EDUR 7130
Presentation 6b
Validity

Presentation Content

1. **Relation between Validity and Reliability**
2. **Empirical Validity vs. Content Validity (or Logical Validity)**
3. **Content Validity**
4. **Empirical Validity: Review of Terms**
5. **Empirical Validity: Internal Structure and Dimensionality** Under development.
6. **Empirical Validity: Published Examples of Relations with Other Variables**
7. **Self-Test**

The discussion below uses various terms interchangeably to refer to measuring devices (i.e., tests, instruments, questionnaires, scales, ratings, measures, measured scores, and measuring tools).

1. Relation between Validity and Reliability

Brief review of these two concepts.

Reliability

The extent to which scores from a measuring device, or ratings, are consistent or agree.

Example

Comparison of three scales to measure my weight.

Scale 1: 226 lbs.
Scale 2: 225 lbs.
Scale 3: 224 lbs.

Given the level of agreement among scores, they appear to be reliable (i.e., scores are consistent and agree closely).

Validity

The extent to which scores from a measuring device, or raters, represent true scores; the degree to which measured scores reflect or represent the construct these scores were designed to measure.

Example

A single scale designed to measure weight in pounds. My known weight is 225 lbs. The scale is tested three times, 30 seconds apart, in sequence.

Scale Reading Time 1: 125 lbs.
Scale Reading Time 2: 125 lbs.
Scale Reading Time 3: 125 lbs.

Despite the perfect agreement, thus the perfect reliability, this scale does not accurately represent my true weight of 225 lbs. therefore the scores provided are invalid.

Note that in the above example reliability is perfect, $r_{xx} = 1.00$, but the scores lack validity. This demonstrates that high reliability does not guarantee high validity; so high reliability tells us nothing about the validity of scores.

If reliability is low, e.g., r_{xx} = .36, validity cannot exist because low reliability means scores are not consistent, scores do not agree, and scores vary widely, so those scores cannot be trusted to represent the true score, therefore validity must be low. Given this, low reliability means validity will be low too.

This can be illustrated with the scale example. This time scale readings differ widely, as shown below.

Scale Reading Time 1: 343 lbs.
Scale Reading Time 2: 57 lbs.
Scale Reading Time 3: 125 lbs.

Since there is no consistency among weights, these weights are not reliable. Additionally, we don't know which, if any, of these weights best reflects my true weight, so we cannot trust any of these weights to be a measure of my true weight, therefore the scale provides invalid – untrustworthy – weights. Low reliability leads to low validity.

Summary

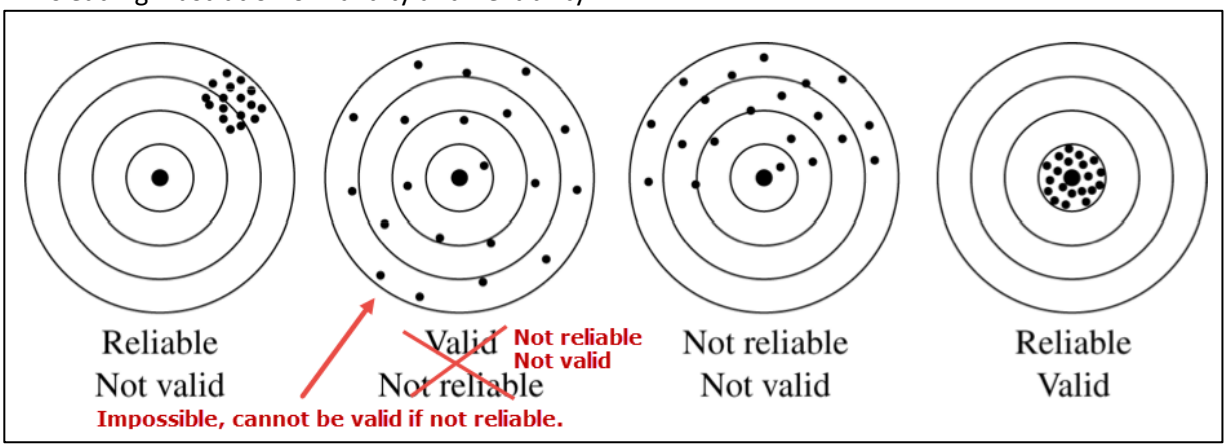
- Reliability means only that scores **agree** or are **consistent**, low reliability means validity is low, however, high reliability does not provide any indication about validity.
- Validity means that the scores **measure what they were designed to measure**; scores **accurately** reflect what they should be measuring, i.e., measured scores are in close agreement with the true score (i.e., recall true score discussion in notes 6a Reliability).
- Scores may be **reliable but not valid** (e.g., scales show my weight consistently at 35 lbs., but this weight is not accurate for me since my true weight is 225 lbs.; a clock that always reports the time as 10:00 is perfectly consistent, but not valid).
- Scores must be **reliable to be valid** (i.e., there is no way to show a valid measure of my weight if scores are inconsistent, e.g., I weigh 225, but scales provide inaccurate weights: 185, 255, 177, 221).

Table 1: Relation between Validity and Reliability

		Validity	
		Low	High
Reliability	High	Scores reliable/consistent/agree, but not valid (example scales always indicate I weigh 5 lbs., yet my true weight is 225 lbs., scales are perfectly consistent and completely wrong).	Scores are valid and reliable, scores are consistent/agree and measure what they are designed to measure (e.g., my true weight is 225 lbs. and every time I weigh myself the scales report values close to 225 lbs. such as 224, 223, 226, 225).
	Low	Scores are not consistent, and therefore cannot be valid (example, scales repeatedly report my weight incorrectly, 125 lbs., 23 lbs., 389 lbs.),	Not a possible combination, cannot have high validity if reliability is low. Same problem as shown at left – low reliability leads to low validity.

Figure 1 is a common graphical display, popular in research textbooks, used to illustrate the relation between validity and reliability. I present this display to clarify one misleading component, which is noted in red within the display. It is not possible for scores to be valid if they are not reliable. The other three examples, however, do work well to illustrate validity and reliability assuming the center of the target is the true score estimated via measurement.

Figure 1: Misleading Illustration of Validity and Reliability



2. Empirical Validity vs. Content Validity (or Logical Validity)

What is the difference between empirical and content validity?

- Summary:
  - **Content validity** refers to the **process used to develop items** for the instrument or scale, while
  - **empirical validity** refers to the **process and result of testing score behavior**.
- **Content, or Logical, validity** means one uses logical arguments or reason, rather than data, to suggest the instrument's items will provide useful scores. Logical validity does not employ predictions or hypothesis testing of how instrument scores will behave. Instead, one explains the theory/rationale/logic for why particular items are included; explains the various dimensions of constructs measured and how items match those dimensions; critically reviews items to ensure wording is clear and items fit construct dimensions as defined, and that there are sufficient number of items per dimension/domain to adequately assess that dimension or construct; and one seeks input from others about the adequacy of questionnaire/scale items.
- **Content validity** a misleading term in the sense that it does not assess whether obtained scores closely match true scores. Rather than evaluate scores, content validity refers to the process of instrument development.
- **Empirical validity** means one collects data to test predictions about how scores from an instrument behave. One must show that scores from an instrument behave in a predictable manner before those scores can be shown to be valid. Often this is done through forming hypotheses and testing those hypotheses, e.g., test anxiety scores should correlate negatively with math achievement scores (e.g., Pearson r = -.33, this supports validity evidence for test anxiety scores).
- While both empirical and logical validity are needed, empirical evidence for validity is critical for assessing validity of scores.

3. Content Validity

Content validity refers to the process of developing measuring instruments (e.g., tests, questionnaires, scales, etc.).

Content validity, also logical validity, is demonstrated by:

- **Defining** and describing fully the construct to be measured and identifying the various dimensions (or domains/traits) of that construct. For example, one should explain how test anxiety is defined and identify the various dimensions (cognitive and physiological) that should be included on a test anxiety scale.
- Next one should develop/select items to measure that construct, and address **item validity** to show that **each and every** item aligns well with the construct intended and also appears to be suitable for the targeted population; the item should be clearly and concisely written and at an appropriate language level.
- One then addresses **sampling validity** to show that each dimension [or domain or objective or trait] of a construct has an adequate number of suitable items to allow one to measure well each dimension, e.g., to

measure test anxiety one should include a number of items for each of the cognitive dimension and the physiological dimension of anxiety.

- In addition to item and sampling validity, some may also address **face validity** which entails expert judgment that the items (and instrument, including instructions and layout) appear appropriate for the targeted population.

In short, content validity entails

1. Determining purpose of scale, test, or instrument
2. Defining the construct to be measured
3. Identifying and explaining the domains of the construct
4. Developing/selecting a pool of items that fit the construct dimensions
5. Reviewing items, with experts, and field testing items and instrument with a pilot study

Example 1

For a classroom test one should:

- clearly determine which domains or objectives will be included on that test, and
- align each item on the test with an objective (domain or dimension) for the test (e.g., Educational Research Test 2 will cover inferential statistics, reliability, validity, the quantitative research matrix, so each of these four domains should have relevant items on the test), and
- there should be enough items for each domain to provide an adequate sampling of that domain so one can be sure the test assesses student understanding of that domain (e.g., there should be several items each for inferential statistics, reliability, validity, and quantitative research), and lastly
- content "experts" should agree that the test items align with the dimensions identified and adequately represent that dimension (e.g., other teachers review the test beforehand to identify problematic items and determine whether enough item exist for each domain sampled).

Example 2: Published Study with Focus on Scale Development

Menon, S.T. (2001). Employee empowerment: An integrative psychological approach. Applied Psychology: An International Review, 50, 153-180.

Purpose: Menon developed a scale to measure the construct employee empowerment. Menon argued that employee empowerment should consist of three dimensions (hence three sub-scales): (1) perceived control, (2) perceived competence, and (3) goal internalization. The material below shows the steps Menon took to provide evidence for content validity of his scales.

Note: See presentation video for discussion of this information.

This example will focus on the Perceived Control sub-scale.

1. Purpose of instrument p. 155

> This article addresses some of the concerns outlined above. It aims to clarify the definitional and conceptual issues surrounding the empowerment construct by proposing an employee-centred psychological approach. After a brief review of existing literature on empowerment, an integrative psychological perspective on employee empowerment is developed. The results of a measure development study based on this integrative approach are then presented.

2. Perceived control defined and described p. 161

Perceived control refers to beliefs about autonomy in the scheduling and performance of work, availability of resources, authority and decision-making latitude. Perceived competence denotes self-efficacy and confidence

3. Scale dimensions identified
4. Develop item pool for each dimension with adequate sampling of each dimension, p 162

Item Generation. In this stage, the intention was to generate a large pool of items for possible inclusion in the scale. In the present formulation, as empowerment is envisaged as a multidimensional construct, items that tap all three dimensions needed to be included. Given the dearth of empirical precedent, the bulk of the items had to be written anew. Dwyer and Ganster's (1991) scale of perceived control, Paulhus's (1983) sphere-specific measures of perceived control, Jones's (1986) measure of generalised self-efficacy, and Hill, Smith and Mann's (1987) scale for computer efficacy were referred to for initial guidance. Initially, an item pool of 60 items was generated, 20 items for each dimension.

5a. Expert item analysis, expert review of entire instrument, p. 162

Expert Review. The 60 items were then evaluated by a panel of two faculty members and three doctoral students. The faculty members, both familiar with the content area of empowerment, were first asked to review each item in terms of its relevance to the domain of empowerment. This initial screening resulted in a reduced list of 40 items for further consideration. The doctoral student reviewers were then provided with the definition of empowerment developed for this research and were asked to judge each item with regard to (a) its relevance to the empowerment construct as defined, (b) conceptual ambiguity, (c) sentence clarity, (d) conciseness, (e) the subscale to which it belonged, and (f) social desirability. Each item was ranked on all the above dimensions and a mean rank was calculated by averaging the ranking of the three reviewers. For each dimension, the highest ranking five items were selected to form the final list of 15 items to be included in the questionnaire. At the time of questionnaire

5b. Pilot test with feedback, p. 162 and 169

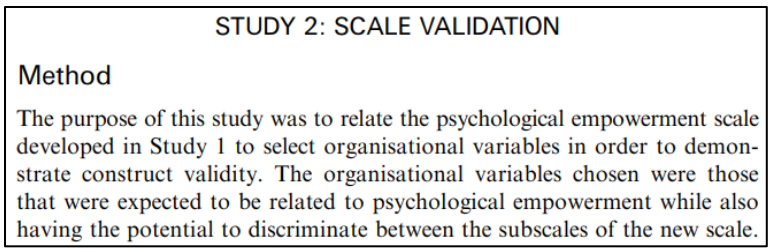
The pilot test is designed to allow for critical review of item performance. This step is taken to identify weak items for revision or deletion; this step also can be used to identify other instrument problems (e.g., poor instructions, feedback from participants about problems and recommended solutions, etc.). At this stage some preliminary evidence for reliability and validity of scores will be assessed as part of the process to identify problematic items.

Menon conducted two studies: the first to assess and refine items and eliminate poor items, the second to further review and validate items. The second was more detailed and involved thorough assessments of reliability and validity. Some of the validity evidence Menon presented will be discussed below in the empirical validity section.

STUDY 1: MEASURE DEVELOPMENT

Method

The measure development process was patterned on the De Vellis (1991) procedure for scale development. The major stages are described in the following sections.



STUDY 2: SCALE VALIDATION

Method

The purpose of this study was to relate the psychological empowerment scale developed in Study 1 to select organisational variables in order to demonstrate construct validity. The organisational variables chosen were those that were expected to be related to psychological empowerment while also having the potential to discriminate between the subscales of the new scale.

As this example shows, there is a logical, systematic process to providing evidence for content validity.

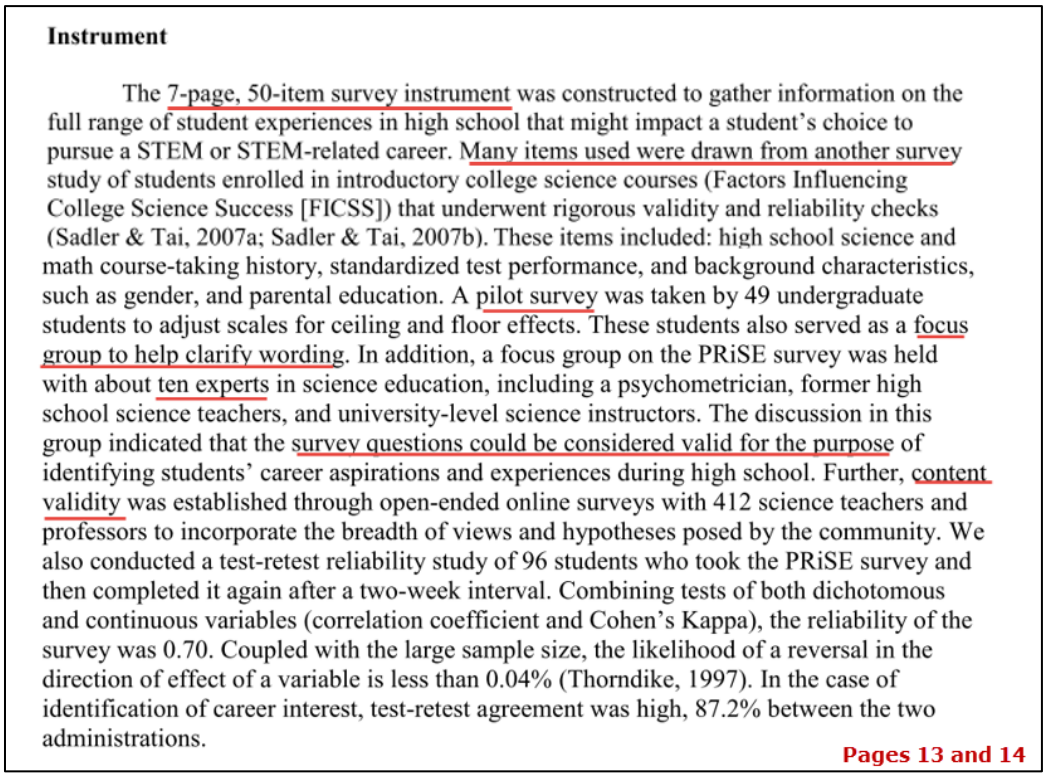
Example 3: Published Study with Focus on Research (Scale Development Secondary)

Fuchs, T. T., Sadler, P. M., & Sonnert, G. (2015). High School Predictors of a Career in Medicine. Journal of Career and Technical Education, 30(1), 9-28.

Purpose: Authors attempted to identify variables that predict career choice in medicine. Authors developed a questionnaire for this study and addressed content validity, but since their study focus was not on questionnaire development, their presentation on content validity is necessarily less detailed than the Menon example above.

Note: See presentation video for discussion of this information.

The authors describe purpose and construction of instrument in the Instrumentation section of the Method.



Instrument

The 7-page, 50-item survey instrument was constructed to gather information on the full range of student experiences in high school that might impact a student's choice to pursue a STEM or STEM-related career. Many items used were drawn from another survey study of students enrolled in introductory college science courses (Factors Influencing College Science Success [FICSS]) that underwent rigorous validity and reliability checks (Sadler & Tai, 2007a; Sadler & Tai, 2007b). These items included: high school science and math course-taking history, standardized test performance, and background characteristics, such as gender, and parental education. A pilot survey was taken by 49 undergraduate students to adjust scales for ceiling and floor effects. These students also served as a focus group to help clarify wording. In addition, a focus group on the PRiSE survey was held with about ten experts in science education, including a psychometrician, former high school science teachers, and university-level science instructors. The discussion in this group indicated that the survey questions could be considered valid for the purpose of identifying students' career aspirations and experiences during high school. Further, content validity was established through open-ended online surveys with 412 science teachers and professors to incorporate the breadth of views and hypotheses posed by the community. We also conducted a test-retest reliability study of 96 students who took the PRiSE survey and then completed it again after a two-week interval. Combining tests of both dichotomous and continuous variables (correlation coefficient and Cohen's Kappa), the reliability of the survey was 0.70. Coupled with the large sample size, the likelihood of a reversal in the direction of effect of a variable is less than 0.04% (Thorndike, 1997). In the case of identification of career interest, test-retest agreement was high, 87.2% between the two administrations.

Pages 13 and 14

The passage above shows steps taken by the authors to address content validity:

1. Possible predictors of career choice solicited from 412 science teachers and professors
2. Items selected from another questionnaire
3. Items selected addressed many constructs that may predict student career choice
4. Critical review of items:
    (a) focus group of students in pilot study to clarify wording

(b) 10 experts reviewed instrument

5. Assessed whether instrument was suitable for targeted population (HS students) via expert review
6. Pilot study conducted to assess functioning of questionnaire

While not as detailed as the earlier example from Menon, the authors of this study do provide sufficient evidence of their attempt to address the conceptual issues with questionnaire development and appear to provide a thorough item develop and critical review process.

4. Empirical Validity: Review of Terms

Empirical validity refers to the use of data, and also predictions/hypotheses, to assess score behavior from a measured construct.

The logic is simple – if data behave in a predictable manner, then that provides some proof the instrument is measuring the construct as intended. The closer the match between data behavior and what was predicted, the stronger the evidence for validity.

There are two general processes for assessing empirical validity:

- Internal Structure, or Relations among Indicators (scale items), and
- Relations with Other Variables, which has many possible approaches.

Internal Structure material is still under development.

Various terms that describe Relations with Other Variables are presented below, and examples from published studies are provided in section 6. As you will see, there is often overlap in practice across methods for examining Relations with Other Variables.

Convergent Validity

Defined: The degree to which two scales designed to measure the same construct produce scores that converge or agree.

More than two scales can be compared, and different methods for measuring the construct can also be compared. Different methods may be illustrated, for example, by a self-administered paper scale vs. a professional evaluation performed by a psychologist. Another example of different methods: self-ratings may be compared to peer ratings.

This notion of convergent validity, that measurement of one construct by two or more methods should produce moderately to strongly correlated scores, was introduced by Campbell and Fiske (1959).

Examples:

- Two parallel forms of an algebra test should produce scores that are highly correlated.
- Two measures of depression, Beck Depression Inventory (BDI) and Center for Epidemiologic Studies Depression Scale (CES-D), should provide similar, correlated assessments for a sample of respondents.
- Measuring test anxiety with two different scales (e.g., TAS – test anxiety scale vs. STABS – Suinn test of anxiety behavior scale) should produce correlated scores.
- Degree of implementation of professional learning communities is evaluated in all schools in Fulton County using two methods: (a) employee self-assessment via administration of the PLC-Assessment Revised scale and (b) program evaluation executed by independent evaluators. Ratings from both methods are compared to assess agreement.

Some authors depart from Campbell and Fiske's more narrow definition and argue that convergent validity is established when one correlates scores from one measured construct to theoretically distinct, but related constructs.

For example, one would expect the scores from mathematics test anxiety would correlate negatively with scores from mathematics self-efficacy. Both are distinct constructs but are also theoretically related. While this definition is common, Campbell and Fiske's conceptualization of convergent validity (i.e., one construct measured in different ways) will be the default conceptualization of convergent validity in this presentation, and the idea that multiple constructs can be theoretically linked will be considered a characteristic of **construct validity**, described below.

Divergent Validity

Defined: Two or more constructs that are theoretically unrelated should produce scores that correlate weakly, i.e., scores from these unrelated constructs should diverge or not agree.

Hypotheses formed to assess divergent validity should provide an indication about the strength of relation expected, whether near zero, weak, or moderate. See the first example below for an illustration.

Examples:

- Scores from two measures of test anxiety (TAS and STABS, see above) should correlate highly and therefore ; however, both measures should diverge from a measure of academic self-efficacy and produce both weaker and negative correlations, and both should diverge from a measure of religiosity and produce near zero correlations. Combined, these three predictions form an example of **construct validity**.
- Test scores in EDUR 7130 should be unrelated to – diverge from – scores from a measure of job satisfaction and a measure of job autonomy.

Discriminate Validity

Two possible definitions are offered below.

(a) Fine Distinctions Among Dimensions or Methods

Defined: Discriminate validity, according to Campbell and Fiske (1959), refers to the ability of a scale to discriminate, or identify differences, when a construct is measured by multiple dimensions or methods.

Many equate **Discriminate** and **Divergent validity**. However, Campbell and Fiske offered a description of discriminate validity that is more nuanced. As noted above, **divergent validity** is evidenced when constructs are predicted to be unrelated or weakly related to other constructs, and data support these predictions. **Discriminate validity** refers to the ability of a construct to show distinctions among dimensions or domains that form that construct, or to show distinctions among methods used to measure that construct.

The expectation is that scores from the different dimensions or methods will correlate, and perhaps strongly, but the correlation will be weak enough to show that the dimensions or methods do, in fact, differ.

Example:

Test anxiety has at least two dimensions

(a) cognitive, how one thinks about anxiety,
"I worry about the consequences of failing this test"
"I expect poor grades or failure when taking a test"
(b) physiological, how one reacts to anxiety,
"My stomach feels in knots before an important test"
"I get nervous during difficult tests"

Scores from both dimensions should correlate strongly, perhaps with a correlation between .50 and .80 for example, but the correlation should not be so high that the two dimensions become indistinguishable or are viewed by respondents as the same thing. A correlation of .85 or larger could be

viewed as indicating the two dimensions are the same – that is, it is not possible for respondents, when providing answers, to discriminate between cognitive and physiological aspects of test anxiety.

In short, while scores between dimensions are expected to correlate, the correlation should not be perfect or near perfect because one would not expect everyone to have the same cognitive and physiological anxiety reactions to tests. Some may experience more worry but have little physical reactions while others may have more severe physical reactions but less worry. If these two dimensions are truly distinct, then it is possible for some respondents to score higher on one dimension than the other. Given this, the scores from both dimensions should discriminate – show that the two dimensions produces scores that are slightly different thus confirming there are two dimensions to test anxiety. Such as distinction illustrates the nuances outlined by Campbell and Fiske.

(b) Discrimination Among Groups

Defined: Construct scores should be able to clearly distinguish groups with known or theoretical differences if those differences are related to the construct. Construct scores should be able to predict group differences.

Some authors argue that **discriminate validity** also means a scale should be able to provide scores that clearly distinguish groups with known differences, or theoretically assumed differences, on the construct measured by the scale.

Example:

- Assume a shorter version of the Beck Depression Inventory (BDI) is created and scores must be tested to assess validity and reliability. A sample of participants are selected in a known-group validity test. For the depression-known group, 15 individuals are selected who have been diagnosed with depression (i.e., spoken with a doctor about depression symptoms, completed the HAMD-17 scale and obtained a score of 24+ signifying depression, etc.). For the non-depression group, 15 individuals with no symptoms of depression and no clinical diagnosis of depression are selected. Both groups are administered the shortened BDI scale. If BDI scores are statistically higher for the depression-known group than scores for the non-depression group, then this finding provides evidence that the shortened BDI can discriminate between those with depression and those without depression, thus discriminate validity is obtained.
- A teacher identifies the 6 best readers/writers and 6 worst readers/writer in her classroom. If the Georgia's Milestones English Language Arts (ELA) test scores demonstrate discriminate validity, then the 6 best readers/writers should have higher ELA scores than the 6 worst readers/writers.

In summary, **discriminate validity**, as described above, refers to measuring devices demonstrating the ability to make fine distinctions among similar constructs, or dimensions of constructs, or methods for measuring a construct. It can also mean that scores from measuring devices are able to discriminate among groups that have more or less of whatever is assessed by the measuring device, i.e., the measuring device can predict group differences.

Criterion-related Validity

Defined: Scores from an instrument are correlated with a standard (or a "gold standard"), or criterion, to judge the validity of scores from the instrument.

If the instrument and standard produce similar scores, criterion validity evidence is obtained. There are two general types of criterion-related validity, **concurrent** and **predictive**, with timing of criterion usually being the distinction.

Concurrent Validity

Defined: Scores from an instrument correlate well with a related standard, or criterion, and both sets of scores are collected at about the same time.

Sometimes concurrent validity overlaps with **convergent validity** and **construct validity**, discussed below.

Examples:

(a) Scale to Measure Weight

A new scale is built to measure weight in pounds and ounces. Readings from this scale must be validated, or calibrated, with items of known weight. A number of items with various known weights are used as standards, or criteria, for judging the scale. These items range in weight from 0.5oz to 500lbs. with many weights between these extremes. Each item is weighed, the scale reading is recorded, and this is done for each item obtained. Scale reported weight is compared with the known weight for each item to assess level of agreement between the two; the higher the agreement, the greater concurrent validity.

(b) Beck Depression Inventory (BDI)

The BDI consists of 21 items. A shorter version is created with 7 items and is identified as the BDI-7. In this example, scores from the BDI-7 must be evaluated for evidence of validity. The comparison standard, or criterion, are scores from the original BDI. Both are administered to a group of individuals at about the same time, or only one or two days apart (to ensure little change in depression symptoms), and scores from both are compared to assess agreement, or correlation. The higher the agreement, the greater the concurrent validity evidence for the BDI-7. This example also illustrates **convergent validity**, so both concurrent and convergent overlap with this example.

Predictive validity

Defined: Scores from an instrument correlate well with a related standard, or criterion, that occurs in the future (or possibly the past).

Most authors of measurement texts explain that the primary difference between **predictive** and **concurrent validity** is when the criterion occurs. With **concurrent**, the criterion occurs nearly simultaneously with administration of the scale examined. With **predictive validity**, the criterion occurs sometime in the future. It is also possible to have retrospective predictive validity, or postdictive validity, in which an instrument's scores are used to "predict" past criteria/events.

Examples:

(a) Job Autonomy and Turnover

Van den Broeck et al. (2010) developed a scale to measure job autonomy which is the degree to which one has control over decisions and actions on their job (e.g., "I have to follow other people's commands" or "If I could choose, I would do things at work differently"). They sampled 261 employed individuals who had a minimum of three years of working experience and administered their autonomy scale. After six months turnover data were collected from HR managers at each business sampled; 31% of the sample had ceased their employment. They found that job autonomy scores significantly predicted turnover; those with lower autonomy ratings were 2.7 times more likely to quit their jobs. Turnover, six months later, is the future criterion in this example.

(b) Emotional Intelligence and Academics

Several researchers have examined whether emotional intelligence measures (e.g., EQ-i) predict academic performance. Newsome et al. (2000) administered the EQ-i to 180 college students and then months later collected their college GPAs (which is the future criterion in this example). The correlation of EQ-I scores with GPA was r = 0.01, which suggests almost no relation between emotional intelligence and college GPA; EQ-i had no validity evidence for predicting GPA. Van Rooy and Viswesvaran (2004) conducted a meta-analysis of 11 studies with a combined sample size of 1,370 students and found that emotional intelligence correlated at a low level, r = 0.09, with academic outcome measures. This finding suggests little predictive validity evidence for emotional intelligence predicting academic performance.

(c) Warning Behaviors and the Postdiction of School Shooters

Meloy et al. (2014) used a typology of eight warning behaviors in a retrospective study to postdict likely school shooters vs. other students of concern. The warning behaviors included, for example, planning attacks, "fixation/preoccupation with a person or a cause" (p 204), "warrior mentality" (p 204), and leakage communications suggesting intentions. They identified a sample of school shooters and a comparable sample of other students who were of concern for violence in their schools. Once the two samples were identified, the authors then reviewed court documents and news reports for each student and coded their eight warning behaviors based upon evidence provided in court documents and news reports. They found strong evidence for identifying or distinguishing between those who were school shooters and students of concern. For example, 100% of shooters displayed warrior mentality while only 16% of the students of concern displayed a warrior mentality. The phi correlations ranged between .61 and .88 for the successful postdictors (note, phi correlation is Pearson r between two dichotomous, binary, variables). The two groups – shooters vs. non-shooters – serve as the criterion in this example.

Construct Validity

Defined: Construct scores correlate, as hypothesized/predicted, with other constructs or observed variables.

Evidence for construct validity is assessed by developing predictions – hypotheses – based upon theory of how scores from a measuring device should behave relative to other constructs or observed variables. The degree to which these hypotheses are supported by data directly reflects the degree to which validity evidence is provided. This, of course, is how evidence is established with other types of validity too. One could view construct validity as the archetype that subsumes other types of validity types; the other types are special cases of construct validity. Construct validity evidence requires that one theoretically predict how construct X will relate to constructs Y and Z, and maybe observed variables W and V.

Examples:

(a) Beck Depression Inventory (BDI)

Recall the example for **concurrent validity** presented above: a 7-item version of the BDI is developed and called the BDI-7. Concurrent validity evidence is established if a sample of scores between the BDI and BDI-7 correlate strongly. To extend this study with additional predictions would further build the case for **construct validity**. It is not enough that BDI-7 correlates with BDI, one must also show scores from BDI-7 correlate, as predicted/hypothesized with related and, perhaps, unrelated variables. For example, research shows that depression is likely to correlate weakly to moderately, and negatively, with life satisfaction; moderately and positively with suicide ideation; and positively to obesity. In addition to collecting BDI and BID-7 scores, data on life satisfaction, suicide ideation, and obesity would be administered too. Scores from these measures would then be correlated to BDI-7 to learn whether depression, as measured by BDI-7, correlates as expected with BDI, life satisfaction, suicide ideation, and obesity.

(b) Test Anxiety Scale

To provide validity evidence for the test anxiety scale (TAS), the following predictions are made: (a) TAS should correlate positively and strongly with another measure of test anxiety, the Suinn test of anxiety behavior scale (STABS). This correlation, as previously noted, is also evidence for **convergent validity** and **concurrent validity**. In addition, (b) TAS should correlate negatively, but moderately, with academic self-efficacy, and (c) correlate near zero with a measure of religiosity which is expected to be unrelated to test anxiety. These latter two predictions, (b) and (c), are examples of **divergent validity**, as previously noted. Combined, these three predictions form an example of **construct validity**.

In summary, the key to construct validity is the formation and testing of hypothesized relations with the measured construct. Also, note that one hypothesis – one examined relationship – between the measured construct and another

variable/construct is not sufficient. Construct validity requires multiple assessments of how the measured construct behaves relative to other constructs or observed variables. It is the accumulation of evidence, the process and successful outcomes, that provides confidence in construct validity claims.

Internal Structure

To be added.

5. Empirical Validity: Internal Structure and Dimensionality

Internal structure refers to construct dimensionality (or domains, like test anxiety has both a cognitive and physiological components), measurement equivalence (or invariance, whether different groups interpret scale items and respond similarly), and reliability. This presentation will focus exclusively on construct dimensionality. See Rios and Wells (2014) for more on measurement invariance and reliability, and how both relate to internal structure.

(a) Dimensionality

Under development.

(b) Reading Factor Analysis Results

Under development.

6. Empirical Validity: Published Examples of Relations with Other Variables

The purpose of this section is to provide examples of validity evidence in published research.

(a) Two or More Instruments Designed to Measure the Same Construct

This type of validity involves testing whether scores, ratings, or classifications from an instrument correlate with scores from a second instrument that measures the same, or a highly similar construct. By comparing scores, ratings, or classifications between the two – instrument 1 vs. instrument 2 – one provides evidence for **convergent** and **concurrent validity**. For this assessment to be effective, the comparison instrument, or criterion, must have strong evidence for reliability and validity.

Procedure:

- Administer both instruments to the same group of people at nearly the same time.
- Second, examine degree of agreement or relation between the scores from both instruments to obtain an estimate of the **validity coefficient**.

Validity Coefficient (usually Pearson's r)

- The correlation obtained is known as the **validity coefficient**, and it is an index that measures evidence for concurrent validity.
- The validity coefficient will typically range from 0.00 to 1.00 although negative values are possible but rare and whether a negative validity coefficient is useful depends upon the scaling and scoring of the two instruments. If higher scores on one instrument indicate more of the construct, but lower scores on the other instrument indicate more of the construct, then a negative correlation would be expected. Normally, however, positive relations are expected.
- It is possible to have other measures of validity coefficients depending upon the nature of the rating system used. For example, if the instrument requires classification into groups (e.g., test determines presence of COVID-19, yes/no), then the validity coefficient could be percentage correct classification (e.g., 88% correctly identified as having COVID-19), or a measure designed for qualitative/categorical measurement, such as Cohen's kappa which was very briefly discussed in the video 6b for reliability.

Example 1

Dadfar, M., & Lester, D. (2017). Cronbach's α reliability, concurrent validity, and factorial structure of the Death Depression Scale in an Iranian hospital staff sample. International journal of nursing sciences, 4(2), 135-141.

Purpose: "The aim of this study was to explore the performance of the Farsi version of the Death Depression Scale [DDS] with an Iranian convenience sample of nurses (n = 106)." Scores from the DDS were compared with other death-related measures to assess criterion-related evidence for validity.

Note: See presentation video for discussion of this information.

In addition to administering the DDS, the authors administered five other theoretically related scales that assess death thoughts and concerns. Correlations between DDS scores and scores from the other five measures are show in Table 4 below.

Table 4
Descriptives of all scales and correlations with the DDS.

Scales	Mean	SD	Cronbach's α	Pearson r with DDS
Death Concern Scale (DCS)	72.72	10.82	0.73	0.40**
Collett-Lester Fear of Death Scale (CLFDS)	99.15	25.14	0.94	0.39**
Templer's Death Anxiety Scale (DAS)	8.27	2.71	0.60	0.50**
Reasons for Death Fear Scale (RDFS)	57.70	14.23	0.90	0.35**
Death Obsession Scale (DOS)	30.74	12.35	0.95	0.44**

*Two-tailed $P < 0$ 0.01.

The authors provide the following brief interpretation of results and conclude, in the Discussion section, that these results provided evidence for **concurrent validity**. This could also an example of **convergent validity**.

> The DDS correlated 0.40 with the DCS, 0.39 with the CLFDS, 0.50 with the DAS, 0.35 with the RDFS, and 0.44 with the DOS, indicating good construct and criterion-related validity. Concurrent validity for the DDS with the other scales, were significant (See Table 4). Pages 137-138

Example 2

Stratton, R. J., Hackston, A., Longmore, D., Dixon, R., Price, S., Stroud, M., ... & Elia, M. (2004). Malnutrition in hospital outpatients and inpatients: prevalence, concurrent validity and ease of use of the 'malnutrition universal screening tool'('MUST') for adults. *British Journal of Nutrition*, *92*(5), 799-808.

Purpose: Test ease of use of MUST (malnutrition universal screening tool) and the degree to which classifications and scores from MUST agree with classifications from other malnutrition measures (i.e., assess convergent or concurrent validity).

Note: See presentation video for discussion of this information.

The authors provide a short definition of concurrent validity on page 801: "A tool can have concurrent validity if it shows good to excellent agreement with other tools or with a reference standard" [i.e., criterion].

The MUST measure provides scores that are ordinal, which often work with Pearson correlation to assess concurrent validity. However, MUST provides only three categories of scores:

- low risk,
- medium risk, and
- high risk for malnutrition.

Pearson correlation is not a good analysis tool when there are only three categories, so instead the authors use a nominal measure of agreement, Cohen's kappa (see brief discussion in notes 6a Reliability), to assess level of agreement among instrument classifications.

Below are sections of Stratton et al.'s (2004) study in which they discussed questionnaire development (content validity) and reliability (degree of classification reproducibility across users).

Table 3 contains estimates of agreement between MUST and other measures of malnutrition, the criteria used to assess the validity of MUST scores. The authors explain, in the Table 3 footnote, how to interpret the measure of agreement, kappa, with values of .40 to .75 viewed as fair or acceptable, and scores over .75 as excellent.

In the absence of a 'gold standard' for malnutrition it is difficult to establish the validity of nutrition screening tools. However, 'MUST' has content validity (comprehensiveness of the tool), face validity (issues which are relevant to the purpose of the test) and internal consistency. 'MUST' has some predictive validity, e.g. predicting length of hospital stay, mortality and discharge destination of groups of hospital patients (King *et al.* 2003; Wood *et al.* 2004) and general practitioner visits and hospital admissions in free-living individuals (Stratton *et al.* 2002). 'MUST' also has excellent reproducibility (κ 0·809–1·000) between users (nurses, health care assistants, doctors, nursing and medical students) in different health care settings across the UK (Elia, 2003; Stratton *et al.* 2003*a*). However, the concurrent

Page 801

801

Reference Standard = Criterion

validity of this tool needs investigation. Concurrent (correlational) validity involves comparison of a tool with another validated criterion measure or reference measure. A tool can have concurrent validity if it shows good to excellent agreement with other tools or with a reference standard, (e.g. assessed by κ, a chance-corrected measure of agreement). 'MUST' has been shown to have excellent agreement with a dietitian's assessment of malnutrition (Elia, 2003), but whether it has agreement with other previously published tools used in the UK in adults is unknown and requires study. Generally there is little information about the concurrent validity of other tools and when this has been established, it has usually been through comparison of only a few tools in one particular patient group and health care setting (e.g. Correia *et al.* 2003). A comparison of the ease of use of 'MUST' with other tools is also warranted. Therefore, the aims of the present series of studies were: (1) to compare the prevalence of malnutrition risk assessed by 'MUST' and a variety of other published tools in both hospital outpatients and inpatients; (2) to investigate the concurrent validity of 'MUST' with these other published tools and to assess whether the same patients are identified as malnourished; (3) to compare the ease of use of 'MUST' with these other published tools.

In addition to Cohen's kappa, the authors also present simple percentage agreement between measures. The percentage agreement ranges from low of 67% to a high of 92%. Most agreement levels are over 77%.

Table 3. Concurrent validity between the 'malnutrition universal screening tool' ('MUST') and other nutritional screening procedures in hospital and community settings*

Cohen kappa

Percent Agreement

Tool comparison	Categories (*n*)	Patients	*n*	Agreement %†	Agreement κ‡	SE
Community setting						
'MUST' *v.* MEREC	3	Outpatients	50	92	0·893	0·077
'MUST' *v.* HH	3	Outpatients	50	84	0·711	0·105
Hospital setting						
'MUST' *v.* NRS	3	Medical <65 years old	75	89	0·775	0·072
'MUST' *v.* NRS	2	Medical <65 years old	75	92	0·813	0·073
'MUST' *v.* MST	2	Medical <65 years old	75	88	0·707	0·091
'MUST' *v.* MNA-tool	2	Medical >65 years old	86	77	0·551	0·081
'MUST' *v.* MNA-tool	2	Surgical	85	80	0·605	0·083
'MUST' *v.* SGA	3	Medical	50	72	§	§
'MUST' *v.* SGA	2	Medical	50	92	0·783	0·102
'MUST' *v.* URS	3	Surgical	52	67	0·255	0·101
'MUST' *v.* URS	2	Surgical	52	77	0·431	0·130

MEREC, *MEREC Bulletin* tool; HH, Hickson and Hill tool; NRS, nutrition risk score; MST, malnutrition screening tool; MNA-tool, short-form mini nutritional assessment screening tool; SGA, subjective global assessment; URS, undernutrition risk score.

* For details of tools and procedures, see Figure 1, Table 2 and pp. 801–803.

† Percentage of patients placed in the same malnutrition risk category by the two tools. Disagreements in categorisation between tools were not systematically biased, except between MUST and MNA-tool in medical and surgical patients (two categories, $P=0·0005$) and MUST and URS (two categories, $P=0·039$).

‡ κ 0·400–0·750 fair–good; κ > 0·750 excellent agreement beyond chance (Landis & Koch, 1977).

Page 804

§ As the observer did not categorise any patients as high risk with SGA, κ was not calculated for the three category comparison.

Table 4, below, shows the counts of agreement and disagreement between MUST and another measure called malnutrition screening tool (MST), which provides only two categories, No Risk and Risk, so the authors combined MUST categories Medium and High into one and used Low Risk as the other category. For **concurrent validity** the issue is the degree to which MST No Risk agrees with MUST Low Risk, and MST Risk agrees with MUST Medium + High Risk.

The counts of agreement and disagreement are provided below in Table 4 and marked with red. For Low Risk and No Risk, the two scales agreed on 49 study participants, and for the Risk and Medium + High Risk, the scales agreed on 17 participants. The two scales disagreed on 5 participants (Risk with MST, but Low Risk with MUST) and 4 participants (No Risk with MST and Medium + High Risk with MUST). There were 49 + 17 = 66 agreements and 5 + 4 = 9 disagreements. In total, 75 participants were evaluated with both instruments, so the percentage agreement was 66 / 75 = .88 or 88%. That shows good agreement between the classification of these two scales which provides evidence for convergent validity of MUST with the criterion measure MST. As shown in Table 3, the Cohen kappa measure of agreement for these two measures is .707.

Table 4. Cross-tabulation of malnutrition risk according to the 'malnutrition universal screening tool' ('MUST') and the malnutrition screening tool (MST)*

Categories showing agreement, 49 + 17 = 66, so total agreement is 66 / 75 = 88%

		'MUST' (two categories)		
		Low risk	Medium + high risk	Total
Medical <65-year-old patients (*n* 75)†		*n*	*n*	*n*
MST (two categories)	No risk	49‡ =agree	4 =disagree	53
	Risk	5 =disagree	17 =agree	22
Total		54	21	75

* For details of tools and procedures, see Figure 1, Table 2 and p. 802
† κ 0·707. Similar total proportion identified as at risk by two tools (28 % 'MUST', 29 % MST), but individual patients categorised differently by the two tools.
‡ Agreements.

Page 805

(b) Predicting Future Behavior

This type of evidence – using scale scores to predict future behavior – is typically identified as **predictive validity**, as explained previously. And if theory is used to link measured scores with several future measured behaviors, then **construct validity** is also assessed.

Procedure:

- To illustrate this type of validity process, assume we are testing an instrument that measures **student adjustment to college**.
- First, administer the adjustment to college instrument to a sample of freshmen college students about halfway through their first semester at college.
- Second, after some time passes obtain scores from identified criteria. These criteria should be theoretically or logically linked to college adjustment. For example, one who adjusts well to college is more likely to
    - persevere through the first year, second year, third year, and eventually graduate;
    - obtain higher GPA averages for each year in college (freshman, sophomore, etc.);
    - have higher motivation to learn;
    - score higher on measures of college satisfaction;
    - have less depression symptoms while in college; and
    - pursue graduate degrees.
- Third, analyze relationships between instrument scores and the future criteria scores using whatever analysis is appropriate (e.g., Pearson correlation, t-tests, Cohen's kappa, etc.). If Pearson's correlation is used, or something similar like ICC, then then obtain results are called **validity coefficients**.
- Each of the criteria noted above – perseverance, GPA, motivation, etc. – could be measured at different times throughout their college career and correlated to initial student adjustment scores to learn whether those

adjustment scores correlated with criteria scores as hypothesized. If the relations found are as anticipated, that would provide evidence of predictive validity for the student adjustment to college measure.

Evidence from predictive usually results in weaker measures of association than found with convergent validity (i.e., correlating scores from two instruments designed to measure the same construct).

Example 1:

Morgan, R. (1989). Analyses of the Predictive Validity of the SAT® and High School Grades from 1976 TO 1985. ETS Research Report Series, 1989(2), i-16.

Purpose: To learn whether SAT (originally called scholastic aptitude test, now scholastic assessment test) scores predict Freshmen GPA (F-GPA). For many, the time difference between the SAT test and freshmen GPA, the criterion, is one or two years.

Note: See presentation video for discussion of this information.

Students normally complete the SAT during their junior and senior years of high school. Colleges and universities would require submission of SAT scores because, in theory, SAT scores help predict success in college. Combined with high school performance, HS GPA, these two predictors could be used to assess whether one is likely to do well in college.

Morgan provides the following table of correlations of freshman GPA (F-GPA) with SAT-V (verbal), SAT-M (math), multiple SAT (V and M combined), high school record (HS GPA), and multiple correlation (SAT-V, SAT-M, and HS GPA combined). Data from 10 years, 1976 to 1985 are provided. As Table 4 shows, SAT-V correlations with F-GPA range from a low of .32 to a high of .39, and SAT-M provides a similar range of correlations. The combined score for SAT has a range of correlations with F-GPA of .38 to .45. It appears that HS GPA is a stronger predictor of F-GPA with correlations ranging from .48 to .52.

Table 4. Estimates of the Yearly Correlations for SAT Scores and High School Record with F-GPA with Associated Standard Errors and Slopes of the Best-Fitting Lines for the Correlations Page 7

Pearson correlations

Year	*SAT-V*	*SAT-M*	*Multiple SAT*	*High School Record*	*Multiple Correlation*
1976	.39(.01)	.38(.01)	.45(.01)	.52(.01)	.58(.01)
1977	.37(.01)	.38(.01)	.44(.01)	.51(.01)	.57(.01)
1978	.37(.01)	.38(.01)	.44(.01)	.50(.01)	.57(.01)
1979	.36(.01)	.36(.01)	.42(.01)	.50(.01)	.56(.01)
1980	.35(.01)	.37(.01)	.42(.01)	.49(.01)	.56(.01)
1981	.35(.01)	.33(.01)	.40(.01)	.51(.01)	.56(.01)
1982	.36(.01)	.35(.01)	.41(.01)	.51(.01)	.57(.01)
1983	.34(.01)	.36(.01)	.41(.01)	.50(.01)	.56(.01)
1984	.32(.01)	.35(.01)	.40(.01)	.48(.02)	.55(.01)
1985	.32(.01)	.32(.01)	.38(.01)	.52(.01)	.57(.01)

Example 2:

Bothma, C. F., & Roodt, G. (2013). The validation of the turnover intention scale. SA Journal of Human Resource Management, 11(1), 1-12.

Purpose: This study was conducted to learn whether a reduced version of the Turnover Intention Scale with only 6 items, TIS-6, would provide reliable and valid scores. The sample consisted of 2,429 employees in an information technology company.

Note: See presentation video for discussion of this information.

The TIS-6 was administered to this sample and 4 months later data were collected from the company about who had left their jobs. In total, after 4 months 84 employees resigned. Bothma and Roodt argued that if the TIS-6 is predictive, then mean scores for turnover intention should be higher for those who resigned when compared to those who were still employed. Results of their analysis is presented below.

> There was a significant difference in the turnover intention scores of those employees who resigned (M = 5.14, SD = 1.26) compared to those who stayed (M = 4.13, SD = 1.28): $t(170) = 5.20$, $p \leq 0.001$ (two-tailed). The difference in the means (mean difference = 1.01, 95% CI: 0.63 to 1.39) has a large effect ($\eta_p^2 = 0.14$). This finding supports the criterion-predictive validity of the TIS-6 to predict actual turnover. **Page 7**

Results show that, after 4 months, those who resigned did have significantly higher scores on the TIS-6 than those who remained, thus lending predictive validity support to the TIS-6.

Bothma and Roodt repeated this analysis after 4 years at which time a total of 405 employees had resigned. Results are reported below. Again, mean scores for the TIS-6 were statistically higher for the group who resigned, although this time the difference was much smaller suggesting the predictive validity of the TIS-6 is not as strong as for the 4-month period. This suggests the TIS-6 is more predictive of immediate behavior than more distant future behavior.

> The data profiles of the 405 employees who resigned from the ICT company over the 4-year period *after* the survey was conducted were compared with the data profiles of 405 employees drawn randomly from the remaining sample (n = 2024) who stayed with the company. Independent-sample t-tests were conducted to compare the different variable scores of those employees who resigned versus those who stayed. The following analyses (displayed in Table 5) provide evidence that turnover intention scores can be used as a proxy for actual labour turnover. **Page 8**

Note that this example uses discrimination between two groups to examine **predictive validity** evidence. This fits the second definition of **discriminate validity**, that measured scores should be able to discriminate between known, or theoretically assumed, group differences.

(c) Theoretical Relations with Other Constructs

As described above, this process is consistent with **construct validity**, **convergent validity**, **divergent validity**, and **discriminate validity**. One should theorize how scores from the measured construct will correlate with both related and non-related constructs, then test those expectations to assess validity evidence.

Procedure:

- Identify all relevant constructs – the measured construct of interest and those that should be related and others that may only be weakly related or not related.
- Administer all instruments to a group of people at nearly the same time.
- Statistically analyze data to obtain estimates of relations among variables of interest.

Example 1

Menon, S. (2001). Employee empowerment: An integrative psychological approach. Applied psychology, 50 (1), 153-180.

Purpose: Menon developed a scale to measure employee empowerment. He defined empowerment to consist of three dimensions (or sub-scales): perceived control, perceived competence, and goal internalization. His purpose was to assess evidence for reliability and validity of scores.

Note: See presentation video for discussion of this information.

Below, in Table 3, are correlations between Menon's three dimensions, or sub-scales, and another measure with related dimensions. Do these correlations seem to match what one would expect?

Menon's sub-scales:

Perceived Control = amount of autonomy one has on job (e.g., can make decisions independently)
Perceived Competence = belief in one's skill and ability to perform job-related tasks
Goal Internalization = degree to which employee has adopted goals/objectives of the company

This presentation will focus on Perceived Control.

Criteria used to assess Perceive Control validity:

- Helplessness = likely to be inversely related, negative correlation expected since having control is opposite of helplessness
- Impact = like control, so should be positively related
- Self-determination = also like control, so should be positively related
- Competence = unclear how this is related, not sure correlation direction or strength, although not likely to be strong
- Meaning = to see value in work, see comment above for competence, relation with perceived control unclear

TABLE 3
Correlations of Empowerment Subscales with the Spreitzer Scale and Helplessness

	Empowerment subscales		
	Perceived Control	*Perceived Competence*	*Goal Internalisation*
Spreitzer scale (12 items, $\alpha = .84$)	.66	.32	.53
Helplessness (6 items, $\alpha = .86$) = have little power on job	−.74	−.17**	−.52
Spreitzer subscales			
Impact = can influence job decisions (3 items, $\alpha = .93$)	.75	.12*	.40
Self-determination = can initiate, have power on job (3 items, $\alpha = .74$)	.53	.16**	.33
Competence (3 items, $\alpha = .72$)	.13*	.66	.27
Meaning = sees value in work (3 items, $\alpha = .85$)	.30	.18**	.48

* $p < .05$
** $p < .01$
All other correlations significant at $p < .001$

page 168

Correlations for perceived control column are consistent with the expectations I formed above. Given the consistency with expectations, this provides good evidence for **construct validity**, and also **convergent** and **divergent validity**. Similar hypotheses could be formed for competence and goal internalization.

Example 2

Van Hooff, M. L., Geurts, S. A., Kompier, M. A., & Taris, T. W. (2007). "How fatigued do you currently feel?" Convergent and discriminant validity of a single-item fatigue measure. Journal of Occupational Health, 49(3), 224-234.

Purpose: Authors developed a single-item measure of fatigue:

"How fatigued do you currently feel?"
Response Options: 1 = "not at all" to 10 = "extremely"

Their purpose "was to establish the convergent and discriminate validity of a single-item measure of daily fatigue…in a daily diary context."

Below, in Table 1, they identify constructs that should display convergent validity, all with positive correlations, and four items that should display weak or no correlation and therefore evidence divergent validity (or as they label it discriminate validity).

Most variables in Table 1 should be identifiable with the names provided. However two may not be clear, so those are explained below.

POMS = profile of mood states, "well-validated instrument to measure fatigue" (p. 225)
Daily WHI = daily work-home interference, extent work bothers private, at-home time

Do the correlations, highlighted in the red column, demonstrate evidence of construct validity of the single-item fatigue measure?

Table 1. Means, standard deviations and correlations of the variables under study

	Measure	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
Convergent validity															
	1 fatigue report mark	1.60	0.44	1											
Daily	2 POMS	4.73	1.59	.80**	1										
	3 daily WHI	1.38	0.36	.45**	.53**	1									
	4 daily sleep complaints	1.40	0.88	.45**	.45**	.32**	1								
	5 daily work-related effort	5.25	1.75	.47**	.28**	.34**	.22*	1							
Global	6 global fatigue	1.89	0.59	.51**	.52**	.33**	.36**	.11	1						
	7 global health complaints	2.56	2.33	.35**	.43**	.29**	.30**	–.03	.66**	1					
	8 global WHI	1.02	0.42	.55**	.58**	.66**	.37**	.34**	.63**	.48**	1				
	9 global job pressure	2.44	0.51	.16	.19	.20*	.10	.27**	.06	.03	.26**	1			
Discriminant validity															
Daily	10 daily work pleasure	6.76	0.95	–.02	–.08	–.24*	–.06	–.08	–.23*	–.15	–.21*	–.09	1		
Global	11 global job control	3.23	0.43	–.12	–.11	–.28**	–.20*	–.05	–.17	–.13	–.28**	–.25**	.25*	1	
	12 global social support	2.46	0.69	.02	.08	–.17	–.10	–.07	–.12	–.21*	–.18*	–.07	.15	.18	1
	13 global motivation to learn	2.60	0.43	–.02	.04	–.04	–.07	.00	–.18*	.03	–.04	.21*	.26**	.13	.27**

Number (%) of missing values between 0 (0%) and 22 (18,3%); mean number of missing values=13 (10.8%).
*p<0.05; **p<0.01.

page 229

Van Hoof et al. argue that the results do support the validity of the single-item measure. Their discussion appears below (pp. 232-233).

Convergent validity

The results provided evidence for the convergent validity of the single-item fatigue report mark. First, crude correlations revealed a very strong association between the report mark and the alternative multiple item measure (POMS). This result was confirmed using multilevel analysis. Although this analysis also revealed some statistically significant variations in the strength of this association across the time of the day and days of the week, the relevance of these variations can be questioned, as they only explained one percent of additional variance.

Second, the fatigue report mark was substantially correlated with other, supposedly related, daily variables: daily work-home interference, daily sleep complaints and daily work-related effort. These findings were confirmed by means of multilevel analysis. Moreover, this latter analysis showed that the associations between the report mark and these daily variables were stable across the observation period, and, thus, did not depend on the day they were measured.

Finally, the fatigue report mark was related to three out of the four global variables included to investigate its convergent validity. It was substantially correlated with global fatigue, global health complaints and global work-home interference. No significant association was

Discriminant validity evidence

The results also support the discriminant validity of the single-item fatigue report mark, as it revealed only non-significant or weak relationships with measures supposed to tap constructs other than fatigue. Correlations show that this measure is not significantly related to daily work pleasure and multilevel analysis revealed only a weak negative association with this variable. The report mark was also unrelated to any of the global measures incorporated to examine discriminant validity (i.e., global job control, global social support, global motivation to learn).

7. Self-test

This section provides questions to help test yourself on some of the basic ideas of validity.

Question

If I tell you test scores are valid, what does this tell you about reliability?

Answer

If validity is present for scores, the scores must be reliable.

Question

If I tell you test scores are not valid, what does this tell you about reliability?

Answer

It tells us nothing about reliability since scores may be consistent but invalid (e.g., my scales reporting a weight of 35 every time, yet I weigh 200).

Question

If I tell you test scores are not reliable, what does this tell you about validity?

Answer

If the scores are not reliable, then the scores cannot be valid.

Question

If I tell you test scores are reliable, what does this tell you about validity?

Answer

Nothing is known about validity; reliability is a necessary condition for validity, but not sufficient. We need more information to show validity. For example, my scales can be very reliable showing me with a weight of 100, 100, and 100 three times in a row, but this is far from my true weight, hence the scales give reliable scores, but not valid scores.

Question

Which is this, empirical or logical validity?

A researcher designs an instrument to measure test anxiety and includes items to measure both psychological (thoughts and worry) and physiological (physical reactions such as sweating and heartbeat) components because research and theory suggest that both components are important to measure when assessing test anxiety.

Answer

Logical Validity: No data were collected; instead, the researcher provides a logical reason why certain items were included on the scale.

Question

Which is this, empirical or logical validity?

After developing a set of items to measure both psychological and physiological components of test anxiety, several experts in psychology are asked to review critically each item and decide if it appears to measure test anxiety. Their feedback will be used to determine which items to retain or eliminate.

Answer

Logical Validity: No predictions about how instrument scores will behave were tested; instead, the researcher collets expert opinion about the usefulness of each item.

Question

Which is this, empirical or logical validity?

A researcher expects that if scores from the test anxiety scale truly measure test anxiety, then there should be a negative correlation between test anxiety scores and mathematics final exam scores, i.e., the higher test anxiety, the lower mathematics final scores. The Pearson correlation between these two variables was $r = -.49$.

Answer

Empirical Validity: Tested the prediction that there would be a negative association between anxiety and math scores. This is an example of evidence for **construct validity** and **concurrent validity**.

Question

Which is this, empirical or logical validity?

Previous research shows that females tend to display more test anxiety that males. Scores from the test anxiety instrument were collected and compared by sex, and results of a t-test show that females tended to have higher levels of test anxiety (Females $M = 8.75$, Males $M = 5.67$, $t = 3.19$, $p < .05$) immediately before a mathematics final exam.

Answer

Empirical Validity: Tested whether there would be a sex difference in test anxiety consistent with previously reported findings about test anxiety. This is an example of evidence for **construct validity**, and **discriminate validity** according to some authors.

Question

Which is this, empirical or logical validity?

A researcher is developing a test to measure students' achievement in educational research. The researcher notes that the following topics were covered in class: hypotheses, variables, sampling, and statistics. For each of these four areas, the researcher writes a total of 6 questions: that is, 6 questions covering hypotheses, 6 on variables, 6 on sampling, and 6 on statistics, for a total of 24 questions. After writing the questions, the researcher has a colleague read each question to ensure that each is appropriate for educational research and appears to cover something relevant toward hypotheses, variables, sampling, or statistics.

Answer

Logical Validity: Test items developed according to a rubric or content domain, and evaluated according to expert opinion, but scores from instrument were not tested empirically.

Question

Which is this, empirical or logical validity?

The researcher who developed the educational research achievement test noted above obtained scores from this test and from another already-validated achievement test of educational research to see if the scores align for each student tested. Each student completed both tests. A Pearson r showed a correlation of r = .61 between scores of both tests for each student.

Answer

Empirical Validity: Tested whether there would be a correlation between two instruments, one newly developed and one already developed and validated - this is an example of evidence for **convergent validity** and **concurrent validity**.

Another Example of Empirical Validity

One could hypothesize that achievement test scores should be positively related to number of hours studied. If the scores do show a positive correlation to number of hours studied, then that provides some evidence for empirical validity, specifically **construct validity**. In addition, one could further hypothesize that achievement in educational research would be positively related to logical reasoning ability. If one was able to administer a test of logical reasoning ability and then correlate the scores from the educational research achievement test with scores from the logical reasoning test, and if the correlation was positive as expected, then this would also be evidence for the **construct validity** (and **concurrent** and **convergent** validity) of the scores from the educational research test.

Question

If the validity coefficient equals r = .89, is this strong or weak evidence for concurrent validity?

Answer

This is strong evidence for concurrent validity. Scores from both instruments are displaying a similar pattern.

Question

If r = .13, is this good evidence for concurrent validity?

Answer

If r = .13 then that is a very weak correlation, so little evidence of concurrent validity.

Question

Suppose we have two measures of happiness: Scale A and Scale B. Look carefully at the scale response options below.

Happiness A using the following response scale:

1 = very unhappy
2 = unhappy
3 = so-so
4 = happy
5 = very happy

Happiness B using the following response scale:

1 = very happy
2 = happy
3 = so-so
4 = unhappy
5 = very unhappy

Example items:

Scale A

	Very Unhappy	Unhappy	So-so	Happy	Very Happy
My life overall	1	2	3	4	5
My work environment	1	2	3	4	5
My marriage	1	2	3	4	5

If this person was generally happy, how would they respond to Scale A?

Possible responses = 4, 4, 5 = Mean response of about 4.33

Scale B

	Very Happy	Happy	So-so	Unhappy	Very Unhappy
How I see myself	1	2	3	4	5
Experiences with colleagues	1	2	3	4	5
Day-to-day life at home	1	2	3	4	5

If this person was generally happy, how would they respond to Scale A?

Possible responses = 1, 2, 1 = Mean response of about 1.33

The two Happiness instruments use reversed scaling response options for their respective items.

If a group completed both instruments, what correlation should be expected if both Happiness instruments produce valid scores?

Answer

Correlation should be **negative** between scores from these two Happiness instruments due to the reversed scaling responses employed.

Question:

Now suppose the correlation between the two instruments is -.89, is this strong or weak evidence for concurrent validity?

Answer

Despite being negative, Pearson's r = -.89 represents strong evidence. Interpretation of correlations (or any scores) requires that one must understand what the scores represent. Sometime high scores indicate positive attributes, and sometimes low scores indicate positive attributes. It is not as simple as saying "all validity coefficients must be positive" – sometimes that just isn't the case, as this example illustrates.

Question

If we wish to establish the predictive validity of SAT scores, how might we do that?

Answer

To provide evidence for predictive validity of SAT scores, the typical approach is to obtain SAT scores from students who took the SAT during their high school years, then correlate those scores with freshmen GPA in college. If the SAT is predictive of collegiate success, then there should be a positive correlation between SAT and future GPA.

Question

What is the criterion in this SAT example?

Answer

Future GPA would be the criterion by which we judge the predictive ability of SAT scores.

Question

If we wish to establish evidence of validity for GRE scores, what criterion might we use?

Answer

Since GRE scores are used to screen for graduate school admission, we should expect to find that GRE scores can usefully predict graduate school performance. Thus, one possible criterion to assess predictive validity of GRE scores would be graduate school GPA. Another would be whether GRE score differences exist between those who graduate or fail to graduate from graduate school.

References

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. Measurement and evaluation in Counseling and Development, 36(3), 181-191.

Meloy, J. R., Hoffmann, J., Roshdi, K., & Guldimann, A. (2014). Some warning behaviors discriminate between school shooters and other students of concern. Journal of Threat Assessment and Management, 1(3), 203.

Menon, S.T. (2001). Employee empowerment: An integrative psychological approach. Applied Psychology: An International Review, 50, 153-180.

Newsome, S., Day, A. L., & Catano, V. M. (2000). Assessing the predictive validity of emotional intelligence. Personality and Individual differences, 29(6), 1005-1016.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. Psicothema, 26(1), 108-116.

Van den Broeck, A., Vansteenkiste, M., De Witte, H., Soenens, B., & Lens, W. (2010). Capturing autonomy, competence, and relatedness at work: Construction and initial validation of the Work-related Basic Need Satisfaction scale. Journal of occupational and organizational psychology, 83(4), 981-1002.

Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. Journal of vocational Behavior, 65(1), 71-95.