attempted to define a construct of academic English language proficiency ….”
- “A strong construct theory … within a nomological network … did not result from the process and therefore the construct itself was not a good basis for subsequent research.”
- “However, Kane’s organizing concept of an ‘interpretive argument’ which does not rely on a construct, proved to be successful.”

For ILR testing, content validity is an absolute prerequisite before the test can have construct validity (and concurrent validity).

1 2 (3)

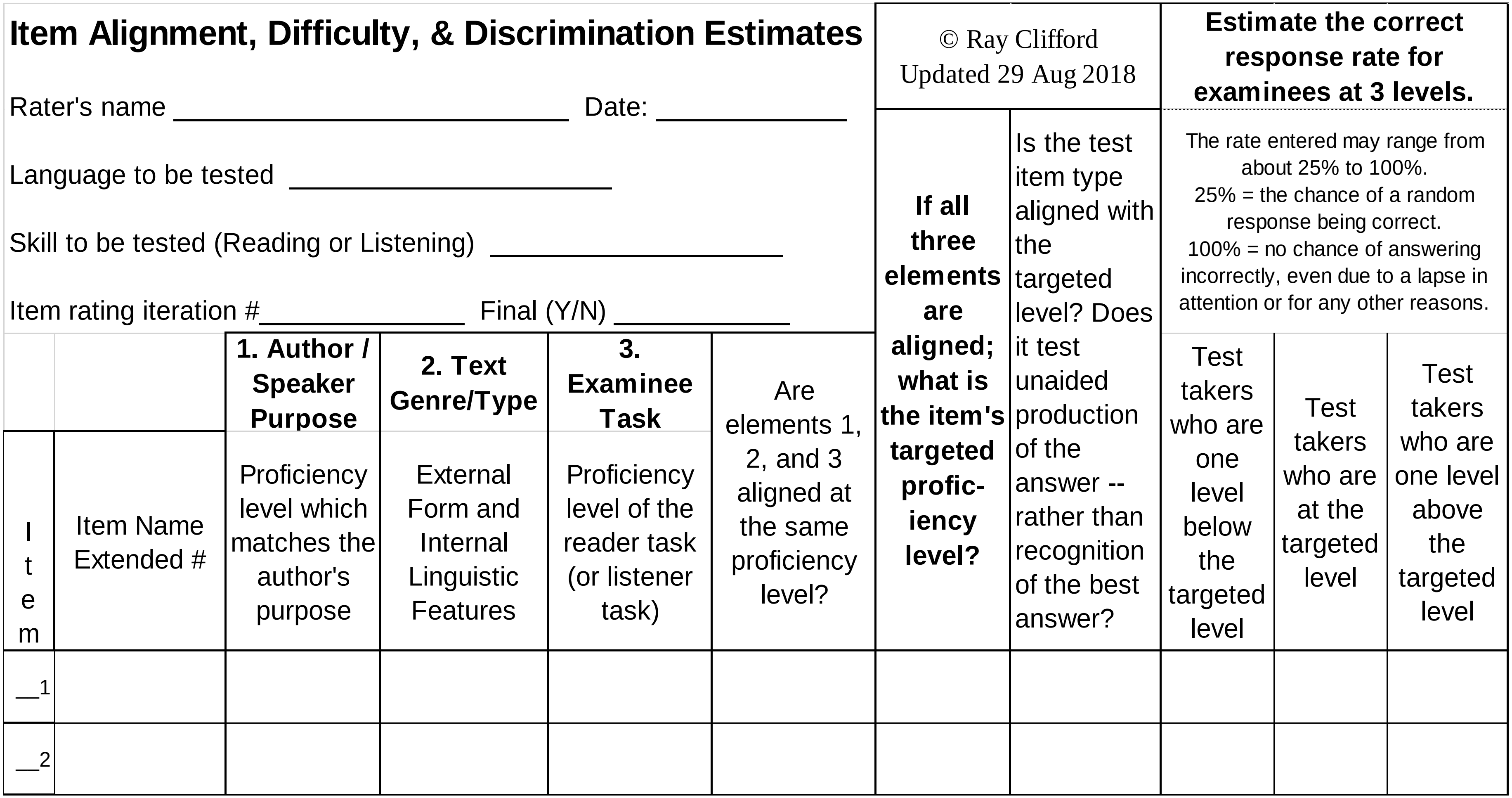


Content Validity

Construct Validity

Concurrent Validity

Judges' Content Rating Sheet



Item Alignment, Difficulty, & Discrimination Estimates

Rater's name ____________________________ Date: _____________

Language to be tested _______________________

Skill to be tested (Reading or Listening) _____________________

Item rating iteration #______________ Final (Y/N) ____________

© Ray Clifford
Updated 29 Aug 2018

Estimate the correct response rate for examinees at 3 levels.

The rate entered may range from about 25% to 100%.
25% = the chance of a random response being correct.
100% = no chance of answering incorrectly, even due to a lapse in attention or for any other reasons.

		1. Author / Speaker Purpose	**2. Text Genre/Type**	**3. Examinee Task**						
Item	Item Name Extended #	Proficiency level which matches the author's purpose	External Form and Internal Linguistic Features	Proficiency level of the reader task (or listener task)	Are elements 1, 2, and 3 aligned at the same proficiency level?	**If all three elements are aligned; what is the item's targeted profic-iency level?**	Is the test item type aligned with the targeted level? Does it test unaided production of the answer -- rather than recognition of the best answer?	Test takers who are one level below the targeted level	Test takers who are at the targeted level	Test takers who are one level above the targeted level
__1										
__2										

Content Validity

For construct validation,

ILR testers should see IRT statistics like these from an English Reading Test –

With 3 Levels & 4 testlets of 5 items each at each level; n = 680.

NATO Level	Logit value of Testlet A	Logit value of Testlet B	Logit value of Testlet C	Logit value of Testlet D	Standard Error of the model in Logits
3	+ 1.8	+ 1.8	+ 1.8	+ 1.8	.04
2	+ 0.3	+ 0.3	+ 0.3	+ 0.2	.04
1	- 1.5	- 1.6	- 1.6	- 1.7	.06

Content Validation

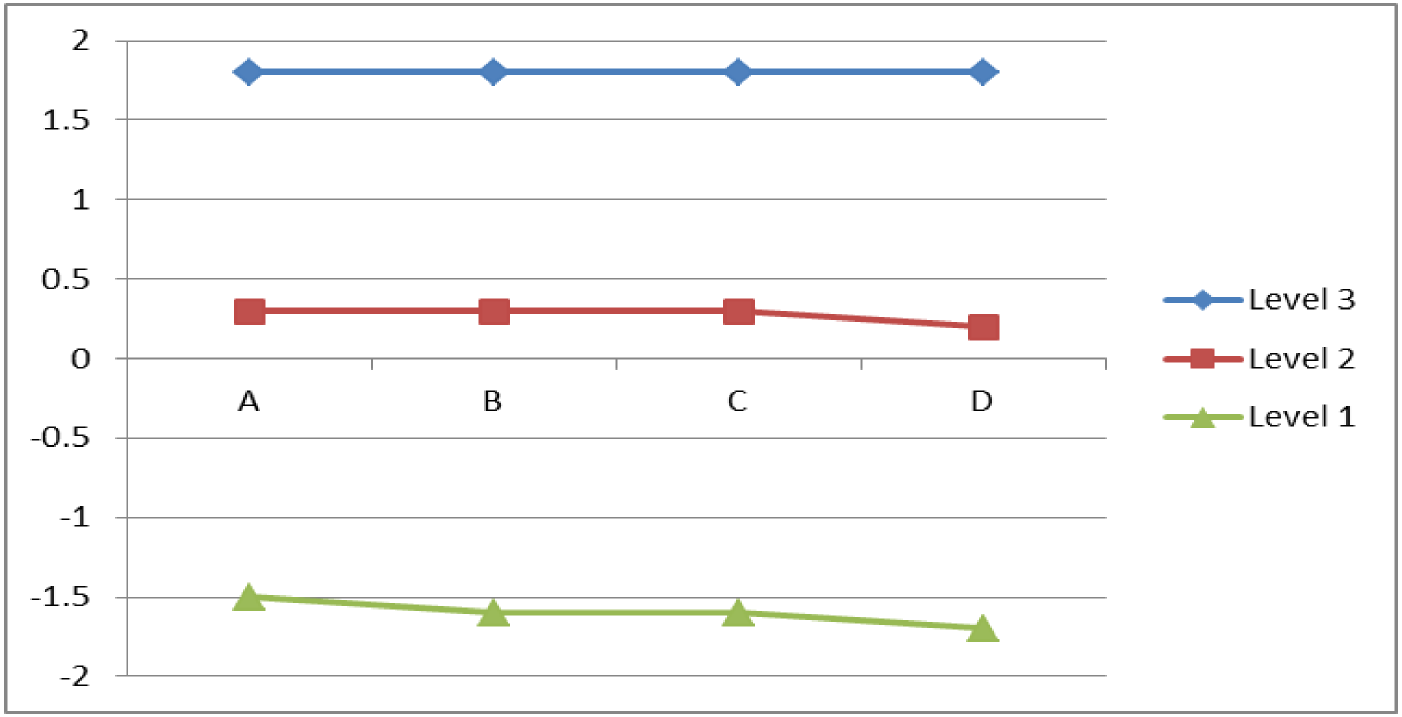
Construct Validation

A Clear Hierarchy of Difficulty

Testlet difficulties are within +/-.02 logits of each other.

Standard Error of Measurement < .06

Vertical distance between clusters > 1 logit



A Note on Scoring and Alignment

- In ILR language tests, only answers to fully-aligned test items can be interpreted.

What is a person's proficiency level?

- If s/he can't answer a Level 3 inference question about a Level 2 text?
- If s/he can answer a Level 3 inference question about a Level 1 text?
- If s/he can't answer a main idea question about a Level 3 text?
- What is s/he can answer a main idea question about a Level 3 text?

Content Validation

Construct Validation

Obtaining direct evidence of an ILR test’s validity requires full alignment, including “floor-and-ceiling” scoring.

What: = How Tested: = How Scored:

ILR	Aligned Test Design	A Score for Each Level	&	Pick the Highest Level Passed
Level 3	Level 3 Test	Level 3 Score		Passed at 3 ?
Level 2	Level 2 Test	Level 2 Score		Passed at 2 ?
Level 1	Level 1 Test	Level 1 Score		Passed at 1 ?

Content Validation → Construct Validation

Some Realities of Language Learning

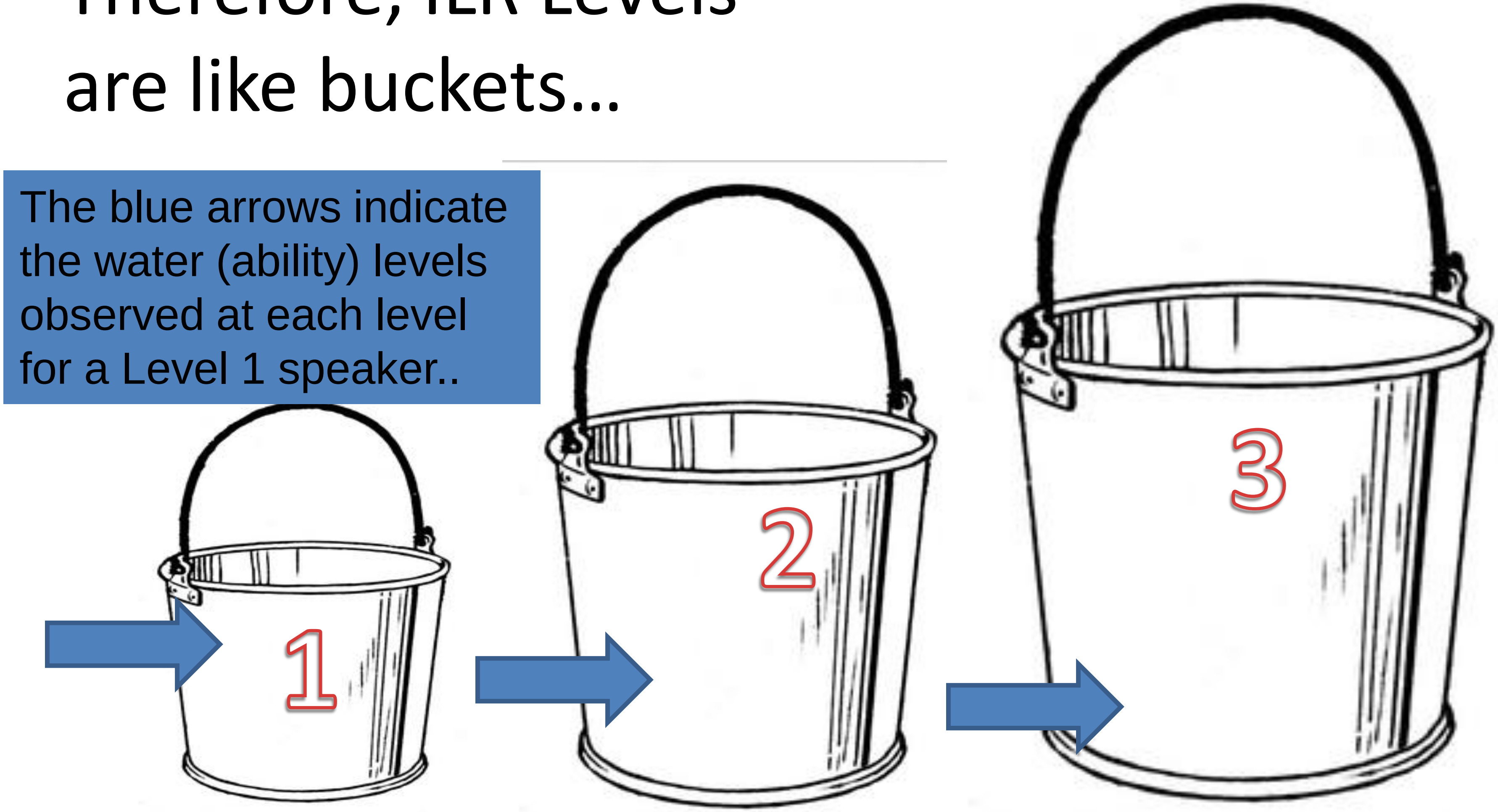
1. Language learners do not completely master one proficiency level before they begin learning the skills described at the next higher level.
2. Usually, learners develop conceptual control or even partial control over the next higher proficiency level by the time they have attained sustained, consistent control over the lower level.

Content Validation

Construct Validation

Therefore, ILR Levels are like buckets…

The blue arrows indicate the water (ability) levels observed at each level for a Level 1 speaker..



Some Level 2 skills will develop before Level 1 is mastered and some Level 3 skills will develop before Level 2 is mastered.
But the buckets will still reach their full (or mastery) state sequentially.

Content Validation → Construct Validation

Not having separate scores for each level can lead to inaccurate results.

- An ILR proficiency rating of a 1, a 2, or a 3 can only be assigned if **all** of the Task, Condition, and Accuracy expectations associated with that level have been met.
- Therefore, it is more important to know how much “water” (ability) is in each bucket, than it is to know how much “total water” there is in all three buckets.
- Let’s look at a simplified example...

Which learner is more proficient?

- Alice received a **TOTAL** score of 35 or **58**% on a 3-level test.

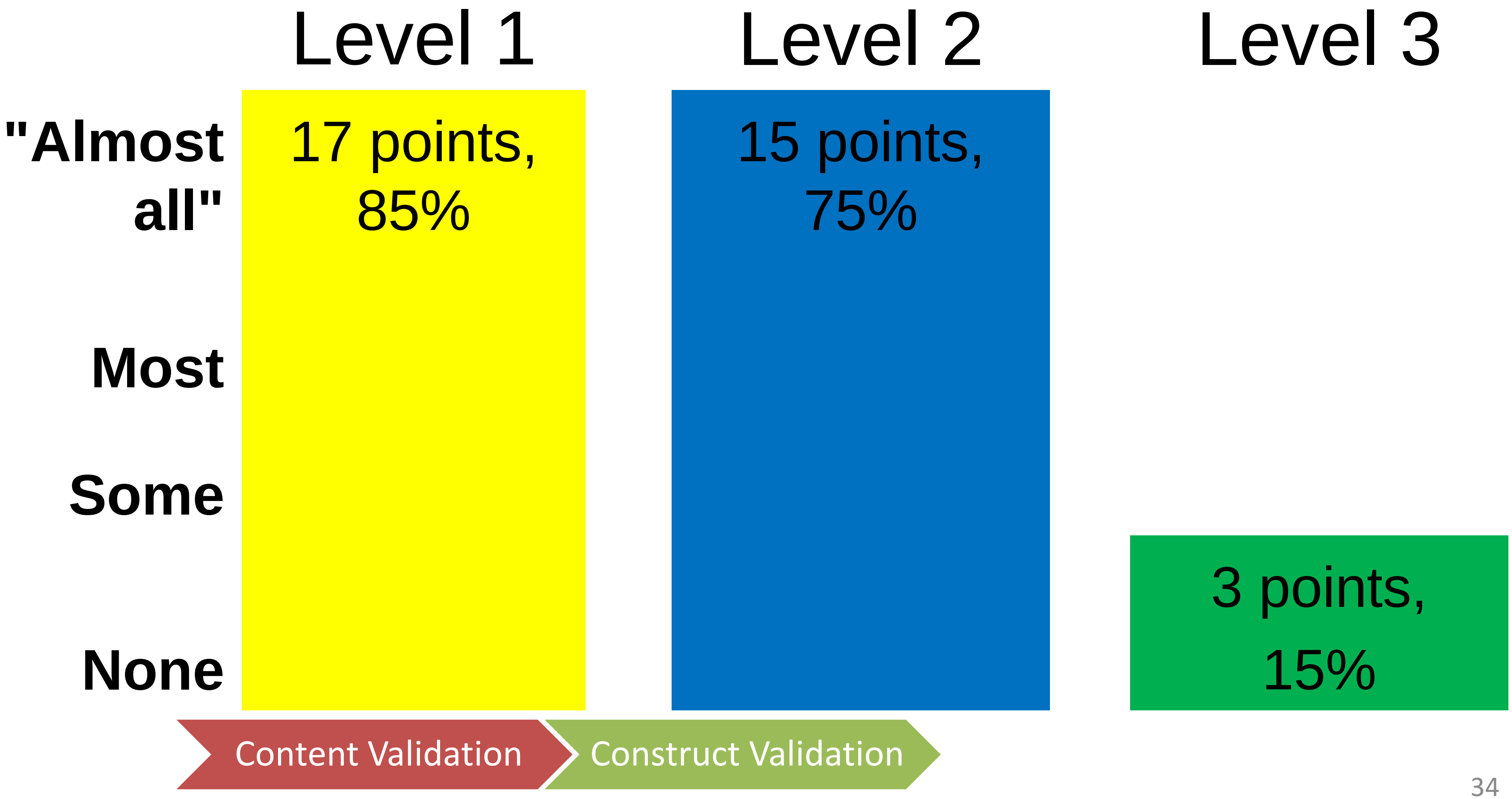
- Bob received a **TOTAL** score of 37 or **62%** on the same 3-level test.

- Before we decide, let’s look at their level-by-level test scores.

Example A: Alice’s total score = **58%**

ILR Proficiency Level = 2

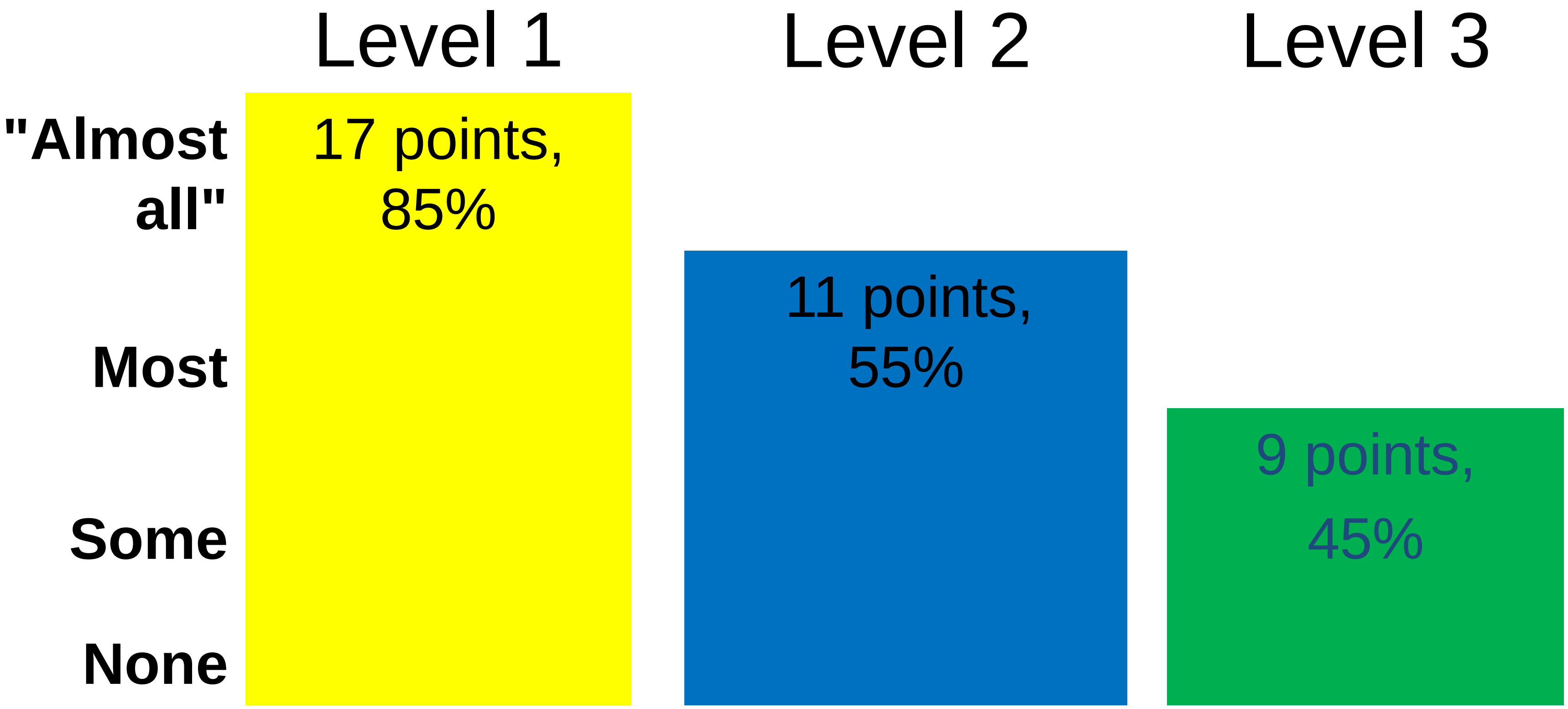
More precisely, Level 2 with Random abilities at Level 3



Example B: Bob’s total score = **62%**

C-R Proficiency Level = 1+

More precisely, Level 1 with Developing abilities at Level 2



Content Validation → Construct Validation

Which learner is more proficient?

- **Alice is more proficient.** She received a TOTAL score of 35 (58% overall), but met the Criterion-Referenced, TCA requirements for both **Level 1 and Level 2** !
- Bob received a total score of 37 (62% overall), but he only fully satisfied the TCA requirements for **Level 1 – and not for level 2.**

Content Validation

Construct Validation

Summary

- ILR tests measure observable behaviors.
- Test takers' communication behaviors are judged against a validated hierarchy of ILR Task, Condition, and Accuracy (TCA) requirements.
- Creating ILR tests requires a full alignment of the ILR TCA criteria across both the test design and the scoring protocols.
- In valid tests, the test items will have a separate item difficulty cluster for each ILR level being tested – and those clusters will not overlap in difficulty.

Questions to Consider

1. Do all ILR tests align with the ILR Task, Condition, and Accuracy criteria?
2. Do all ILR scoring protocols align with the ILR Task, Condition, and Accuracy criteria?
3. Do the items targeting each ILR level cluster in difficulty?
4. Do those clusters produce the expected non-overlapping hierarchy of Difficulty?
5. If not, why not?

For more information on the empirical validation of the ILR (and the derivative ACTFL and NATO) testing criteria, see:

Clifford, R. & Cox, T. (2013) “Empirical Validation of Reading Proficiency Guidelines”, *Foreign Language Annals*. Vol. 46, No. 1.

Cox, T. & Clifford, R. (2014) “Empirical Validation of Listening Proficiency Guidelines”, *Foreign Language Annals*. Vol. 47, No. 3.

Clifford, R. (2016) “A Rational for Criterion-Referenced Proficiency Testing”, *Foreign Language Annals*. Vol. 49, No. 2.

Thank you for coming to this
non-standard presentation
on test validation designed for
people with standards.

I hope you don’t leave feeling testy.