**02: Brief Review of Reliability and Validity**

**1. Reliability: Defined**

 Reliability has two meanings, that measured scores are **consistent**, and/or measured scores **agree**. Consistency is the degree to which sets of scores show similar patterns, and agreement is the extent to which scores are similar, i.e., differences between scores are small in absolute value. Reliability is the degree to which a measuring device provides consistent scores or scores that agree. Reliability does not address whether scores represent what they are supposed to measure – that is an issue of validity – instead, reliability only assesses whether scores are consistent, in agreement, or both.

Example 1

My weight as recorded by home scales on Thursday morning upon three successive measurement attempts:

205

205

205

These weights are very consistent and in agreement.

Reliability can vary by **degree**.

Example 2

My weight as recorded by home scales on Friday morning upon three successive measurement attempts:

205

206

204

This set of measurements show less consistency and agreement than the weights on Thursday, so less reliability.

Example 3

My weight as recorded by home scales on Saturday morning upon three successive measurement attempts:

112

285

59

Much less consistent and much less agreement, so no reliability. Maybe I broke my scales yesterday.

**2. Reliability: Coefficient Values**

 Since reliability can vary by degree, indices of reliability should reflect this and provide an indication of the extent to which scores demonstrate high or low reliability. There are several methods for calculating reliability, some of which will be described and illustrated in presentations that follow on test-retest and internal consistency.

In most cases the coefficients calculated to measure reliability range from a low of 0.00 to a high of 1.00 with scores closer to 1.00 indicating greater reliability. In general, we can use the following scale for interpreting reliability coefficients, e.g., Pearson r or intra-class correlation coefficient (ICC) for test-retest, and Cronbach’s alpha for internal consistency.

|  |  |
| --- | --- |
| Coefficient Value |  |
| 0.90 to 1.00 | Excellent reliability; expected for professionally develop tests uses for placement and evaluation |
| 0.70 to 0.90 | Acceptable, good reliability for research purposes and typical educational settings |
| 0.60 to 0.70 | Usable level of reliability for research, but also questionable |
| 0.00 to 0.60 | Generally unacceptable  |

**3. Reliability: Consistency vs. Agreement**

 **Consistency** refers to the relative position of scores across two sets of scores. Consistency is an assessment of whether two sets of scores tend to rank order something in similar positions. **Agreement** refers to the degree to which two sets of scores agree or show little difference in actual scores; the lower the absolute difference, the greater the agreement between scores.

Pearson r is designed to provide a measure of consistency. Loosely described, this means we may use Pearson’s r to assess whether relative rank appears to be replicated from one set of scores to another.

Pearson r does not assess magnitude of absolute differences and can therefore present a misleading assessment of reliability when test-retest scores, parallel-forms scores, or multiple rater/judge scores show large differences. As Example 4 shows, Pearson r is .91 between the two sets of scores for Relative Reliability. A Pearson r of .91 suggests an excellent level of test-retest reliability. Note, however, that the actual scores are very different (Mean for Test 1 = 77.50, mean for Test 2 = 16.62) with about a 50-point difference, on average.

As a contrast to Relative Reliability, or Consistency, note scores in the section of Example 4 labeled Absolute Reliability, or Agreement. With these scores there is little difference from administration 1 and 2, so the scores display good agreement and therefore high reliability (i.e., Pearson r = .98).

Example 4: Relative vs. Absolute Reliability

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Relative Reliability,Consistency |  | Absolute Reliability, Agreement |
| Student |  | Test 1 | Rank 1 |  | Test 2 | Rank 2 |  | Test 1 | Test 2 | Difference |
| 1 |  | 95 | 1 |  | 44 | 1 |  | 95 | 92 | 3 |
| 2 |  | 90 | 2 |  | 22 | 2 |  | 90 | 91 | -1 |
| 3 |  | 85 | 3 |  | 20 | 3 |  | 85 | 83 | 2 |
| 4 |  | 80 | 4 |  | 19 | 4 |  | 80 | 79 | 1 |
| 5 |  | 75 | 5 |  | 10 | 5 |  | 75 | 78 | -3 |
| 6 |  | 70 | 6 |  | 9 | 6 |  | 70 | 72 | -2 |
| 7 |  | 65 | 7 |  | 8 | 7 |  | 65 | 64 | 1 |
| 8 |  | 60 | 8 |  | 1 | 8 |  | 60 | 61 | -1 |
|  |  | Test 1 and 2 Pearson r = .91 |  | Test 1 and 2 Pearson r = .98 |

In some situations, one desires a measure of consistency. For example, when comparing student performance on the ACT and SAT, a measure of consistency would be helpful to know whether the general ranking, or relative position of students, remains similar despite the ACT and SAT having different scoring scales. If raters are asked to independent rate something, such as observed anti-social behavior, and if the raters develop and use different rating scales, Pearson r could assess whether scores obtained from the two rating scales and raters provided similar relative ratings of those observed for anti-social behavior.

When comparing parallel scales or tests, or when assessing stability of scores from a scale or test, a preferred measure is one that considers both relative performance (consistency) and absolute performance (score agreement). Pearson r does not provide a measure that addresses both conditions. As Example 4 shows, Pearson’s r works for judging consistency, whether two sets of scores tend to rank people or objects similarly but fails as a measure of test-retest agreement when scores show little agreement. For judging agreement, one uses the Intraclass Correlation Coefficient (ICC). We will learn about the ICC in the next set of notes.

**4. Reliability: Methods for Assessing Reliability**

 There are many formal ways to determine reliability of scores. Below, briefly described, are a few of the more commonly employed approaches. Methods for calculating reliability will be presented in separate presentation notes.

**Test-retest Reliability**

Test-retest reliability assesses the **stability** of scores across time; the more stable scores over time, the greater the consistency, so the greater reliability. Note the term **stability** used above. Test-retest reliability estimates are also known as the **coefficient of stability** since test-retest scores should be stable (unchanging) over time.

Test-retest reliability establishment steps:

* Administer the **same measuring device** (or some test, scale, inventory, measuring device) to the **same group** of people on two different occasions.
* Let appropriate time elapse between administrations (scale dependent).
* Obtain scores from the two occasions.
* Correlate the **paired scores** using Pearson’s correlation coefficient, or, better, the ICC since the ICC can be used to assess agreement of scores across time (i.e., determine whether scores appear to be similar across time).

**Paired scores** – each person will provide two scores, one score from the first administration and the second score from the second administration. Without paired scores, it is impossible to calculate test-retest or parallel forms reliability. Example 5 below shows paired scores for four individuals.

Example 5: Paired Scores

|  |  |  |
| --- | --- | --- |
| Person | Time 1 Score | Time 2 Score |
| Bryan | 81 | 83 |
| Marijke | 93 | 91 |
| Gunther | 88 | 88 |
| Marlynn | 90 | 92 |

Mean scores must also be similar in test-retest, and same is true for equivalent-forms reliability. If mean scores between sets of scores differ, this suggests the scores from the two sets do not agree and therefore do not display test-retest reliability. As noted, Pearson r cannot detect mean differences in scores, but the ICC can and therefore should be used.

Scores from studies in which a scale or test is administered in a **pre-test** and **post-test** format are inappropriate for test-retest types of reliability assessments.

Question

Why is it inappropriate to assess test-retest reliability from a pretest to posttest study?

Answer

With pre-test to post-test types of studies, test-retest reliability is not ideal because test-retest assumes scores should remain stable, consistent, over time. However, pre-test/post-test studies usually implement a treatment between the pre-test and the post-test, and usually this treatment is designed to change scores from pre to post. If one expects, because of the treatment, that scores will change between pre and post, then test-retest is not appropriate since scores should change and therefore not exhibit stability.

Question

Suppose an instrument measures something that is highly variable over short time periods (such as test anxiety), would test-retest be appropriate for that variable?

Answer

No, test-retest reliability is designed to determine the stability of scores over time, so something that fluctuates over time is inappropriate for test-retest reliability. Instruments designed to measure highly variable constructs, such as test anxiety, may not be suitable for test-retest reliability since test-retest focuses upon stable traits, such as IQ.

**Equivalent-forms Reliability**

Equivalent-forms (also called parallel-forms or alternate forms) reliability is designed to assess whether two forms of a scale or test provide similar scores. The two forms of the scale or test should measure the same thing, the same construct or variable, but have differently worded items.

Examples of Similar Test Items

Example 1: Assessing Basic Multiplication Understanding

Form A

What is 5 × 7?

Form B

What is 4 × 8?

Example 2: Assessing Reliability Coefficient Understanding

Form A

Which reliability estimate is strongest?

.35

.56

.81

Form B

Which reliability estimate is weakest?

.21

.49

.75

Example 3

If you took the ACT, SAT, or GRE more than once, you would not take the same form of the test, but all forms supposedly provide similar scores for an individual. Thus, anyone who took the SAT twice is likely to get similar scores for each subtest (e.g., 50 verbal first time, 52 verbal the next time).

The procedure for establishing equivalent-forms reliability is nearly identical to the procedure for test-retest reliability.

 Equivalent-forms reliability establishment steps:

* Have two (or more) forms of an instrument (scale, test, etc.)
* Administer the **both forms** to the **same group** of people at roughly the same time (e.g., within a few hours or days).
* Obtain scores from the two or more forms.
* Correlate the paired scores using Pearson’s correlation coefficient or use the ICC which is better since it also can be used to assess agreement of scores.

Question

Why is it important to administer both forms to the same group of people?

Answer

True score assessment works by assessing the true score for the same individuals on both forms – true scores should not change within a person. Correlation and ICC require paired scores, so scores must be paired by participants otherwise any deviation between pairs of scores will be due to differences in participants and not due to differences in forms. Also, paired scores eliminate a large source of variability – individual differences across participants in performance.

Question

Two forms of a test are administered to the same group of people. Scores from both forms are correlated. Suppose the correlation is very high (r = .95), does this mean the two forms are equivalent?

Answer

Correlation coefficient of .95 is very high and would indicate high level of equivalence between forms. However, recall the difference in reliability between equivalence and agreement. It is possible for scores to be highly correlated, yet means be very different -- if that is the case, then the forms are not equivalent.

Example

Listed are five people and the scores they received on form A (first score) and form B (second score).

These scores show a similar pattern, yet there is little to no agreement, however the Pearson r = .98.

|  |  |  |
| --- | --- | --- |
| Participant | Form 1 | Form 2 |
| Person A | 90 | 50 |
| Person B | 80 | 40 |
| Person C | 70 | 30 |
| Person D | 50 | 20 |
| Person E | 30 | 10 |

The equivalent forms reliability would be .98, but as the scores are not equivalent between Form 1 and Form 2, so the forms are not equivalent despite the high correlation. If we use the ICC to estimate agreement, the reliability coefficient as estimated by the ICC would be much lower than .98.

As you can see, a high Pearson r is not enough with equivalent forms (or test-retest) – these forms must also produce similar mean scores too.

In sum, to show equivalent forms reliability, one must show that the pattern of scores from two tests are similar (have a high correlation), and one must also show that scores from the two tests are similar, that is, they must have close agreement.

Equivalent-forms reliability is not practical for most research situations since it is difficult, impractical, or disruptive to administer two forms in many educational research studies.

**Internal Consistency Reliability**

One of the most practical and employed forms of reliability estimation is **internal consistency** which refers to the level of consistency in responses to items that are designed to measure the same latent variable or construct. **Construct** refers to a conceptualized variable that is measured using responses to items designed to measure that construct. Thus, one constructs scores for the conceptualized variable, or latent variable, by forming a **composite score** from several items.

Examples of constructs include IQ, mathematics self-efficacy, life satisfaction, work autonomy, and science motivation. Each of these conceptualized variables are formed by taking a composite score from responses to many items or indicators (recall that items designed to measure a construct or latent variable are known as indicators).

Internal consistency means that items designed to measure the same latent variable produce scores that are similar. Stated differently, if items are internally consistent, then if we know how someone responds to one item, then we can predict how they are likely to respond to another item.

Example 1 of Internal Consistency

Test Anxiety Items appear below. **Logical Consistency Check:** To assess whether items are likely to generate internally consistency responses, do the following:

* Assume you have high levels of test anxiety (take an extreme position to test whether items provide similar scores)
* Answer each item
* Determine whether responses are similar for each item

If each of the items generates similar responses (e.g., scores of 5 or 6), then the items are generating internal consistent responses.

**Instructions**: Please indicate, on the scale provided, how true each statement is for you **immediately before taking an important test**.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|   | Not True of Me |   |   |   |   | Very True of Me |
| 1. I have an uneasy, upset feeling. | 1 | 2 | 3 | 4 | 5 | 6 |
| 2. I'm concerned about doing poorly.  | 1 | 2 | 3 | 4 | 5 | 6 |
| 3. I'm thinking about the consequences of failing. | 1 | 2 | 3 | 4 | 5 | 6 |

The construct of test anxiety could be formed by taking the overall mean response to the three items. So, for example, maybe student Beth responded to items 1, 2, and 3 with these scores: 5, 4, 5.

For Beth, her test anxiety composite score would be (5+4+5)/3 = 4.66

Internal Consistency establishment steps:

* Have several items, must be more than one item, designed to measure a single construct.
* Administer items a group of participants.
* Calculate internal consistency on responses to all items designed to measure that single construct from each participant. Note: One should NOT calculate internal consistency for a total instrument since not all items are designed to measure the same construct.
* There are several ways to calculate internal consistency, but Cronbach’s alpha most common.

Cronbach’s Alpha (α)

* an advanced form of the split-half reliability method and a more general form of the KR formula for reliability
* ranges from 0.00 to 1.00
* closer to 1.00 the more consistent are responses from items
* closer to 0.00 the less consistent are responses from items

Question

What is the cut-off level one usually expects for scores to be judged internally consistent?

Answer

For most research purposes Cronbach’s alpha should be .70 or larger, and for professionally developed instruments (such as SAT, GRE, CRCT), one expects alpha to be .90 or better. This is true for any measures of reliability considered – test-retest, ICC, Cronbach’s alpha.

**Scorer/Rater Reliability (or, more accurately termed Agreement)**

**Intra-judge reliability** refers to the consistency with which one judge assigns scores to the same assessment, **inter-judge (inter-rater)** reliability refers to the consistency with which two or more judges assign scores to the same assessment.

For intra-judge, one judge must score something more than once to establish intra-judge reliability. To establish intra-judge, two or more scores of the same assessment are needed to show consistency.

Question

A teacher reads an essay and scores it based upon predefined scoring criteria as defined within a scoring rubric. For example, using an essay grading rubric, the teacher reads Bryan’s essay and scores it an 8 out of 10.

Does this establish intra-judge reliability? Explain why or why not.

Answer

To show reliability, one must show that scores are similar, so **more than one score** on that essay is needed to assess intra-judge reliability. The teacher must read Bryan’s essay once and score it, then let time pass and read Bryan’s essay again and score it again. If the two essay scores are similar, say an 8 and 9 out of 10 for the two readings, that would show agreement. However, if the scores are dissimilar, such as 8 and 4, then that shows little agreement so there is low intra-judge reliability and this suggests problems with scoring essays.

Inter-judge (inter-rater) consistency is determined by comparing scores from two or more judges to learn whether the scores are similar/agree -- if the scores are similar, then there is evidence for inter-judge reliability. Key is that scores must be on the same observations (e.g., same essay, diving event, etc.).

Example

Olympic judges for high dive competition rates one dive from one diver and we examine whether those ratings are similar.

**5. Validity: Defined**

 Recall that **measurement** is defined as the systematic process of assigning scores or labels to categories of a variable. The measurement process may produce scores (e.g., statistics test 1 scores, GRE scores, science self-efficacy scores, etc.) or labels (e.g., her sex is female, his political party affiliation is Tory, he is unemployed, that flower type is daisy, etc.).

In the broadest sense **validity** is the degree to which a measurement process produces scores or labels that reflect what measurement process was intended to measure.

For example, if measuring student skill in multiplication, scores from a test of multiplication are more valid if they truly measure multiplication skill than skills in reading, spelling, addition, or any other content. We expect students with strong understanding of multiplication to receive high scores on the test while those with weak understanding of multiplication should correspondingly receive low scores on the test.

As another example, if we use a scale to measure job satisfaction, then we expect that those who obtain high scores on the scale should have high levels of job satisfaction and those who obtain low scores on the scale should have low levels of job satisfaction. If measured scores from the scale do not reflect the actual job satisfaction experience of employees, then the scale lacks validity.

In short, validity is the degree to which obtained scores, or labels, reflect what they were designed to measure.

**6. Relation between Validity and Reliability**

 **Reliability** is the degree to which a measuring device produces scores that display consistency or agreement. **Validity** is the degree to which a measuring device produces scores that represent or measure that for which the device was designed to measure.

Question

How do validity and reliability differ?

Answer

* Reliability only means that scores **agree** and are **consistent**.
* Validity means that the scores **measure what they were designed to measure**; scores **accurately** reflect what they should be measuring.
* Scores may be **reliable but not valid** (my scales show my weight consistently at 35 lbs., but this weight is not accurate for me because I weigh 200lbs; a clock always reports the time as 10:00 is perfectly consistent, but not valid)
* Scores must be **reliable to be valid** – no way to show a valid measure of my weight if scores are inconsistent, e.g., I weigh 200, but scales are inaccurate: 185, 255, 177, 221, etc.)

Question

Is it possible to have reliability without validity?

Answer

Yes, we can have reliability without validity. For example, assume I weigh 200 pounds. If I step on my scales repeated one morning and see these weights, what do we know about the scale?

95

95

95

Scale produces scores that agree and is consistent (reliable), but not accurate (invalid), so there is reliability without validity.

Question

Is it possible to have empirical validity without reliability? Empirical validity means data are collected to assess validity.

Answer

One cannot have validity without consistency of scores. For example, if I weigh myself on scales and step off, then step on again, and I repeat this action four times, I might get scores such as:

254

 65

233

186

There is no consistency to these scores, and there is no accuracy across the scores either. Without **some** consistency one cannot have validity of measurement.

**7. Validity Assessed: Logical and Empirical Evidence**

 Reliability for a sample can be quickly assessed using one or more of the procedures outlined above (e.g., test-retest, internal consistency), but validity requires a multi-step process that includes both logical validity and empirical validity.

What is the difference between empirical and logical validity?

* **Empirical validity** means one collects data to tests 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 negative with math achievement scores (e.g., Pearson r = -.33, this supports validity evidence for test anxiety although additional evidence would likely be needed).
* **Logical validity**, also known more commonly as **content** **validity**, means one uses logical arguments or reason, rather than data, to suggest instrument items will provide useful scores. Logical validity refers to the overall process used to develop items for the test or scale. Logical validity does not employ hypothesis testing of how instrument or test scores will behave. Instead, one explains the theory and rationale for why particular items are included; explains the various dimensions of constructs measured and how items match those dimensions; and gets input from others about the adequacy of questionnaire/scale/test items.
* Empirical is typically the more critical of the two, but both are needed to build confidence in a scale or test.

Methods for establishing logical and empirical validity will be presented in a separate presentation.

Examples of Logical Validity

(a) When developing a scale of measure test anxiety, a research first identifies a clear definition of test anxiety as suggested by theory and reason, and then determines dimensions of test anxiety (e.g., psychological and physiological reactions one might have when anxious about tests) as suggested by theory, reason, and prior research.

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

(c) After developing a set of items to measure both psychological and physiological dimensions of test anxiety, several experts in psychology are asked to review critically each item and decide if that item appears to measure at least one of the dimensions of test anxiety. Their feedback will be used to determine which items to retain, revise, or eliminate.

(d) An instructor is developing a test to measure students’ achievement in educational research. The instructor 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 instruction 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.

Examples of Empirical Validity

(a) 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. This tested prediction of a negative relation between test anxiety and final exam scores is an example of evidence for **construct validity**.

(b) Previous research shows that females tend to display more test anxiety that males. Scores from the newly developed test anxiety scale 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. This tested prediction of sex differences in test anxiety is an example of **construct** **validity**.

(c) The instructor above who developed the educational research achievement test administered two tests to his students. The first test he constructed as described above in the Logical Validity example, and the second test was one previously developed and validated in another course on educational research. Both tests were designed to assess knowledge of the same content: hypotheses, variables, sampling, and statistics. Scores from both tests were correlated with Pearson’s r and the correlation produced was r = .71. In this example the instructor 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 **concurrent validity** and **convergent** **validity**.

Review of Relation between Validity and Reliability

Question

If scores are valid, then what does this imply reliability?

Answer

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

Question

If scores are not valid, then what does this imply reliability?

Answer

Nothing can be determined 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 scores are not reliable, then what does this imply validity?

Answer

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

Question

If scores are reliable, then what does this imply about validity?

Answer

Nothing is known about validity; reliability is a necessary condition for validity, but not sufficient. More information is needed 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.

**References**

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.