Concept of Reliability

The concept of reliability is of the consistency or precision of a measure

- Weight example

Reliability varies along a continuum, measures are reliable to a greater or lesser extent

Not an all or nothing quality

Concept of Reliability

The opposite of consistency and precision is variability due to *random measurement error*

Reliability is lack of random measurement error

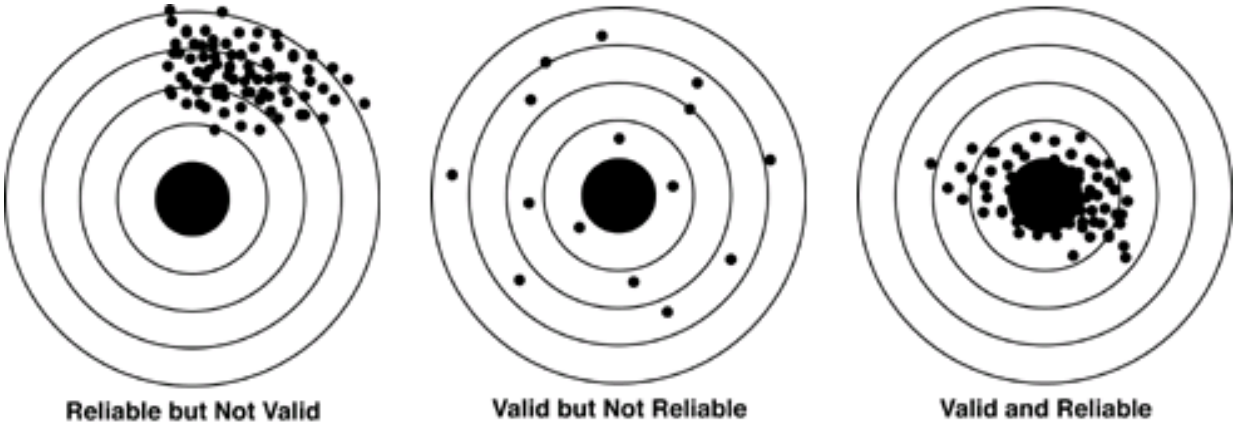
Random error is unexplained variation that is *not systematic*

If variability is random, there will be some overestimates and some underestimates

On average estimate is accurate

Concept of Reliability

Target analogy



http://ccnmtl.columbia.edu/projects/qmss/measurement/validity_and_reliability.html

Theoretical Foundations

Classical Test Theory (CTT)

Observed Score	=	*True Score*	+	*Error*
X_o	$=$	X_t	$+$	X_e

Note: many texts use $X = T + E$

Theoretical Foundations

Reliability is the proportion of the observed score variance, s_o^2, that is due to the true score, s_t^2

The smaller the error variance, s_e^2 ,the greater proportion that is due to true score variance and the higher the reliability

If proportion is 1.0, then no error variance – perfect reliability

If proportion is 0.0, then all error variance – no reliability and all noise

Theoretical Foundations

$$\text{Reliability} = \frac{\text{True}}{\text{True} + \text{Error}}$$

$$R_{xx} = \frac{s_t^2}{s_t^2 + s_e^2}$$

$$= \frac{s_t^2}{s_o^2}$$

Note: your text uses R_{xx} as the symbol for reliability but most texts use ρ_{xx} (rho) or r_{xx}

Attenuation

Measurement error *attenuates* correlations

- Imagine if a score was only random error
- If observed scores are a function of true scores and measurement error, degree of error will cloud estimation of the relationship between two variables
- Example: child's age and reading ability

Attenuation

Remember that measurement error will increase the variance of the observed score, so the denominator in the correlation coefficient will be larger

This makes the estimate of the correlation smaller in magnitude

$$r_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{c_{xy}}{s_x s_y}$$

Attenuation

$$r_{x_o y_o} = r_{x_t y_t} \sqrt{R_{xx} R_{yy}}$$

$r_{x_o y_o}$ is the correlation estimated from the data (between observed scores), $r_{x_t y_t}$ is the correlation between the true scores (if we could know them), R_{xx} and R_{yy} are the reliabilities of the two measures

Attenuation

Example 1: say the reliability for my guess at the age is .6 and the measurement of reading ability is .5 and that the true score correlation is .4

$$r_{x_o y_o} = r_{x_t y_t}\sqrt{R_{xx}R_{yy}}$$

$$= .4\sqrt{(.5)(.6)} = .4\sqrt{.3} = .4(.548) = .21$$

When the true score correlation is .4, the estimated correlation is .21—a substantial underestimate—almost half the value!

Attenuation

Example 2: say the reliability for my guess at the age is .9 and the measurement of reading ability is .9 and that the true score correlation is .4

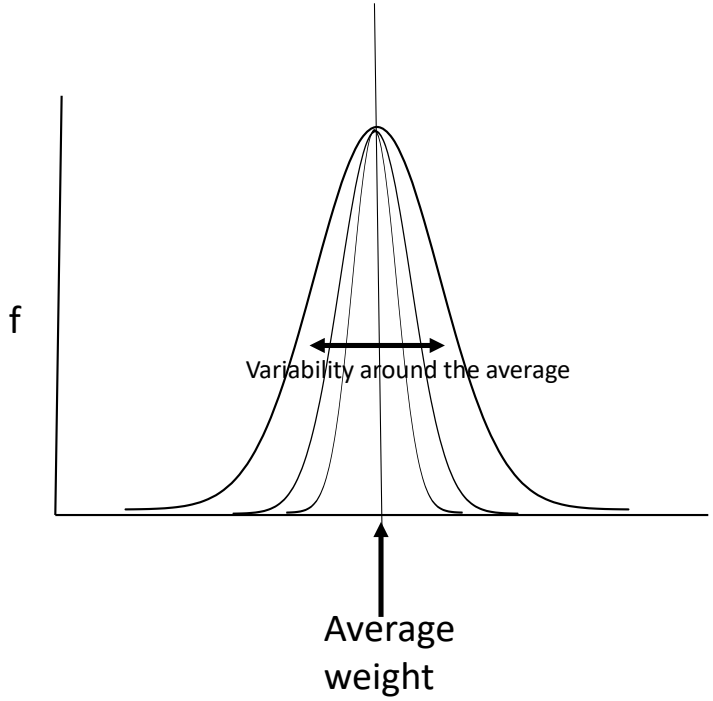
$$r_{x_o y_o} = r_{x_t y_t}\sqrt{R_{xx}R_{yy}}$$
$$= .4\sqrt{(.9)(.9)} = .4\sqrt{.81} = .4(.9) = .36$$

When the true score correlation is .4, the estimated correlation is .36—not nearly as bad

Means

Remember that random measurement error sometimes leads to overestimates and sometimes leads to underestimates

On average the estimate will be accurate



Means

Comparing means

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If X_1 and X_2 observed scores have larger variance (s_1^2 and s_2^2) than their true score counterparts, then the denominator will be larger and the t will be smaller, so ***less likely to be significant***

Means

Comparing means

$$d_{x_o} = \frac{\left|\bar{X}_{o1} - \bar{X}_{o2}\right|}{\sqrt{\frac{s_{o1}^2 + s_{o2}^2}{2}}}$$

Also seen in the estimate of the effect size, which gives the magnitude of the group difference (where $o1$ and $o2$ subscripts indicate observed values for group one and two)

Estimating Reliability

Test-retest reliability

- Repeat the test two or more times to see how similar the measurements are
- Calculate the correlation between the measurement occasions
- Problem is that in the interval between the measurement occasions the attribute may have changed
- Small time interval needed in between measurements without contamination from recall

Estimating Reliability

Parallel tests

- Two tests are parallel if their true scores are the same and they have the same standard deviation
- Theoretical notion, because it is not possible to know with absolute certainty that two tests are exactly parallel

Estimating Reliability

Alternative forms reliability

If we could create two parallel or alternative forms of a measure, we could estimate reliability of the measure without repeated measurements

e.g., standardized tests, like the SAT and GRE, use alternative test forms

Estimating Reliability

Split-half reliability

- Can develop a larger test and correlate two halves
- Problem is how best to split up the test
  - e.g., what if the first half and second half differ?

Estimating Reliability

Domain sampling theory (model)

- What if we considered a set of items from a test to be from a larger pool (domain, population) of items from the same test
- We could think of every item as a small parallel test, a *testlet* or *subtest*

Estimating Reliability

Domain sampling theory (model)

If we view each item as good representations of the true score and each as a random selected item from a domain or population of possible items, then we can relax the assumption that each test is strictly parallel

Instead we only need to think of them as on average equally representing the domain

Estimating Reliability

Internal reliability

The domain sampling idea allows us to use the correlations among items to gauge the reliability of a measure

This is the basis of *internal reliability*, such as the type of reliability assessed by Cronbach's alpha

Cronbach's Alpha

Preliminary steps

- Generate descriptive statistics, including means, standard deviations (and/or variances, skewness and kurtosis)
- Obtain frequency tables and histograms
- Check for errors in entry, coding, etc.
- Variables do not need to be normally distributed, but when they are highly skewed or kurtotic or they respondents have not used the full range of values, you may want to consider the wording of that item.
- Check correlations to confirm scoring direction is correct and potentially eliminate items that are supposed to correlate that do not

Cronbach's Alpha

Cronbach's alpha (Cronbach, 1951) is an estimate of internal reliability (sometimes called the "consistency coefficient")

Conceptually based on the proportion of true score to total observed score variance

$$R_{xx} = \frac{s_t^2}{s_t^2 + s_e^2} = \frac{s_t^2}{s_o^2}$$

Cronbach's Alpha

If we can estimate the proportion of the observed score variance that is due to measurement error, then we can estimate reliability

$$\text{Proportion error} = 1 - \frac{s_e^2}{s_o^2}$$

Cronbach's alpha (α) raw score form is:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_X^2}\right)$$

k = number of items, s_i^2 is the variance for each item, and s_x^2 is the variance for the composite scale score (as a sum of the items)

Cronbach's Alpha

The domain sampling model conceptualizes the items (testlets or subtests) as retests, so that the average correlation between these subtests is a measure of reliability

Cronbach's alpha in the standardized form is:

$$\alpha = \frac{k\overline{r}_{ii'}}{1+(k-1)\overline{r}_{ii'}}$$

$\overline{r}_{ii'}$ is the average correlation among all pairs of items, and k is the number of items

Cronbach's Alpha

The standardized coefficient alpha is the alpha for the set of items after they have been standardized (converted to z-scores) and will be equal or higher than the raw score version

Raw score alpha assumes the variances of the of the items are equal, and if they are not, the raw score estimate will be smaller than the standardized estimate

Usually similar, but when items are on very different scales (e.g., some 5-point and some 9-point scales), the difference may be larger

Cronbach's Alpha

Composites scores calculated by the sum or mean tend to weight items with larger variances more heavily

Standardizing items before computing the composite will equally weight them, because variances are all equal to 1

In most applications, researchers do not bother to do standardize items, sometimes because the original metric is lost (e.g., average of items on a 7-point no longer between 1 and 7, but are z-score values instead)

Cronbach's Alpha

What is an acceptable alpha? Exceeding .70 is widely mentioned as a cutoff for acceptable reliability, but what is "acceptable" or "good" depends heavily of the consequences of using a measure with some certain level of reliability.

Many scales with an alpha of .70 can be improved, however.

And this value has been grossly over applied and over stated.

Cronbach's Alpha

The .70 criteria is commonly attributed to Nunnally (1978), a highly regarded psychometrician, but using.70 as a standard was clearly not his intention:

> what a satisfactory level of reliability is depends on how a measure is being used. In the early stages of research . . . one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .70 or higher will suffice. . . . In contrast to the standards in basic research, in many applied settings a reliability of .80 is not nearly high enough. In basic research, the concern is with the size of correlations and with the differences in means for different experimental treatments, for which purposes a reliability of .80 for the different measures is adequate. In many applied problems, a great deal hinges on the exact score made by a person on a test. . . . In such instances it is frightening to think that any measurement error is permitted. Even with a reliability of .90, the standard error of measurement is almost one-third as large as the standard deviation of the test scores. In those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard. (pp. 245-246)

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill. Quote also given by: Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria what did they really say?. *Organizational research methods*, *9*(2), 202-220.

Kuder-Richardson 20 (KR_{20})

The KR20 (Kuder & Richardson, 1937) is a special case of Cronbach's alpha when the items are binary (e.g., yes/no or correct/incorrect)

It is equivalent to the raw score form of Cronbach's alpha, so computation of α for a set of binary items will give the same result as the KR_{20}

Cronbach's Alpha: Some Properties

- Cronbach's alpha is an estimate of internal reliability or consistency and does not indicate stability over time necessarily
- Alpha is a *lower bound estimate* of reliability, and actual reliability may be higher
- Alpha is equal to the estimate of reliability from all possible split halves
- Alpha assumes unidimensionality—if the measure really assesses more than one hypothetical construct (or factor), the estimate may be incorrect (lower than for each factor)

Cronbach's Alpha: Some Properties

- A more heterogeneous group will have a higher alpha than a more homogeneous group, all other things equal
- Speeded tests may inflate alpha (Lord & Novick, 1968), related to the homogeneity phenomenon above
- Test length affects alpha—longer tests are more reliable
  - Consider a single-item test vs. multiple item test
  - Think about domain sampling—larger sample of items should be a better estimate of the population of items

Cronbach's Alpha: Some Properties

Spearman-Brown prophecy formula

$$R_{xx-revised} = \frac{nR_{xx-original}}{1+(n-1)R_{xx-original}}$$

n is the factor by which the size is increased

If length is increased from a 10-item test is increased to 20 items (with the same average inter-item correlation), n = 2, because the length is increased by a factor of 2

Cronbach's Alpha: Some Properties

Spearman-Brown prophecy formula

If length is increased from a 10-item test is increased to 20 items (with the same average inter-item correlation), n = 2, because the length is increased by a factor of 2. Assume the original reliability $R_{xx\text{-original}}$ is .6.

$$R_{xx-revised} = \frac{nR_{xx-original}}{1+(n-1)R_{xx-original}}$$

$$= \frac{2(.6)}{1+(2-1).6} = \frac{1.2}{1.6} = .75$$

Cronbach's Alpha: Some Properties

Spearman-Brown prophecy formula (using average inter-item correlation)

$$R_{XX} = \frac{k\bar{r}_{ii'}}{1 + (k-1)\bar{r}_{ii'}}$$

k is number of items, and $\bar{r}_{ii'}$ is the average inter-item correlation

Cronbach's Alpha: Some Properties

Spearman-Brown prophecy formula (using average inter-item correlation)

$$R_{XX} = \frac{k\bar{r}_{ii'}}{1+(k-1)\bar{r}_{ii'}}$$

$$= \frac{5(.4)}{1+(5-1).4} = \frac{2}{2.6} = .77$$

Cronbach's Alpha: Some Properties

Spearman-Brown prophecy formula (using average inter-item correlation)

$$R_{XX} = \frac{k\bar{r}_{ii'}}{1+(k-1)\bar{r}_{ii'}}$$

$$= \frac{20(.4)}{1+(20-1).4} = \frac{8}{8.6} = .93$$

Cronbach's Alpha: Some Properties

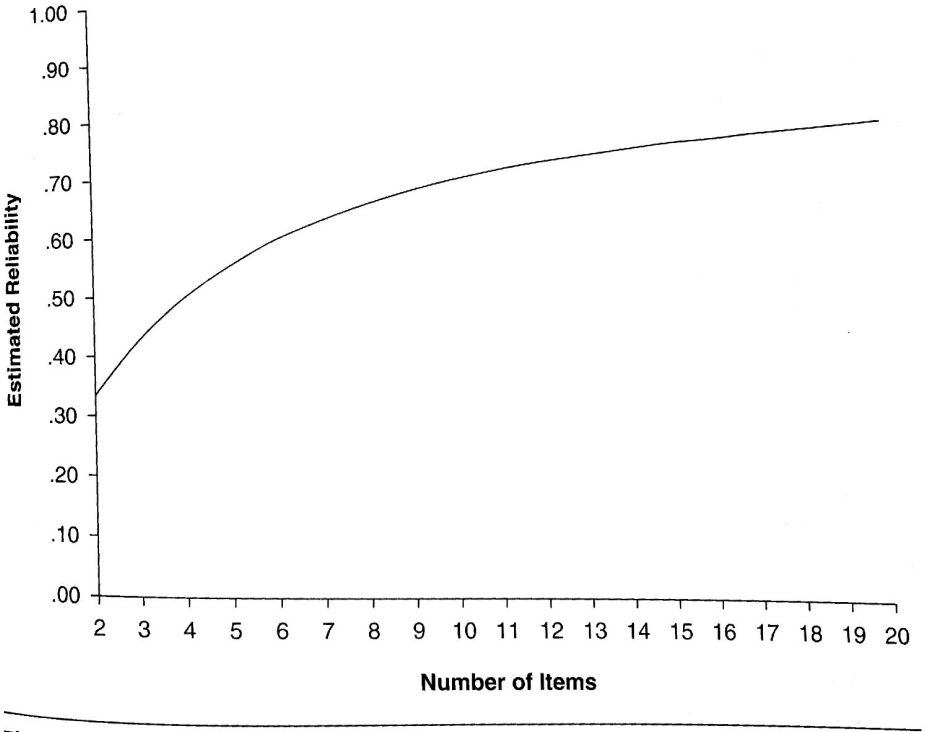


Figure 6.3 The Association Between Number of Items and Reliability (for a Test With an Average Interitem Correlation of .30)

Furr & Bacharach (2014, p. 151)

Cronbach's Alpha: Some Properties

- Does not indicate that alpha is "biased" by the number of items, but it may be difficult to reach acceptable reliability with short scales even if inter-item correlation is fairly high
- Longer scales may still have high reliability even though some items are not so good
- Good idea to also look at average inter-item correlation and item-total statistics because of the sensitivity to length