Research Methods (HDFS 3390),
Alan Reifman, Texas Tech University
Reliability and Validity

RELIABILITY: A measure consistently yields the same result.

- If you took a ruler and repeatedly measured your Smartphone, you would always get the same length.
- Ruler measurements of a physical object are perfectly reliable.
- Social scientific measurements of people (e.g., self-esteem, relationship conflict) are not perfectly reliable.
- People may be more or less tired, more or less able to concentrate, in a better or worse mood, etc., when filling out questionnaires.

VALIDITY: A questionnaire or test measures what it intends to measure. In other words, it measures accurately.

- We determine how valid a test is by comparing people's test scores to their success on real-world tasks in the relevant domain.
- For example, because the SAT seeks "to measure a student's potential for academic success in college," [1] the SAT's validity could be shown if high SAT-scorers obtained high college GPA's and low SAT-scorers obtained low college GPA's.
- Other examples:
  - eHarmony seeks to help singles form successful romantic relationships. To assess the validity of the eHarmony method, the company has compared eHarmony couples to couples who got together in other ways on their relationship satisfaction.
  - For many occupations, the state of Texas requires an exam to get a license. Examples include barbering and cosmetology. How do we know if the barber/ cosmetologist tests are valid? In other words, what real-world criteria could we use to see if high exam-scorers are actually good at styling and cutting hair?

---

[1] Kobrin and colleagues (2008), cited in Espenshade and Chung (2010).

Even Some Well-Established Measures Have Questionable Validity

The article *BMI [Body Mass Index] is a Terrible Measure of Health* provides an example:

The goal … should be to identify people with excess fat, since that fat has been associated with bad health outcomes. But the BMI is a function of a person's weight and height. Weight includes fat, but it also includes bones, muscle, fluids and everything else in the body... [One study] found that 47 percent of people classified as overweight by BMI and 29 percent of those who qualified as obese were healthy as measured by [other indicators such as blood pressure and cholesterol]... Using BMI alone as a measure of health would misclassify almost 75 million adults in the U.S., the authors concluded.

BMI calculator

Measures Must Be Reliable and Valid to Be Used in Research

- **A measure can be reliable (consistently yields the same result), but not valid (not measuring what it's supposed to be measuring). In other words, a measure can be consistent, but consistently bad.**
  - **Using the previous slide's example, Body Mass Index is probably reliable, but of "terrible" validity as a measure of health.**
- **If a measure is valid, it probably is reliable, as well. An unreliable (inconsistent) measure would have a hard time predicting real-world outcomes.**
- **Target-shooting analogy**

Both Reliability and Validity are Based on the Correlation Statistic (How do two variables go together?)

Positively correlated: As one variable goes up, so does the other. They both follow the same pattern. Knowing where a person stands on one variable, you know roughly where he/she stands on the other (maximum = +1.0).

- ***Example:* The more hours one studies before a test, the higher the score he or she will likely get.**

Not at all correlated (zero correlation): Knowing where a person stands on one variable tells us nothing about where he/she stands on the other. Someone who has a high score on one variable is equally likely to have a high or a low value on the other.

- ***Example:* A person's number of sneezes per week is (probably) uncorrelated with the percent of a person's shirts that are blue.**

Negatively correlated: As one goes up, the other goes down. They follow an *inverse* pattern. Knowing where a person stands on one variable, you again know roughly where he/she stands on the other (minimum = -1.0).

- ***Example:* The higher the winter temperatures where one lives (e.g., Miami), the fewer the heavy jackets people buy.**

Graphical depictions of positive, zero, and negative correlations. Note that the correlation (symbolized *r*) is based upon the slope of the best-fitting line (line which comes closest to all the points) and degree to which points are close to the line vs. being scattered.

A song to nail down our understanding of correlation, best-fit lines, upward and downward slopes, etc.

Fitting the Line

Lyrics by Alan Reifman

(May be sung to the tune of "Draggin' the Line," James/King)

Plotting the data, on X and Y,
Finding the slope, with most points nearby,
We want to find the angle, of the trend's incline,
Fitting the line (fitting the line),

Upward slopes make r positive,
Slopes trending down, make it negative,
From minus-one to plus-one, r can feel fine,
Fitting the line (fitting the line),
Fitting the line (fitting the line),

Points align, how will the data shine?
If you have upward slopes, it'll give you a plus sign,
Fitting the line (fitting the line),
Fitting the line (fitting the line),

How strongly will your variables relate?
Is there a trend, or just a zero flat state?
You want to know what your analysis will find,
Fitting the line (fitting the line),
Fitting the line (fitting the line),

Points align, how will the data shine?
Your r will be minus, if the slope declines,
Fitting the line (fitting the line),
Fitting the line (fitting the line),

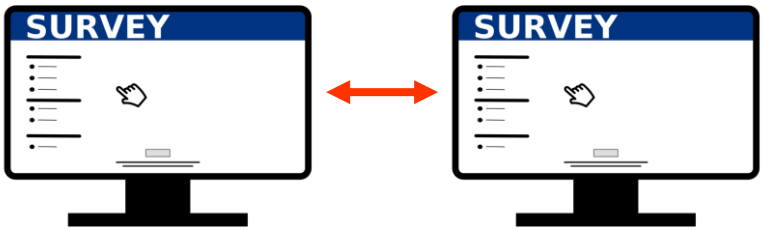
(Guitar solo)

Points align, how will the data shine?
If you have upward slopes, it'll give you a plus sign,
Fitting the line (fitting the line),
Fitting the line (fitting the line)…

Facebook album of Dr. Reifman meeting Tommy James, after the latter's concert at the 2013 South Plains Fair.

Different Types of Reliability (Consistency) for Different Research Strategies

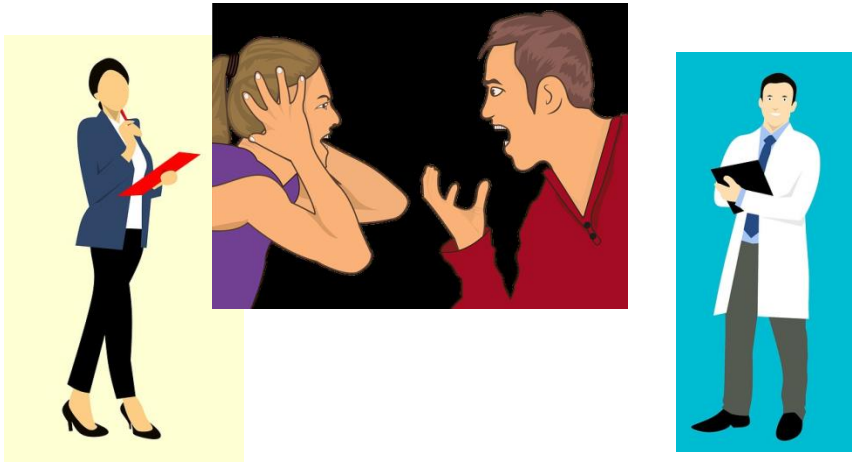
TEST-RETEST RELIABILITY



- Give same measure to people twice, days, weeks, or months apart.
- Compute correlation between scores at Time 1 (T1) and Time 2 (T2).
- Large positive correlation* tells us people who scored highly at T1 also scored highly at T2 (and those with low scores at T1 also had low scores at T2).
- CONSISTENCY OVER TIME

*As one example, test-retest reliability correlations for people who took the SAT more than once have been found to be .77 for whites and .90 for blacks (Vars & Bowen, 1998; in Jencks & Phillips, *The Black-White Test Score Gap*, p. 471, footnote 22).

INTER-RATER RELIABILITY



- Have two raters independently view interaction of couple or family (live or video).
- Each rater counts behaviors such as interruptions, affirming statements, etc., or gives global ratings (e.g., 1-10 scale) on warmth, conflict, etc.
- Compute correlation between scores of Rater 1 and Rater 2.
- CONSISTENCY ACROSS RATERS

Images: Mohamed Hassan & Mustafa Shehadeh, https://pixabay.com/

One More Type of Reliability, for When a Test is Given on Only One Occasion (i.e., Can't use test-retest)

INTERNAL CONSISTENCY (ALPHA, α): When there is high internal consistency (maximum = 1.0), how a person answers any one item tells you how he/she answered the others.

Hendrick & Hendrick Love Attitudes Scale

(Shown are the highest and lowest reliability subscales from one particular study, out of the full set of 6 subscales; the others are Eros, Ludus, Pragma, and Agape)

Storge

(love that started as friendship)

Alpha (α) = .83

Correlations between items are fairly high

	A	B	C	D
A. Our love best b/c it grew from friendship	--			
B. Our friendship merged into love	.62	--		
C. Our love really a deep friendship	.42	.36	--	
D. Satisfying/ developed from good friendship	.80	.67	.40	--

Mania

(obsessive/ possessive love)

Alpha (α) = .65

Correlations between items much lower

	A	B	C	D
A. When partner doesn't pay att'n, I feel sick	--			
B. Trouble concentrating	.28	--		
C. Can't relax if I think partner w/ someone else	.38	.25	--	
D. If partner ignores, get attention back	.39	.24	.38	--

These results are from an undergraduate research project by Matt McCord, a former 3390 student.

The more you agree with the statement in Item A, the more you should agree with that in Item B, etc. That's what we mean by "internal consistency."

Zeller and Carmines (1980) state in *Measurement in the Social Sciences*: "In general, as the average correlation among the items increases and as the number of items increases, alpha takes on a larger value" (p. 56; Table 3.2A).

Types of Validity

(From most to least important, in Dr. Reifman's view)

TYPE	DEFINITION	EXAMPLE
Predictive (Criterion Related)	Test scores should correlate with real-world outcomes (as we discussed at beginning of this unit)	SAT (V) & first-year grades correlation = .36; SAT (M) & first-year grades correlation = .35
Construct: *Convergent*	Test should correlate with other similar measures administered in a single testing session (if you can't follow participants for months in the real world)	SAT should correlate with other academic ability tests
Construct: *Discriminant*	Test should <u>not</u> correlate with irrelevant variables	SAT should not correlate with political attitudes
Content	Covers the necessary range of material	Math test should cover arithmetic, algebra, geometry, trigonometry, etc.
Face	Items look like they are covering proper topics	Math test should not have history items

Source for SAT validity correlations: David Owen (with Marilyn Doerr), *None of the Above: The Truth Behind the SATs* (1999, revised and updated edition; p. 197)

Reliability correlations tend to be much larger than validity correlations.

Why might this be so?

RELIABILITY (test-retest)

VALIDITY (predictive/criterion)

Another song…

Reliable and Valid

Lyrics by Alan Reifman

(May be sung to the tune of "Don't Stop [Thinking About Tomorrow]," Christine McVie, for Fleetwood Mac)

When selecting a questionnaire,
Psychometrics have to be sound,
You can make your own, if you have to,
But try to use one already around,

Make… it… re-liable and valid,
Make… it… the best that you can find,
It will help, strengthen your research,
Measurement's prime, measurement's prime,

(Guitar solo)

To assess re-li-a-bility,
Use test-retest with two occasions,
Use alpha for a one-time test, and,
Inter-rater for observations,

Make… it… re-liable and valid,
Make… it… the best that you can find,
It will help, strengthen your research,
Measurement's prime, measurement's prime,

(Guitar solo)

To assess a test's validity,
There are many forms to make your case,
They may or may not be statistical,
Predictive, construct, content and face,

Make… it… re-liable and valid,
Make… it… the best that you can find,
It will help, strengthen your research,
Measurement's prime, measurement's prime,

Make… it… re-liable and valid,
Make… it… the best that you can find,
It will help, strengthen your research,
Measurement's prime, measurement's prime,

Ooh, make your tests sound,
Ooh, make your tests sound,…
(Fade out)

More Advanced Resources for Graduate Students

- **Discussion of reliability and validity in terms of Classical Test Theory (Lesa Hoffman)**
- **Useful chapter:** John, O. P. & Benet-Martínez, V. (2000). Measurement, scale construction, and reliability. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339-369). New York, NY: Cambridge University Press.