

## LIES, DAMNED LIES...

Rick Grieve  
PSY 562  
Western Kentucky University  
(with a huge THANKS to Tim Thornberry)

1

---

---

---

---

---

---

---

---

**“STAT” IS A FOUR-LETTER WORD...**

“There are three kinds of lies: lies, damned lies, and statistics.”  
– Mark Twain

Stats are important and they can be tricky

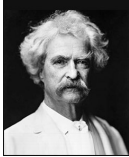
You have to know them in order to know good psych tests

Be a little skeptical when you hear people using stats

Be on the lookout for misleading terms or misrepresented data

People can use stats to push their agenda – this is true of test developers, too

[http://www.huffingtonpost.com/peter-h-gleick/misrepresenting-climate-s\\_b\\_819367.html](http://www.huffingtonpost.com/peter-h-gleick/misrepresenting-climate-s_b_819367.html)



2

---

---

---

---

---

---

---

---

**WHAT ARE PSYCH TESTS?**

Carefully chosen samples of behavior

A numerical or categorical system of measure is applied according to pre-established standards to describe or make inferences about behavior

There are few if any **constants** in psychology – a lot of times, we are dealing with **variables**

3

---

---

---

---

---

---

---

---

## TWO TYPES OF VARIABLES

**Discrete** – finite ranges of values (or infinite, but **countable** range)

- Can be dichotomous (sex, yes or no, coin toss), polytomous (gender identity, race, marital status)

**Continuous** – infinite ranges, really can't be counted, can divide into infinity

- A lot of these in psych (degree of depression, extraversion)
- Measurements are **approximations** – there is **error** that we need to consider!
- We need to estimate the effects of error and inform patients about error – stats can help us do this

4

---

---

---

---

---

---

---

---

## MEASUREMENT

Before we use stats, we need to know how to measure our construct or behavior of interest

Application of **rules for assigning numbers to objects or attributes**; transforms the qualities of attributes into numbers

To measure something, we need some kind of **scale** (height, shoe size, ranking system)

We have to agree on a particular scale or way to measure something

- E.g., if I say I am 1,803 mm tall, that's not very meaningful

Systematic measurement allows for easier confirmation and analysis of phenomena = more objective, more scientific

So, it's important to know the pros and cons of different types of measurements, different scales



5

---

---

---

---

---

---

---

---

## PROPERTIES OF SCALES

### 1. Magnitude

- Property of "moreness"
- A particular instance of an attribute represents more, less, or equal amounts compared to another instance
- E.g., height (tall, taller, tallest; first, second, or third place;  $5 > 2$ )

### 2. Equal Intervals

- A scale has this property if the difference between two points at **any** place on the scale has the same meaning as the difference between two other points
- E.g., height, weight (difference between 100 and 105 lbs is the same as 130 and 135 lbs)
- Straight-line (or linear) relationships
- \*Psychological tests **rarely** have this property (IQ does not have equal intervals)

6

---

---

---

---

---

---

---

---

### PROPERTIES OF SCALES (CONT'D)

#### 3. Absolute Zero

- NONE of that property exists (hard to find psychological constructs with an absolute zero)
- Temperature in Kelvin, time, weight

Based on these 3 properties of scales, we can derive 4 types of scales

- Nominal, ordinal, interval, ratio

7

---

---

---

---

---

---

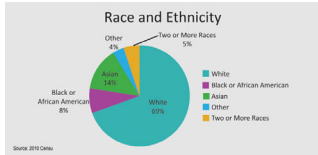
---

---

### TYPES OF SCALES

#### 1. Nominal

- No magnitude, no intervals, no absolute zero
- Serves only to **identify** or name objects – categorical data
- Race, gender, marital status, occupation, home ownership
- e.g., 0 = Single ; 1 = Married; 2 = Divorced
- We cannot run certain descriptive or inferential stats with this type of scale
  - Mean marital status = 1.43 – what does that mean?
- What we can do: **Frequency distributions, percentages**



8

---

---

---

---

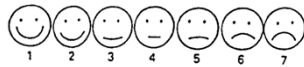
---

---

---

---

### TYPES OF SCALES (CONT'D)



#### 2. Ordinal

- Has **magnitude**
- No equal intervals, no absolute zero
- E.g., rank orders, IQ, self-report pain scales
- You cannot say anything about what the differences between two ranks means (we cannot say how far apart first place was from second place and assume the same difference occurred between second and third place)
- \*Most often, we have to use ordinal scales because we cannot say for certain whether there are equal intervals between points on the scale

9

---

---

---

---

---

---

---

---

### TYPES OF SCALES (CONT'D)

#### 3. Interval

- Scale has **magnitude and equal intervals**
- No absolute zero
- Now we can do some meaningful, interpretable **arithmetic** (Adding)
- E.g., measuring temperature in Fahrenheit or Celsius
- Not a ratio, no absolute zero, so you cannot say 60 degrees is twice as hot as 30 degrees




---

---

---

---

---

---

---

---

10

### TYPES OF SCALES (LAST ONE, I PROMISE)

#### 4. Ratio

- Has **magnitude, equal intervals, and an absolute zero**
- Temperature in Kelvin, speed, sports stats, age
- There is an absolute zero, a point at which **none** of the property exists (0 K = no molecular activity or no heat)
  - Hard to do this with emotions, thoughts, behaviors...
- You can say that I am driving twice as fast as you or that hitting .300 is twice as good as hitting .150 in baseball
- In psych testing, frequencies and reaction times can be ratios




---

---

---

---

---

---

---

---

11

### WHY ARE WE DOING THIS?

Knowing what scale you're working with helps you decide what stats are appropriate

Knowing the limits of measurement in psych testing is important for creating tests or selecting and using tests

Knowing how tests are created and items selected helps you understand tests better and pick good ones!

---

---

---

---

---

---

---

---

12

## TEST DEVELOPMENT

Mostly done by publishing companies (time and money)

- Imagine if you tried to create the WAIS/WISC all by yourself...
- Individuals can submit ideas, though...
- Identify a need

It all starts with the items

- The means by which we sample a test taker's behaviors
- A test is only as good as its items
- Test quality depends on the quality, number, and sequencing of items
- Tests (and test items) are designed for a given purpose

Item analysis – all the tools we can use to evaluate item quality during test development

- Qualitative and quantitative approaches

13

---

---

---

---

---

---

---

---

## QUALITATIVE ITEM ANALYSIS

Rely on reviewer judgment

Evaluate substance and style of items, their accuracy and fairness

Major criteria used

- 1. Appropriateness of item content and format for given purpose and population
- 2. Clarity of expression
- 3. Grammar
- 4. Adherence to basic rules the field has established over time
- Also, minimize bias or offensiveness

14

---

---

---

---

---

---

---

---

## QUANTITATIVE ITEM ANALYSIS

Stats!

Evaluates psychometric characteristics of items based on responses obtained from a sample during test development

Item analysis can occur throughout test development

A lot of criticisms of intelligence tests, specifically, but not limited to, the Wechsler tests, have come about because of quantitative item analysis.

15

---

---

---

---

---

---

---

---

## TEST PLANNING

You decide you need a new test (one does not exist or is not adequate for your needs)

You discuss the new test's **purpose**, the rationale for it, and what sort of inferences should be drawn from the proposed test

You then specify:

- Constructs or knowledge domains the test will assess
- Target population
- Objectives of test items (within overall objective of the entire test)
- Concrete means of gathering and scoring behavior samples
  - Administration methods, formatting of test item stimuli and responses, scoring procedures...

16

---

---

---

---

---

---

---

---

## TEST DEVELOPMENT

After test planning

Generate an item pool, generate administration and scoring procedures

Submit item pool to reviewers for qualitative item analysis

Revise and replace items as needed

Try out items on samples representative of the target population

Evaluate results of trial administrations of the item pool using quantitative item analysis and additional qualitative analysis

Add, delete, and/or modify items

**Cross-validate:** administer to a new sample to see if stats remain stable

17

---

---

---

---

---

---

---

---

## TEST DEVELOPMENT (CONT'D)

Standardize or fix test length and item sequencing

Finalize administration and scoring procedures

Administer test to another new sample

- This one needs to be carefully selected to represent the target population
- Develop norm data, reliability and validity evidence, and item-level analysis stats for the final test

Publish test, administration and scoring manual, and technical manual with stats

\*All these steps may take years! They may need to be repeated!

Tests need to be updated over time so they do not become obsolete, to update norms (Flynn effect), to match changing performance criteria, to prevent cheating, to make the test better...

18

---

---

---

---

---

---

---

---

### ITEM TYPES: SELECTED- AND CONSTRUCTED-RESPONSE ITEMS

- 1. **Selected-Response Items** (aka objective or fixed-response)
  - Close-ended, present limited number of alternatives for test taker to choose
  - MC, T/F, ranking, matching, rearrangement of options, forced choice
  - Typically scored pass/fail (but partial credit is possible)
  - Fairly popular in the US for the past few decades, but not without criticism

Pros: Easy, objective scoring; helps reliability; efficient use of test time; answers can be transformed into a numerical scale for easy quantitative analysis

Cons: Susceptible to guessing (and therefore error) or careless mistakes (both are unrelated to the construct of interest), lying; item construction takes lots of skill, time, and knowledge; less flexible

19

---

---

---

---

---

---

---

---

---

---

### ITEM TYPES (CONT'D)

- 2. **Constructed-Response Items** (aka free response)
  - Open-ended, writing samples, free oral responses, performances, products, essays, fill-in-the-blanks
  - Need thorough instructions and procedural rules for standardization
    - Time limits, medium, names, and length of required response; allowed outside sources
  - Often used in interviews, questionnaires, projectives, behavior observations
  - Usually less-focused, more global assessments

Pros: richer samples, wider range of answers, more creative, elicit authentic behavior (not just choices)

Cons: lower reliability and validity; practical constraints; subjective, complex scoring; poorer standardization, usually shorter (which hurts domain sampling); varying response length affects scores

20

---

---

---

---

---

---

---

---

---

---

### NEW DIRECTIONS IN ITEM ANALYSIS: ITEM RESPONSE THEORY (IRT) AND LATENT TRAIT THEORY

New hotness in psych testing, made possible by technology

Can be used with pencil-paper tests, computer-based tests, and computerized adaptive testing (CAT)

- CAT = individually tailoring item sequences to test taker ability based on prior responses

Not perfect, complex, some professionals don't even know about it

21

---

---

---

---

---

---

---

---

---

---

## GOAL OF TEST ITEM DEVELOPMENT: ITEM VALIDITY

Quantitative and qualitative item analysis tries to document this – does the test item contribute to the test's **purpose**? Does it **discriminate** between test takers as intended?

**Item difficulty** – varies relative to test taker ability level

- May include word difficulty (frequency they are used in a language)
- Normative analysis – what proportion of test takers in the norm sample got it right? (Percentage Passing or p)
- Often transformed to z scores (interval rather than ordinal) – if normal distribution
- Also depends on distractors
- Speed vs power tests

Two approaches for analyzing item difficulty and discrimination (**analyzing item-test regression and IRT**)

22

---

---

---

---

---

---

---

---

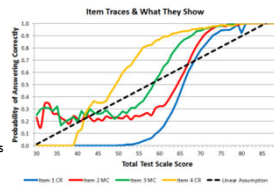
## ITEM-TEST REGRESSION

Calculate the proportion of test takers at each total score level who passed a given item

Visually see how each item functions across scores

Informative, but crude

Dependent on samples and item sets from which the data are obtained



23

---

---

---

---

---

---

---

---

## ITEM RESPONSE THEORY (IRT)

Refers to a variety of newer models for developing and testing tests

More sophisticated, but requires certain statistical assumptions and more extensive data collection

Contrast with **classical test theory (CTT)**

- Old, busted, traditional psychometric methods of test development and evaluation
- Still widely used
- Focuses on total scores on a test rather than individual items

24

---

---

---

---

---

---

---

---



## GOALS OF IRT

Generate items that provide the **max amount of info** possible concerning the ability level of examinees

Give examinees items **tailored** to their ability level

**Reduce number of items needed** to assess the test taker's ability level (without hurting reliability)

- Saves time, money, and frustration of the examinee

---

---

---

---

---

---

---

---

25

## PROBLEMS WITH CTT (THERE ARE PROBLEMS WITH IRT, TOO)

Indexes of item difficulty and item discrimination are **group dependent**

- When testing with different groups, the item difficulty and discrimination values might change (IRT estimates are assumed to not change, thus minimizing group differences on tests)

Fixed-length test scores are **test-dependent**

- Scores are a function of the specific items selected to be included in the test
- Makes comparisons of tests difficult

Stat problems: Reliability of scores gauged using **SEM**, which is assumed to be equal for all examinees (not a good assumption)

---

---

---

---

---

---

---

---

26

## ITEM FAIRNESS

Bias can occur against individuals or groups of test takers

"Fairness" is debatable

Also **qualitative and quantitative** analysis methods

- **Qualitative** – panels of demographically heterogeneous individuals with expertise
  - Stereotypical depictions of subgroups, offensive content, appropriate representation of diverse subgroups, language
- **Quantitative** – look to see if individuals from different groups with the same level of a trait differ in the probability of responding to an item in a certain way ("**differential item functioning**")

---

---

---

---

---

---

---

---

27

## THE FUTURE IS NOW!

### Computerized test item generation

- Plug in your theoretical model, your constructs you want to measure (with a healthy dose of syntax and logic)

### Automated essay scoring (AES)

- Can you imagine using this for WIAT Essay Comp?!

---

---

---

---

---

---

---

---

28



## RELIABILITY AND VALIDITY

---

---

---

---

---

---

---

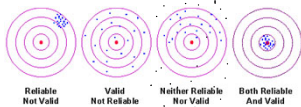
---

29

## RELIABILITY AND VALIDITY

We want tests that are both reliable and valid

- What is reliability?
- What is validity?
- Hopefully, the quizzes you just took are reliable and valid!
- Why did I do that to you? Why do I make you take quizzes and exams?
- Wouldn't life be grand if I didn't give you tests...? .




---

---

---

---

---

---

---

---

30

### WHAT IF...



I invented a mind reader that told me your TRUE KNOWLEDGE about any given subject (testing, psychology, sports, your significant other...)

- No more exams! Just a wave of the Grieve True Knowledge Tester 8000...
- After each lecture, I could just zap you all and write down your true knowledge score to determine your grade
- True scores would vary a bit because some of us know more than others about psychology, testing, etc.



31

---

---

---

---

---

---

---

---

### THE END.



32

---

---

---

---

---

---

---

---

### HERE'S THE PROBLEM...



There is no Grieve True Knowledge Tester 8000

There's no way to know your **true** knowledge of a subject

My test is only an **estimate** of your true knowledge about psychological testing – there could be **error**

In psychology, this is why we call scores on a test the "observed" or "obtained" scores rather than true scores

Hopefully, my test is closely related to your true score

But, some tests measure true scores better than others

We've all had tests where we thought we would do better (or worse) than we did – the observed score was different than what we thought the true score should have been

33

---

---

---

---


---

---

---

---

## ERROR



What are all the reasons a test taker's observed score might differ from his/her "true" score?

- The test itself
- The test taker
- The testing situation/environment

Error is no one's fault – it's an inherent part of testing

True scores vary because of true variability (people differ) and error

Core concept of **Classical Test Theory**

34

---

---

---

---


---

---

---

---

## CLASSICAL TEST THEORY



$X = T + E$  or **Observed score = True score + Error**

Assumption: measurement errors are **random**

- Systematic errors do happen, but they're less likely to make investigators come to the wrong conclusion
- If we made the same mistake over and over (systematically), then eventually we would learn to compensate
- E.g., rubber yardstick vs yardstick that is always 2" too long
- Because error is random, if we tested you on the same test 1,000 times, we would get...

35

---

---

---

---

---

---

---

---


## CLASSICAL TEST THEORY (CONT'D)

A normal distribution - the **mean** should be your **true score**

So if we assess you enough times, we can estimate your true score (i.e., the mean)

Assumption: true score does **NOT** change with repeated testing, variability in scores is **only** because of random error

If we found the standard deviation, we'd get the **standard error of measurement (SEM)**



36

---

---

---

---

---

---

---

---

## STANDARD ERROR OF MEASUREMENT (SEM)

How classical test theory measures error

Standard deviation of errors around a true score

Tells us, on average, how much a score varies from the true score

We can estimate the SEM using the standard deviation of the observed score and the **reliability** of the test

$$S_m = S\sqrt{1-r}$$

where  $S_m$  = the standard error for the measurement  
 $S$  = the standard deviation of scores  
 $r$  = the reliability coefficient

37

---

---

---

---

---

---

---

---

---

---

## STANDARD ERROR OF MEASUREMENT

Why is SEM important?

Where do we find SEM?

38

---

---

---

---

---

---

---

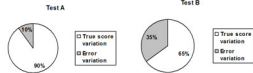
---

---

---

## RELIABILITY

Asks **what proportion** of variation in observed test scores can be attributed to true score variation vs. error variation?



39

---

---

---

---

---

---

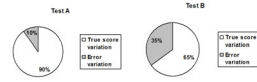
---

---

---

---

## RELIABILITY (CONT'D)



Test A was more **reliable** because it is better at measuring true score variability

Reliability coefficient for Test A is 90/100 or .90 – 10% of observed score variation can be attributed to error variation

What was Test B's reliability coefficient?

If a test was **perfectly reliable**, what would its reliability coefficient be?

If it was **completely unreliable**, what would its coefficient be?

---

---

---

---

---

---

---

---

---

---

40

## RELIABILITY (CONT'D)

Most reliability coefficients are **correlations**

Reliability is also a **ratio**:

- Variance of true scores
- Variance of observed scores
- Remember, observed variance = true score variance + error
- So, reliability will always be < 1

$$\text{Reliability} = \frac{\sigma_{\text{true}}^2}{\sigma_{\text{total}}^2}$$

Reliability can also be thought of as a **percentage** – the percentage of observed variance that is attributable to true score variance (the rest is error)

.40 reliability correlation coefficient = 40% of the variation among examinees can be explained by **real** individual differences; 60% is random or chance factors

---

---

---

---

---

---

---

---

---

---

41

## DOMAIN SAMPLING MODEL

Central concept in classical test theory

We have error problems when we use a **small number** of items to test a larger, more complicated construct or domain (**item sampling error**)

We cannot assess every single possible item in a domain or psychological construct (e.g., spelling) – we have to choose a **sample** of behavior and extrapolate

- We could choose an **infinite** variety of items to test the same thing

Reliability analysis helps us estimate how much error we make when we use your score from a shorter test to estimate your true score

Reliability = **ratio of observed score variance** on a shorter test/sample of behavior **over the full domain true score variance**

---

---

---

---

---

---

---

---

---

---

42

## MODELS OF TEST RELIABILITY

Three major methods of estimating a test's reliability

- Test-retest method
- Parallel forms method
- Internal consistency method

43

---

---

---

---

---

---

---

---

## TEST-RETEST RELIABILITY

Estimates the amount of error associated with administering a test at **two different times (time sampling error)**

How to calculate

- Administer the same test twice, correlate the scores

Assumes that the tested construct is relatively **stable**

- What are some psychological constructs we think might be stable and therefore good for analyzing test-retest reliability?
- What are some constructs that would **not** be good candidates for test-retest reliability?

44

---

---

---

---

---

---

---

---

## PROBLEM WITH TEST-RETEST RELIABILITY — CARRYOVER EFFECTS

Occurs when 1<sup>st</sup> testing session influences scores on 2<sup>nd</sup>

- E.g., remembering their first answers

Could result in an **overestimate** of the true reliability

This is a problem if changes over time are **random** (not predictable, affects some but not all)

Not a problem if **systematic** (everyone's score changes the same from time 1 to time 2)

**Practice effects**

Time interval between tests is vital

- Too short = carryover effects; too long = other potential sources of error

45

---

---

---

---

---

---

---

---

## PARALLEL FORMS RELIABILITY (AKA EQUIVALENT FORMS, ALTERNATE FORMS)

Helps ensure that your test scores are not flukes associated with the **particular set of items** you're using

- Remember domain sampling model? This helps us deal with **item sampling error**

Determines the error variance attributable to the selection of one particular set of items

Compares **two equivalent forms** of a test that measure the **same attribute**

Use different items, but the rules used to select items of a particular difficulty level are the same

Pearson product moment correlation coefficient

One of the most **rigorous** assessments of reliability

46

---

---

---

---

---

---

---

---

## PROBLEMS WITH PARALLEL FORMS RELIABILITY

A lot of **extra work!**

Can be difficult to test the same group of people on the same day

Potential for practice effects

- Can address this by **counterbalancing** administrations



47

---

---

---

---

---

---

---

---

## INTERNAL CONSISTENCY

Evaluate by **dividing a test** into subcomponents

Easier than creating two separate test forms

Evaluate the extent to which different items on the test measure the **same ability or trait**

- Could be lower if we have some **bad items**
- If we're measuring many traits, internal consistency will be low

Different ways to measure

- Split-Half, Cronbach's Alpha, and Kuder-Richardson<sub>20</sub>

48

---

---

---

---

---

---

---

---



## SPLIT-HALF RELIABILITY

How to calculate:

- Give a test, then split it in half, score each half separately, then correlate the two scores

No practice effects!

---

---

---

---

---

---

---

---

49

## PROBLEMS WITH SPLIT-HALF RELIABILITY

Why might split-half not be the best estimate for the addition test's reliability?

- Test gets progressively more difficult, so comparing 1-6 vs 7-12 may underestimate true reliability
- We've shortened the test (6 items vs 12 items) – underestimate the reliability of the full test

Correlation between the two halves of the addition test = .69

- How can we estimate what the reliability would be for the entire test?
- Spearman-Brown Formula

---

---

---

---

---

---

---

---

50

## SPEARMAN-BROWN FORMULA

Estimates reliability for the full test when you split it in half

$$reliability = \frac{2 \times r_{half-test}}{1 + r_{half-test}}$$

Corrected  $r$  for the full test =  $2r_{HALF} / (1 + r_{HALF})$

- =  $2 (.69) / 1.69$
- = .817

**Warning:** Don't use when two halves of the test have different variances

- Use Cronbach's Coefficient Alpha instead

---

---

---

---

---

---

---

---

51

## INTERNAL CONSISTENCY - CRONBACH'S COEFFICIENT ALPHA

More conservative estimate of reliability for split-half method

- Will give you the **lowest** estimate of reliability
- Can tell you if reliability is good but may not tell you much if it's low (true reliability may still be high)
- Do items "hang together"?

$$\alpha = 2 \left[ \frac{\text{Whole test variance} - (\text{first } \frac{1}{2} \text{ variance} \times \text{second half variance})}{\text{whole test variance}} \right]$$

Problems with alpha

- Skewness – responses to items aren't normally distributed

**Most commonly used reliability index**

52

---

---

---

---

---

---

---

---

**VALIDITY** | All we want is to be validated!

53

---

---

---

---

---

---

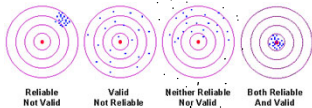
---

---

## RELIABLE AND VALID IS WHAT WE NEED!

We need both if we're going to use psych tests to make important Decisions

We need evidence that a test is meaningful in certain situations for particular purposes



54

---

---

---

---

---

---

---

---

## VALIDITY ACCORDING TO THE JOINT COMMITTEE

- 1985: American Educational research association (AERA), American psychological association (APA), and the National council on measurement in education (NCME)
- Standards for educational and psychological testing – last revised in 2012
- Validity = evidence for inferences made about a test score
- Validity is a unitary concept – represents all the evidence that supports the intended interpretation of a measure
- 3 types of evidence:
  - Content-related
  - Criterion-related
  - Construct-related

55

---

---

---

---

---

---

---

---

## A WORD ON FACE VALIDITY

Joint committee does not recognize this as a legitimate category of validity

The mere appearance that a measure has validity  
Items seem to reasonably relate to the test's purpose

• E.g., math test vs thermometer

We need more evidence than "looks good to me"  
Need systematic, scientific evidence that our measure's scores are related to a certain outcome of interest

Face validity can be important for buy-in and motivation of examinees



56

---

---

---

---

---

---

---

---

## CONTENT-RELATED VALIDITY EVIDENCE

Does the test adequately represent the conceptual domain it's supposed to cover?

Important for educational testing

- Hopefully your first test in here covered the content adequately (not just history)
- Other factors can affect your score, though...
  - Reading level, how the items are worded

Content validity no longer thought of as separate from other types of validity evidence

Only type of evidence (along with face validity) that is logical rather than statistical

Usually based on expert opinion

Can also utilize factor analysis

57

---

---

---

---

---

---

---

---

## CONTENT-RELATED VALIDITY EVIDENCE (CONT'D)

Two related concepts mentioned in the standards...

- **Construct underrepresentation**
  - Failure to capture important components of a construct
  - E.g., only testing memory related information for a psych subject test
- **Construct-irrelevant variance**
  - Scores are influenced by factors irrelevant to the construct
  - E.g., test anxiety, reading comprehension or fluency, illness, poor testing environment

58

---

---

---

---

---

---

---

---

## CRITERION-RELATED VALIDITY EVIDENCE

How well does our test (i.e., predictor) correspond with a particular criterion or outcome?

- Criterion: standard against which a test is compared
- To work, the criterion has to be the "gold standard" measure of a phenomenon
- Often, we learn more about the phenomenon (and how to define it better) by trying to create a good predictor
- Job performance/satisfaction, anxiety/depression, academic achievement, prison recidivism
- A predictor is only useful if it's cheaper/easier to measure than the criterion itself

Does the test score correlate with a well-defined criterion measure?

- If so, we'd say the test has evidence of criterion-related validity

We're using a test as a substitute in some cases to take the place of some criterion that we can't/won't measure

- Two types - Predictive validity evidence and concurrent validity evidence

59

---

---

---

---

---

---

---

---

## CRITERION-RELATED VALIDITY EVIDENCE (CONT'D)

### Predictive validity evidence

- Can the test successfully predict an outcome or criterion of interest?
  - Prison recidivism
- E.g., GRE reportedly predicts educational outcomes in grad school (but not as well as high school GPA...)
- Correlation between sat and first year GPA in college:  $r = 0.36$

### Concurrent validity evidence

- Can the test stand in for a criterion/outcome right now?
- Assess simultaneous relationship between the test and the criterion
  - E.g., job samples (best for blue collar jobs and trade skills), testing for learning disabilities to establish IEPs

60

---

---

---

---

---

---

---

---

## CRITERION VALIDITY EXERCISE

For each of the following tests, come up with

- a criterion or outcome against which the test scores (predictor scores) could be compared in order to establish criterion-related validity evidence
- How could we come up with a study to test whether our predictor relates to our criterion?
- Would this study yield concurrent or predictive validity evidence?

Tests:

- suicide risk scale
- driver skill test
- sales ability indicator
- inventory of substance abuse risk

61

---

---

---

---

---

---

---

---

---

---

## HOW TO EXPRESS THE RELATIONSHIP BETWEEN A TEST AND A CRITERION

### Validity coefficient

- Tells the extent to which a test is valid for making statements about the criterion
- How large does the coefficient have to be?
  - Good question! Rarely do you see a value larger than .60
  - .30 - .40 = high
  - There are soooo many sources of error and other variance in our outcome/criterion variable
  - Think about all the variables that could impact educational outcomes...
  - We can't expect our one predictor (or test) to explain all the variance in educational outcomes... but hopefully it will help us predict better than by chance alone
  - Statistical significance = chances of obtaining that value by chance are low (5 out of 100 times)
  - If a validity coefficient is significant, then we'd say that the predictor helps us predict our criterion better than how we'd do predicting our criterion outcome by chance alone

62

---

---

---

---

---

---

---

---

---

---

## VALIDITY COEFFICIENT (CONT'D)

We should ask ourselves "How much of the variation in our criterion or outcome variable can we predict using our predictor/test score?"

- Answer: square the validity coefficient
- Validity Coefficient<sup>2</sup> = percentage of variation in the criterion we can expect to predict based on our knowledge of the test/predictor score
- So, WKU GPA, then we could explain (.40 X .40 = .16) 16% of variation in college performance (as measured by GPA) using just the sat
  - This means 84% of the total variation in college performance is still left unexplained!



63

---

---

---

---

---

---

---

---

---

---

## VALIDITY COEFFICIENT (CONT'D)

16% may seem like small potatoes, but every little bit helps

- Dunnette (1967) – showed how a simple questionnaire for military screening could save taxpayers millions of dollars every month even though validity wasn't that great

But Sometimes, small validity coefficients aren't useful

You have to ask yourself if it's worth the extra effort, cost to administer this test. etc.

- Are we predicting enough variance to make it worth our while?

Trend in psychology – prove what you're doing is useful (Utility!)

We need to know how to evaluate a validity coefficient

- Because 0.40 doesn't tell us much by itself (could be great or not so great)



64

---

---

---

---

---

---

---

---

## HOW TO EVALUATE A VALIDITY COEFFICIENT

From the joint committee's standards for educational and psychological testing

### 1. Look for changes in the cause of relationships

- The settings in which you use a test (e.g., the GRE) may be different from the experimental settings in which the test was initially validated

The dynamic nature of validity

- just because we find validity in one setting doesn't mean the test has validity in another setting

• Criterion validation logic assumes that whatever relationships we find between the predictor/test and the criterion/outcome during initial experiments will still be there when we're using the test in the real world

- This is usually true, but there may be situations where the relationship changes

- Did the sample population change? Are we comparing apples and oranges?



65

---

---

---

---

---

---

---

---

## HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

### 2. Ask "what does the criterion mean?"

- Is the criterion reliable and valid?

• If we're measuring the correlation of our test with another test with unknown validity, then it doesn't tell us much

• Meaningless garbage + meaningless garbage = a pile of meaningless garbage

- We need to make sure our criterion relates specifically to the use of our test

- SAT → college performance



66

---

---

---

---

---

---

---

---

### HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

**3. Review the subject population in the validity study**

- What if the population in the validity study is different from the population with whom we'll be applying the test?
- Study/real world attrition (Drop out rates)?
  - Employees who quit their job
  - Students who leave college
- Racial differences (think of the first intelligence tests)



67

---

---

---

---

---

---

---

---

---

---

### HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

**4. Be sure Sample Size was adequate**

- Common practice to do an initial validity study with a small, convenience sample
- Can be misleading because chance/error can have a larger impact on your results when you have a small sample
- Small samples can artificially inflate validity coefficients
- Good studies attempt to provide evidence of cross validation
- Cross validation study: assesses how well the test predicts performance on the criterion with a totally new group of subjects
- Larger initial samples increase the likelihood for cross validation



68

---

---

---

---

---

---

---

---

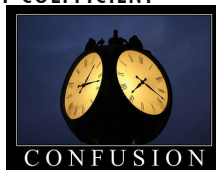
---

---

### HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

**5. Never confuse the criterion with the predictor**

- Many times, people are required to meet a certain level of performance on a predictor in order to be allowed to reach the criterion
- E.g., certain GRE score before admittance to a graduate program
- Thus, they are making the predictor a criterion
- Problem: predictor isn't perfect
- Some people do poorly on the predictor but succeed at the criterion



69

---

---

---

---

---

---

---

---

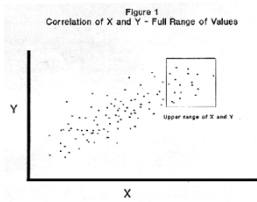
---

---

## HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

### 6. Check for restricted range on both the predictor and the criterion

- Restricted range = all scores on a variable fall close together
- Problem: correlation depends on variability
- Typical grad school GPA: 3.5 – 4.0
- If there's no variability on either your predictor or criterion, then your correlation coefficient won't mean much



70

---

---

---

---

---

---

---

---

## TIME TO FEED THE FLAMES OF HATRED...

The GRE is not a perfect predictor of grad school success

Possible explanations:

1. GRE may not be a valid test for grad student selection
2. Students who are admitted have such a restricted range of ability that it's impossible to find significant correlations
  - People with low GRE scores aren't admitted, so we can't include them in validity studies
3. Grad school grades have a restricted range, too
  - A's and B's only...  $\epsilon$  = failing grade

A better predictor? Undergrad GPA



71

---

---

---

---

---

---

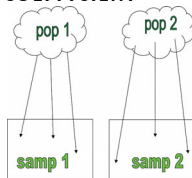
---

---

## HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

### 7. Review evidence for validity generalization

- **Generalizability:** evidence that findings obtained in one setting generalize or can be applied to another setting
  - This is an empirical question! We need proof!
- **Reasons evidence may not generalize between settings**
  1. Differences in how the predictor construct was measured
  2. Differences in how the criterion is defined or measured
  3. Differences in when the test was administered
  4. Different sample demographics
  5. Differences between the setting in which the test was developed and the "real world"



72

---

---

---

---

---

---

---

---



## HOW TO EVALUATE A VALIDITY COEFFICIENT (CONT'D)

### 8. Consider differential prediction

- Predictive relationships may change for different groups
- gender, race, age, national origin, primary language
- Separate validity studies for different groups may be needed



73

---

---

---

---

---

---

---

---

## CONSTRUCT-RELATED VALIDITY EVIDENCE

Relatively new (mid-1950s) – now seen as the unifying, most important type of validity  
 Establishing criterion-related validity can be tough when we can't objectively or clearly define the criterion (e.g., intelligence, love, curiosity, mental health)

Construct-related: researcher tries to define a construct and develop instruments to measure it at the same time

Over a series of studies, You try to gather evidence that your measure means something by showing how it relates to other measures

Two types: Convergent and Discriminant evidence

74

---

---

---

---

---

---

---

---

## CONSTRUCT-RELATED VALIDITY EVIDENCE (CONT'D)

### Convergent Evidence

- Measure correlates well with other tests believed to measure the same construct
- Sort of like criterion-related evidence but there's no clear criterion
- Measures of the same construct converge (note: they do not perfectly overlap... otherwise, why have a new measure?)

### Obtained in two ways:

1. Show that the test measures the same things as other tests used for the same purpose
  - Correlate IQ tests with other IQ tests
2. Demonstrate relationships we would expect if the test is really doing its job
  - If we're really measuring our construct correctly, what relationships would we expect to see between our test scores and other measures?
    - Intelligence – academic achievement (math, reading, writing...), job performance, life expectancy, fMRI studies, marital satisfaction, interpersonal functioning, emotional regulation...



75

---

---

---

---

---

---

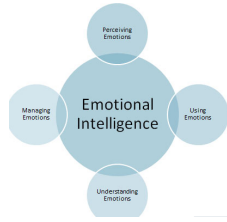
---

---

## CONSTRUCT-RELATED VALIDITY EVIDENCE (CONT'D)

### Discriminant evidence (or divergent validation)

- Demonstration that the measure is unique and has low correlations with measures of unrelated constructs
- Evidence that the test doesn't measure what it's not supposed to measure
- If we're measuring the exact same thing with two measures, we really only need one
- Our new measure has to relate to the old measure but be different and unique at the same time




---

---

---

---

---

---

---

---

76

## HOW RELIABILITY AFFECTS VALIDITY

We can't validate an unreliable test  $r_{xy} \leq \sqrt{(r_{xx})(r_{yy})}$   
 The maximum validity coefficient can be sunk by one of the variables having poor reliability  
 We can have reliability without validity (Thornberry artistic talent tester 3000)  
 We can't have validity without reliability

---

---

---

---

---

---

---

---

77