**03: Validity Evidence**

**1. Standards for Educational and Psychological Testing (1999)**

According to the *Standards* (1999), validity is “the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests” (p. 9).

For example, how does one know that scores from a scale designed to measure test anxiety provides scores that reflect test anxiety? More broadly, how can one demonstrate that scores derived from an instrument are valid?

Validity is established not by a single method, but by providing multiple examples of evidence for scores obtained from an instrument. Validity is established through a process, and it is sample specific, like reliability, so one must provide evidence of validity with each use of an instrument.

Below are some of the more commonly employed means for providing evidence for validity.

**2. Evidence based on Test, Scale, or Instrument Content**

Logical validity, or content validity, stems from the logical/judgmental analysis of items and instrument format. As Goodwin and Leech (2003) explain:

“...this type of validity evidence is based on logical analyses and experts’ evaluations of the content of the measure, including items, tasks, formats, wording, and processes required of examinees. In general, it addresses questions about the extent to which content of a measure represents a specified content domain.” (p. 183)

Detailed steps one takes to establish content validity will be outlined in separate presentations for test construction and questionnaire development. For now, we will consider a summary of steps, listed below, for judging content validity.

1. Delineate purpose of instrument
2. Define constructs to be measured
3. Identify and explain domains (dimensions) of construct
4. Develop item pool to fit domains with adequate sampling of each domain
5. Expert review of each item and review of entire instrument
6. Pilot test with feedback

Haynes, Richard, and Kubany (1995; articled linked in readings for validity on course website) provide a very detailed outline of steps one should take to address content validity. See also Holmbeck and Devine (2009; linked in references below) and Table 1 of Goodwin and Leech (2003; linked in references below) for ideas about expert and participant reviews concerning instrument and items (e.g., item importance, clarity, relevance, bias).

|  |
| --- |
| **Example1: Content Validity Evidence as Presented in a Published Study**  The publication by Menon (2001) focuses on questionnaire development.  Menon, S.T. (2001). [Employee empowerment: An integrative psychological approach](http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-assignments/02-Menon-ST-2001.pdf). Applied Psychology: An International Review, 50, 153-180.  Construct (Latent Variable): Perceived Control  1. Purpose of instrument p. 155  2. Perceived control defined and described p. 159-161  3. Identify and explain domains (dimensions) of construct p. 159-161  Aspect of Perceived Control p.159  4. Develop item pool to fit domains with adequate sampling of each domain, p 162  6. Expert item analysis, expert review of entire instrument, p. 162  7. Pilot test with feedback, p. 162 and 169  Two studies conducted, the first to assess and refine items, the second to further review and validate items. |

**3. Evidence based on Response Processes**

Whiston (2009) explains that evidence based upon response processes focuses on whether

“…individuals either perform or respond in a manner that corresponds to the construct being measured. For example, if an assessment is attempting to measure vocational interest, then the instrument developers might examine whether people are answering the items based on what they like to do rather than what they can do.” (pp. 70-71).

In short, this type of evidence addresses whether respondents view and understand items, instructions, and instrument in same way and respond using anticipated methods.

Examples:

* Teachers using a rubric to assign scores (rating responses in appropriate manner)
* Checking for socially acceptable responses on attitude measures
* Asking respondents how they derived their response choice
* Determining whether mathematic reasoning was used to derive answers (test-taker is responding to items using appropriate processes)
* “Talk-aloud” – asking respondents to explain their reasoning for the answers provided
* Show your work—illustrate how answers were determined
* Demographic item:

Sex \_\_\_

When responding to this item, do respondents think male vs. female, or do respondents think type or frequency? Responses may vary depending upon maturity of respondent. Revise item to make clearer:

What is your Sex? Female \_\_\_\_\_

Male \_\_\_\_\_

This type of validity evidence can overlap with content validity because it is partially concerned with how and why individuals respond the way they do.

One reason for the field test in content validity is to determine whether individuals are reading and interpreting questionnaire items in a similar manner, and whether they attempt to address those items using a framework or schema that aligns with what the scale was designed to measure.

|  |
| --- |
| **Example 1: Continued, Response Processes**  Menon, S.T. (2001).  Construct: Perceived Control  The review, performed by two experts and three doctoral students, addressed Response Process, p. 162 |

**4. Evidence based on Internal Structure**

This type of evidence concerns whether data derived from a measure conform to theoretically expected patterns. Think in terms of constructs and domains – do the domains show distinct response patterns; do different constructs show distinct response patterns?

Methods for addressing internal structure include:

* Correlations among items and scale summated scores (item analysis steps in both test and questionnaire assessment)
* Correlations among domains of a construct
* Exploratory factor analysis
* Confirmatory factor analysis
* Internal consistency is not a measure of internal structure, although some appear to accept it as one indicator of structural adequacy (e.g., Floyd et al. 2005); recall the presentation in the internal consistency notes explaining why internal consistency is not an adequate assessment for uni-dimensionality.

|  |
| --- |
| **Example 1: Continued, Internal Structure**  Menon, S.T. (2001).  Menon performed several assessments of internal structure.  Page 163, Correlations among items (p. 165 Table 1, note correlations in bold)  Page 164, Factor analysis (p. 166 Table 2 factor loadings, note loadings in bold)  Page 164, Internal consistency  Page 167, Confirmatory factor analysis (p. 167 Figure 1 shows loadings and correlations) |

|  |
| --- |
| Item-scale Correlations  This is rarely reported in research, but below is one example:  Frey and Bos 2012 report minimum and maximum item correlations with each dimension of a social competency scale  <http://www.bwgriffin.com/gsu/courses/edur9131/content/Frey_JERO-301-1176-1-PB.pdf>  Table 4, p 34  Factor analysis examples  Frey and Bos 2012 – factor analysis did not go as planned:  <http://www.bwgriffin.com/gsu/courses/edur9131/content/Frey_JERO-301-1176-1-PB.pdf>  See Table 2, p 31 |

**5. Evidence based on Relations to Other Variables**

This type of evidence is used to demonstrate that measured scores behave in predictable patterns besides demonstrating internal structure. Sometimes this type of validity is referenced as criterion-related and includes concurrent, predictive, convergent, and divergent (discriminate) validity.

Researchers predict how obtained scores will relate to other variables. If the predictions are supported statistically, then this provides some evidence for validity of scores.

Concurrent and predictive are loosely defined forms of convergent validity – showing that scores from measure A correlate as expected with measure B.

Divergent validity refers to scores from measure A not correlating with measures that are not related to A. For example, test anxiety is unlikely to correlate with calorie consumption or hair color.

All the above are forms of construct validity.

Examples of evidence based upon relations to other variables:

* Scores from a new test anxiety measure should
  + correlate positively with scores from previously established anxiety measure (convergent validity)
  + correlate negatively with self-efficacy scores (convergent validity)
  + correlate positively with number of items answered incorrectly (convergent validity)
  + be higher for females than males (convergent validity)
  + be higher for an untreated group than a group provided with anxiety reduction training (convergent validity)
  + be unrelated to measures of subject interest, instructor ratings, one’s height or one’s weight (divergent validity)

|  |
| --- |
| **Example 1: Continued, Relations with Other Variables**  Menon, S.T. (2001).  Page 168, Correlations with other measures (p. 168 Table 3)  Convergent  Divergent  Theoretical Predictions   * Spreitzer Scale = empowerment scale, should be strongest with perceived control * Helplessness = lack of control, so should be negative with perceived control * Impact = influence one has, should be strong with perceived control * Self-determination = somewhat like perceived control * Competency = should be strongest with perceived competency * Meaning = likely to correlated best with goal internalization   Page 169, Study 2: Correlations with other measures, see description of variables p 170  p. 171 Table 4   * Centralization = opposite of autonomy, so negative with perceived control * Delegation = positive with control since those in control and delegate * Consulting * Self-esteem = strongest with perceived competence   p. 173 Table 5   * Organizational Commitment = Goal Internalization * Job Involvement = Goal Internalization * Organizational Citizenship Behavior (taking extra roles) = Goal Internalization |

**6. Evidence based on Consequences of Testing**

What are the expected and unexpected results, or consequences, of measurement? This is especially relevant for diagnostic scales used to discriminate among individuals (e.g., reading readiness tests, graduation tests, etc.).

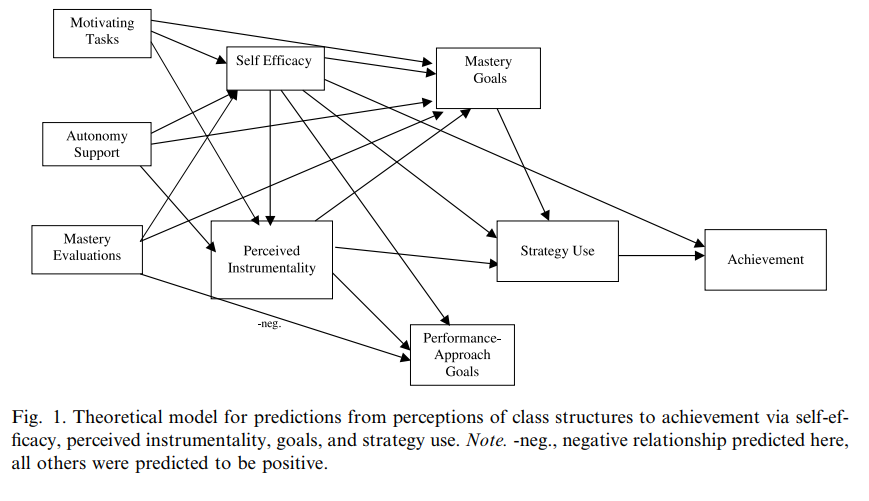
Section to be expanded.

**7. Example of Research Publication**

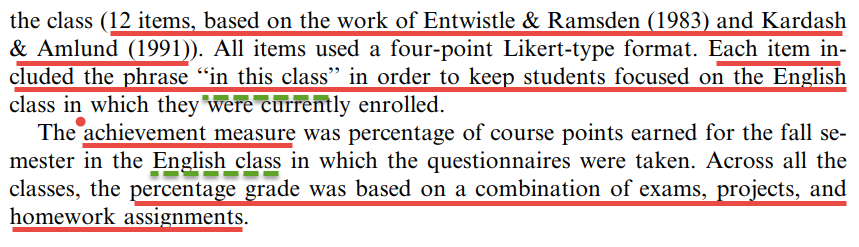
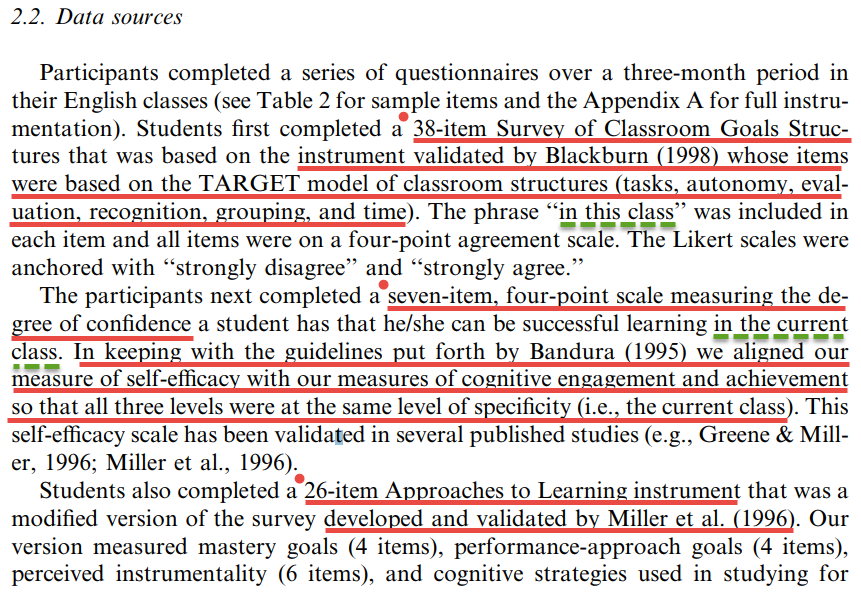
Greene, et al (2004) Predicting high school students’ cognitive engagement and achievement: Contributions of classroom perceptions and motivation. Contemporary Educational Psychology.

<http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/07-validity/07-2004-Greene-Predicting-cognitive-engagement.pdf>

p. 165 Purpose of Study (see below)



* Content Validity, p 469 (see below)
  + Previously developed and validated scales
  + Items based on theory and work of others



* Response Processes

They focused students on their English class by using the phrase “in this class” (underlined in green above).

* Internal Structure, p. 470
  + Confirmatory factor analysis to assess structure of constructs – assess whether they separate as expected
  + Cronbach’s alphas provided for each (Table 2, p 471)
* Relations with Other Variables (Convergent and Divergent Validity)
  + Page 471 (Table 3 p 472), discussion of zero-order correlations (Pearson r) among measures
  + Correlations behave as one would predict
  + Path analysis (p 472) and Figure (p 473) further validates measures by producing relations as hypothesized to some extent

**References**

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Floyd, Phaneuf, & Wilczynski (2005). [Measurement Propoerties of Indirect Assessment Methods for Functional Behavioral Assessment: A Review of Research](http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/07-validity/07-2005-Foyd-Properties-of-Assessments.pdf). School Psychology Review, 34, 58-73.

See Table 1 and discussion on p. 65-66 on Measurement Properties.

Goodwin, LD, & Leech, NL (2003). [The Meaning of Validity in the New Standards for Educational and Psychological Testing: Implications for Measurement Courses](http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/07-validity/07-2003-Goodwin-validity-standards.pdf). Measurement and Evaluation in Counseling and Development, 36, 181-191.

See Table 1 p. 187.

Haynes, S.N., Richard, D.C.S., & Kubany, E.S. (1995). [Content validity in psychological assessment: A functional approach to concepts and methods](http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/07-validity/07-1995-Haynes-Content-Validity-Assessment.pdf). Psychological Assessment, 7, 238-247.

Holmbeck, G. N., & Devine, K. A. (2009). [Editorial: An author's checklist for measure development and validation manuscripts](http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-assignments/02-Holmbeck-Devine-2009.pdf). Journal of Pediatric Psychology, 1-6, 2009.

Whiston, S. (2009). Principles and applications of assessment in counseling (3rd edition). Belmont, CA, USA: Brooks/Cole, Cengage Learning.