Validity and validation of language tests

Rie KOIZUMI (Tokiwa University)
rie-koizumi@mwa.biglobe.ne.jp
July meeting of JALT-Gunma
Sunday, July 20, 2008
at Maebashi Institute of Technology

Today’s presentation

- Validity
- History of conceptualizing validity
- Messick’s (1996) validity framework
- Guidelines to enhance the content aspect
Purposes of language testing

Bachman (1990, p. 13):

(1) In development and use of language tests, to assure that the information, or scores obtained from language tests will be reliable, valid, and useful

(2) In language testing research,
   (a) to formulate and empirically validate a theory of language test performance
   (b) to demonstrate the ways in which performance on language tests is related to communicative language use in its widest sense

Three important concepts

Validity
   - Generally speaking, to what degree a test measures the ability it intends to measure (i.e., test construct) (e.g., Henning, 1987)

Reliability
   - Consistency of rank-order of test-takers

Practicality
   - Feasibility of test
Validity

To be exact, “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” (Messick, 1989, p. 13)

Validation: process of establishing the validity

The most important aspect in test development and use (e.g., Bachman, 1990)

It should always be examined (e.g., Chapelle, 1999).

History of conceptualizing validity

Criterion-related validity

Need to find a criterion (a highly valid test to be compared)

High correlations suggest high validity.

- Concurrent validity
- Predictive validity

Problems: Difficult to find a good criterion
Content validity

- A highly valid test has
  - Content relevance to intended domains
  - Content representativeness
- Problems:
  - Tend to be subjective
  - Tend to have confirmatory bias

Construct validity (1)

- Proposed by Cronbach & Meehl (1955)
  - A highly valid test assess what test developers intended to assess
  - Convergent validity
    - Correlations should be higher when similar abilities are assessed.
  - Discriminant (divergent) validity
    - Correlations should be lower when different abilities are assessed.
Construct validity (2)

- Make hypotheses based on theories and examine them empirically
- “Validity when the other two types of validity cannot be examined” to “Central status in examining validity”
- Include reliability as well as content validity, criterion-related validity

“There are three types of validity (trinitarian view).” to “Validity is a unitary concept but can be divided into aspects. We need to present pieces of evidence for validity.”

“Just to examine one validity type is enough.” to “We need to present multiple pieces of evidence for validity comprehensively.”

Need to show intended interpretations and uses and then to evaluate them
Messick (1989)

- Established the concept of validity
- Include value implications and social consequences of tests in validity inquiry
- A matter of degree, not a “all or nothing” concept
- Validity varies depending on how we interpret and use test scores. Not a test property
- Need to check validity from test development stages until the test is no longer used

Messick’s (1996) validation framework

- Used in several studies (e.g., Chapelle, 1994; Guerrero, 2000; Hasselgren, 2005; Koizumi, 2005; Miller & Linn, 2000).

  - (1) Content aspect [content validity]
  - (2) Substantive aspect [construct validity]
  - (3) Structural aspect [construct validity]
  - (4) Generalizability aspect [reliability]
  - (5) External aspect [criterion-related validity]
  - (6) Consequential aspect [(new)]
Steps for validation

- Chapelle (1999)
  - (a) Make hypotheses about test results
  - (b) report the results in relation to the hypotheses, as evidence for validity
  - (c) make a validity argument by integrating evidence and rationales

Principles of validation

- Present intended interpretation before examining validity
- Showing one piece of validity evidence is not enough.
- Think of alternative interpretations and show evidence that they are not plausible
Example: Validation of a diagnostic test of grammar

- EDiT Grammar Test (English Diagnostic Grammar Test)
- Constructed in cooperation with researchers on SLA and language testing and secondary school teachers, with the aid of the organization called English Language Proficiency Assessment (NPO ELPA)

Main target: Japanese beginning learners of English (8th to 12th year students)

- 72 multiple-choice items
- Intended construct: understanding of “boundaries and internal structures of basic English noun phrases (NPs)”
- Example: Yesterday the window near Ken’s desk was broken.

- Based on Pienemann (1998) and teachers’ experience
- 18 items for each group, 4 sections
Table 1  *Noun Phrase Types According to the Internal Complexity*  

<table>
<thead>
<tr>
<th>Structure</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1  det + head noun</td>
<td><em>my brother / a sweater</em></td>
</tr>
<tr>
<td>Group 2 (det) + premodifier +</td>
<td><em>my green sweater / a boy’s name</em></td>
</tr>
<tr>
<td>head noun</td>
<td></td>
</tr>
<tr>
<td>Group 3  NP (Group 1/Group 2) +</td>
<td><em>a surprise party for my brother</em></td>
</tr>
<tr>
<td>PP</td>
<td></td>
</tr>
<tr>
<td>Group 4  NP (Group 1/Group 2) +</td>
<td>*chances to go to the parties / a couple</td>
</tr>
<tr>
<td>to-infinitive; NP (Group 1/</td>
<td>having a dinner time</td>
</tr>
<tr>
<td>Group 2) + present/past participle</td>
<td></td>
</tr>
<tr>
<td>Group 5  NP (Group 1/Group 2) +</td>
<td><em>things that I bought yesterday</em></td>
</tr>
<tr>
<td>relative clause</td>
<td></td>
</tr>
</tbody>
</table>

**Examples of test items**

**I. Example for matching items (construct: understanding of internal structures)**

- **Instruction:** Choose the most appropriate English phrase for the underlined part of the Japanese sentence.

- **Japanese:** 丘の上の白い建物は、私の学校です。（*The white building on the hill is my school.* [Group 3])

- **Choices:**
  1. the hill on the white building
  2. *the white building on the hill*
  3. the white on the hill building
  4. on the hill the white building
II. Example for making-boundaries items (construct: understanding of boundaries)

Instruction: Where do you make a boundary [boundaries] in a sentence, considering the meaning chunk?

Q: The (1) tall (2) girl speaking (3) to Ken *(4) is my (5) sister. [Group 4]

III. Example for items of inserting be verb or a noun (construct: understanding of boundaries or internal structures)

Instruction: Insert the given word(s) into a sentence.

Q: This (1) book (2) about (3) China *(4)* old (5). <is> [Group 3]
IV. Example for question-making items (construct: understanding of boundaries and internal structures)

Instruction: What is the most appropriate interrogative sentence for the given declarative sentence?
The boy who is playing the piano is Ken. [Group 5]

Choices:
1. *Is the boy who is playing the piano Ken?
2. Who is the boy playing the piano is Ken?
3. Is the boy who playing the piano is Ken?
4. Is the boy is who playing the piano Ken?

Diagnostic information provided to teachers and students

① Grammar level: 1 to 5  
(based on items on Groups 2 to 5)

② NP understanding level: Boundaries OK? Internal structures OK? Both OK?

③ Error characteristics: Effects of L1, difficulty in noun and preposition combination, etc

①: Mainly for students; ② & ③: For teachers
(1) Content aspect

- Examine the content of the test in relation to the test specifications
  - Are the test target domains (or construct) clearly defined?
  - Do test items or tasks match the construct?
  - Is the content of test items or tasks relevant to the test target domains (or construct)?
  - Is it representative?
  - Is the technical quality of the test good? (e.g., clear instructions, appropriate difficulty level)

- Method examples:
  - Analysis of the content (e.g., needs analysis, task analysis, curriculum analysis)
  - Expert professional judgement, Use of analytical frameworks

(1) Content aspect in EDiT Grammar Test

- Hypothesis (H) 1: The construct is clearly defined.
- H2: The test items match the construct.
- H3: The content of test items is relevant to the construct.
- H4: The content of test items is representative.
- H5: The technical quality of the test is good.

- Method used: Expert professional judgement of test development members
- Results: Supported
- Future path: Ask outsiders to examine this aspect
(2) Substantive aspect

Examine how test-takers respond to the items or tasks

- Do test-takers use the processes intended by test developers?
- Do response consistencies or performance regularities reflect domain processes?
- To what degree do relevant factors affect item difficulty and discrimination?
- What does a test appear to measure to the untrained eye?

Method examples:
- Verbal report, observation, questionnaire, & interview
- Discourse analysis

(2) Substantive aspect in EDiT Grammar Test (a)

- H6: The proportion of test items that elicit the intended test-taking processes is high.
- H7: Test-takers on average say the test assesses more grammatical knowledge than reading ability and vocabulary knowledge. (Gra > Read = Voc)

Method used:
- H6: Verbal protocol analysis (7 college students)
- H7: Questionnaire (n = 130; 8th to 9th year)
(2) Substantive aspect in Edit Grammar Test (b)

Results of Questionnaire:
- H6 was partially supported.
- 64% of the items elicited the intended test-taking processes.
- Problems indicated:
  - Answers were reached when the intended knowledge was not used.
  - Inserting item: He (1) wrote (2)* of (3) a (4) teacher. (the name)
    - Test-taker said, “(2), because something is missing after wrote.” (Intended answer: “The name of a teacher is a chunk.”)

(2) Substantive aspect in Edit Grammar Test (c)
- Answers were NOT reached when the intended knowledge was demonstrated.
- Question-making item: The boy who is playing the piano is Ken. [Group 5]
  - 1. Is the boy who is playing the piano Ken?*
  - 2. Who is the boy playing the piano is Ken?
  - 3. Is the boy who playing the piano is Ken?
  - 4. Is the boy is who playing the piano Ken?
    - The translation is perfect, but a test-taker did not select the right answer (see Slide 21 for the example).

Future path: Problematic items will be excluded. A new test version will be created. Similar analysis is required for new items. We aim at making item bank by linking new and old versions in order to shorten the test length
(2) Substantive aspect in EDiT Grammar Test (d)

- Results of Verbal protocol analysis:
  - H7 was partially supported.
  - H7: \(\text{Gra} > \text{Read} = \text{Voc}\)
  - Results: \(\text{Gra} = \text{Read} > \text{Voc}\)
  - Students’ reactions: Too many items; Difficult words and sentences
- Future path:
  - Use of easier words and simpler sentences. Conduct similar analysis

(3) Structural aspect

- Examine internal structure of test responses
- What are relationships among responses to the tasks, items, or parts of the test? Are they consistent with the internal structure of an intended construct?
- What is the observed dimensionality of response data? Is it consistent with the dimensionality of an intended construct?
- Are scoring criteria and rubrics supported by the theory?
- Is the internal structure of the assessment (e.g., score scales, scoring models) consistent with the internal structure of the intended construct?
- Method examples
  - Reliability, item response theory (IRT),
  - Factor analysis, & structural equation modeling
(3) Structural aspect in EDiT Grammar Test

- H8: The internal the structure is unidimensional because the test is intended to assess grammatical knowledge overall.
- Method used: Item response theory (IRT), examining fit ($n = 740$)
- Results: H8 was supported.
  - A high percentage of items (100%) and test-takers (99%) fit the Rasch model.

(4) Generalizability aspect

- Examine differences in test processes and structures over time, across groups and settings, and in response to experimental interventions
  - To what extent can we get the same scores and make the same interpretations across groups and settings?
  - Is there evidence that the score interpretation can be generalizable to the construct domain?
- Method examples
  - Reliability, generalizability theory
  - Differential item functioning (DIF)
  - Multifaceted Rasch analysis
(4) Generalizability aspect in EDiT Grammar Test

- H9: The generalizability coefficients (i.e., reliability) of the test are high for a low-stakes test ($\Phi = .60$).
  - Method used: multivariate generalizability theory ($n = 740$)
  - Results: H9 was partially supported.
    - Enough reliability for each section (.77 to .84) and group (.75 to .81).
    - More items are needed for detailed diagnosis (.34 to .72).
    - Example: Item of inserting *be* verb in Group 3
      - $\Phi = .30$ (6 items); 20 items needed to achieve .60
  - Future path: No. of problematic items will be increased.

(5) External aspect

- Examine relationships of the test scores with other measures and background variables (i.e., the test’s external structure)
  - Can the external patterns of correlations be explained by the intended construct?
  - Are empirical relationships with criterion measures, or the lack of relationships that are consistent with the intended construct?
- Method examples
  - Correlations, convergent and discriminant evidence from multitrait-multimethod (MTMM) comparisons
  - Factor analysis, & structural equation modeling
  - Analysis of variance (ANOVA)
(5) External aspect in EDiT Grammar Test

- H10: There is a relationship between the scores and the ability levels (school year).
- Method used: ANOVA ($n = 740$)
- Results: H10 was partially supported.
  - Intended: $8^{th} < 9^{th} < 10^{th} < 11^{th}$
  - Results: $8^{th} < 9^{th} = 10^{th} = 11^{th}$
    - (no ceiling or floor effects)
- Future path: Detailed analysis of response patterns is needed. The fact that gradual increase was not seen may suggest the diagnosis of this test is necessary.

(6) Consequential aspect

- Examine social consequences of interpreting and using the test scores in particular ways
  - Are there value implications of score interpretation?
  - Are there potential consequences of test use? Are they the intended outcomes or unintended side effects? Are there positive or negative washback effects on teaching and learning?
- Method examples
  - Observation, interview, or questionnaire
  - Comparison between pretest and posttest scores
  - Analysis of teaching materials and documents
(6) Consequential aspect in EDiT Grammar Test

- H11: Test-takers on average say the test has a positive washback on learning.
- Method used: questionnaire (5 Likert scale; \( n = 130 \))
- Results: H11 was partially supported.
  - Example: Was the test useful for your learning?
  - “5” means “Yes, very much.” \( M = 3.38, SD = 1.08 \)
- Future path: A new version will be analyzed again. Need to obtain teachers’ responses

Validity argument of EDiT Grammar Test

- The hypotheses were generally supported. The interpretation and use of EDiT Grammar Test are at least moderately valid.
- Some unpredicted results suggest that it is necessary to revise the test further.
Characteristics of Messick’s framework

- **Strength**: A comprehensive framework to conceptualize six aspects of validity to be examined in validation

- **Weakness**
  - (1) There is no detailed guideline for a relative importance of validity aspects and how to advance validation. Thus, those who conduct validation may feel that there is too much work to do and it is beyond their task (Guion, 1995)
  - (2) Six aspects of validity may be treated in the same way as three validity types (Guion, 1995, p. 25).

Guidelines to enhance the content aspect

- **Write a test specification**
  - Read & Chapelle (2001): Inference, use, and impact should be specified in test purposes.
- **Train item writers**
- **Being very critical**
Role of test specifications in test development stages

- Davidson & Lynch (2002, p. 15)
- (1) Select skill
- (2) Write spec
- (3) Write item/task from spec
- (4) Assemble test and pilot/trial
- (5) Finalize operational measure
- Iterative processes for test revisions

Example of test specifications

- Davidson & Lynch (2002)
- (1) General description
  - General objectives
  - Specific objectives
- (2) Prompt attributes
- (3) Response attributes
- (4) Sample item
  - Directions
- (5) Specification supplement
Writing Multiple-choice items

- How to write good items (Haladyna, 2004, pp. 99-100)
- Also avoid the following (Carr, 2008)
  - Items probably answerable using background or word knowledge
  - Items only answerable using background knowledge
  - Items answerable using common sense
  - Items giving away the answers to other items
  - Overlapping items
  - Items with no single clear right answer
  - Items with overlapping options
  - Items with distractors that are true, or possibly true, but are not contained in the passage
  - Items with “odd man out” options
  - Construct-irrelevant items
  - Items merely requiring word matching instead of reading comprehension

Possible answers

- ① Items probably answerable using background or word knowledge
- ② Overlapping items
- ③ Items giving away the answers to other items
- ④ Items with overlapping options
- ⑤ Items with distractors that are true, or possibly true, but are not contained in the passage
- ⑥ Avoid opinion-based items.
- ⑦ Make distractors plausible.
- ⑧ Keep the length of options about the same.
After Messick (1)

- Proposals for validation procedures
  - Kane (2006)
    - Argument-based approach
  - Mislevy, Steinberg, & Almond (2002)
    - Evidence centered design
  - Bachman (2005)
    - Test utilization argument
  - Chapelle, Enright, & Jamieson (2008)

After Messick (2)

- Test fairness: Kunnan (2004)
- Proposals of code of ethics and practice (ILTA, n.d.; JLTA, 2006)
- Proposals to limit the range of validity
  - Social consequences are important but should not be included in validity. Validity is a test property. Causal relationships should be examined in validation. (Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2004)
Future inquiry

- Present a concept of validity and validation procedures in a simple way while maintaining theoretical coherence
- Cooperative test development: with teachers, content specialists, language testers, statisticians, etc.
- Explore and report distinct (context-based) validation procedures if there are Asian or Japanese testing cultures
- Recognize the expertise of language testing

References (1)


References (2)


Kanatani, K., & English Diagnostic Test (EDiT) Development Group. (2006). Eigo shindan test kaihatsu heno michi—ELPA Eigo Shindan Project no kiseki [Road to the development of the EDiT: Trajectory of the ELPA English Diagnostic Project]. Tokyo: Association for English Language Proficiency Assessment (ELPA).


References (3)


Any questions?

- Thank you for listening!