



## Agenda

- Validity generalization.
  - Why and how
- Introducing construct relationships.
  - Convergent and discriminant evidence.

# Validity Generalization

- Validity is about the **inferences** we make from a test, which means that validity depends on the purpose we are using a test for.
  - Does this mean we have to re-validate a test every time it is used in a new situation?
  - Or can validity evidence obtained in one situation be generalized to another?
- **Validity generalization** is the name we give to the process we use to answer this questions.
  - Focused on **criterion-related validity** evidence.

# Some Historical Perspective

- 1960s-1970s: several IO psychologists conducted reviews of validation studies of selection tests.
- They found LOTS of variability in validity coefficients...
  - Even for the **same** test predicting the **same** job!
- Came to the conclusion that validity was **situationally specific**.
  - You could **not** assume that a test that was a good predictor in one situation would be good in another.
  - New validation study needed for every organization, every implementation of a test.
- Obvious downsides: burdensome and expensive!

## Schmidt & Hunter (1977)

- o Pointed out that many of the validity studies in those early reviews weren't very good.
  - o Small samples = low power
  - o Unreliable measures
  - o Range restriction
  - o Plenty of other flaws!
- o Maybe the variability in validity coefficients is due to variations in study quality, **not** to true differences in validity!
  - o Type II error – failing to find an effect that really is there.
  - o In other words, maybe situational specificity is a methodological artifact.


## Typical Validation Conditions Work Against Validation

- o “when true validity for a given test is in fact constant at .45 in a series of jobs, criterion reliability is .70, the prior selection ratio on the test is .60, and sample size is 68 (the median over 406 published validity studies; Lent, Aurbach, and Levin, 1971), the test will be reported to be valid 54% of the time and invalid 46% of the time.”
  - o - Schmidt & Hunter (1977), p. 530.
- o Using significance tests to label a test as “valid” or “invalid” doesn't help matters!

## Logic of VG

- When we have a number of studies using the same test to predict (pretty much) the same criterion, we can estimate how much of the variance in results is due to measurement error.
  - Loosely analogous to estimating error in one test via its correlation with another test.
- Based in Bayesian probability:
  - Bayes' theorem:  $P(A|B) = \frac{P(B|A) P(A)}{P(B)}$
  - Rewrites the probability of A given B into the probability of B given A and the *prior* probability of A (the probability of A by itself, without any information about B).

## Applying Bayes to Validity

- In other words... what I want to know is the **probability that this test is valid (A)** given the **result of my study (B)**. Bayes' theorem says I can answer this based on:
  - The probability of my result (B) given that this test is valid (A).
    - In other words, the quality of my study – if the test is valid, could I tell? Sample size and reliability influence this.
  - The probability of my result (B) overall.
  - The probability that the test is valid at all (A). 
    - We use information from previous (*prior*) studies about the test to gain information about A.

## More Practically

- We can use meta-analytic techniques to:
  - Estimate the overall probability that a test is valid –  $P(A)$ .
  - Estimate the impact of methodological error on our ability to find a meaningful result -  $P(B|A)$ .
- Estimate and correct for:
  - Variations in sample sizes across studies.
  - Unreliability in the criterion.
    - Why just in the criterion?
  - Range restriction
  - Assumed a normal distribution of errors across studies for the last two (since these tend to go unreported).

## Schmidt & Hunter Concluded:

- For some predictor/criterion combinations, we can explain **nearly all of the variance** in validity coefficients across studies by variations in **study quality!**
  - In other words, true validity is pretty much constant.
- For other combinations, there is meaningful variability across studies left after we control for artifacts.
  - This could be due to artifacts we haven't controlled for yet.
  - Or it could mean that validity is definitely not guaranteed.
- We can put confidence intervals around our “true” (corrected) mean validity coefficient.
  - If these don't include zero, the test is valid to at least **some** extent across studies.

## Issues in VG

- Requires a substantial set of studies linking the same (or similar) predictor to the same (or similar) criteria.
- Can only generalize to situations “similar” to those contained in the original set of studies.
  - e.g., if all previous studies were in customer service jobs, you can’t be confident that it will generalize to mechanical engineers.
- Requires some assumptions in order to make corrections.
  - If you don’t make those assumptions, you can’t account for those types of error.

## VG Today

- Widely used!
  - Sometimes (but not always) acceptable in a court defense in lieu of a validity study.
- Can be broad (e.g., cognitive ability tests) or narrow (e.g., the GRE).
- Some use all of these corrections, some don’t.
- Differences in how “similar” predictors and criteria are interpreted.


# Relationships with Other Constructs

## Construct Relationships

- Earlier, we defined this as:
  - “Does the test as a whole relate to other constructs in the way we theoretically expect it to?”
- Concerned broadly with the construct’s **nomological net**.
  - Criterion relationships are **part** of the nomological net.
- Testing predictions based on theory and prior empirical evidence.



## Convergent Evidence

- If our test really measures X, it should be **highly** related to other tests **that also measure X**.
- Not perfectly related, but highly related. (Why?) 
- If our test really measures X, it should be **reasonably** related to tests that measure **things that are similar to X**.

## Discriminant Evidence

- (sometimes called “divergent evidence”)
- If our test measures X, it should be **(essentially) uncorrelated** with things that have **nothing to do with X**.
  - But most things in psychology are somewhat correlated, so we’ll settle for a small correlation here.
  - Why do we need to test this? What alternative explanations are we trying to rule out here?
- If our test measures X, it should **not be highly** correlated with **constructs that are distinct from X**.

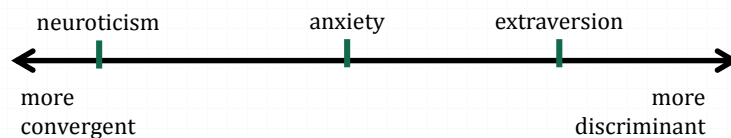


## For Example

- o If we develop a test of neuroticism, it should be:
  - o Highly correlated with other measures of neuroticism.
  - o Moderately correlated with measures of anxiety.
  - o Barely correlated at all with measures of extraversion.
  - o Not too highly correlated with measures of anxiety.
- o That's right: whether a particular relationship is evidence of convergent or discriminant validity may be ambiguous.
  - o This is not really a binary concept!

## Convergence Continuum

- o I find it more helpful to think about convergent and discriminant validity as a continuum:



- o Hypotheses about the **relative** strength of relationships among constructs.

# Questions?

For next time: Construct relationships, and the multitrait-multimethod matrix.

Read: R & M 8.5 & 8.9

Reading Response: Under what circumstances is a *small* correlation between your test and another variable evidence *for* validity?

Lab Friday: Classical Item Analysis