Ch. 5: Validity

Copyright © 2020 Michael Diehr
All Rights Reserved
For use only by students enrolled
in my sections of Psyc 402
through December 2020.
May not be posted, shared or uploaded
online without permission.

Reminders

- Review Chapter 4 - Reliability
- Begin Chapter 5 - Validity
- Introduction to Project assignment
- Discuss Exercise 3 - Literature Search
  - See website for due date

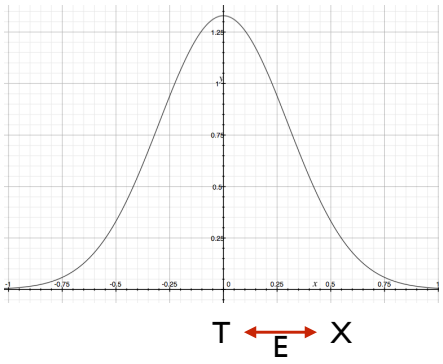
Review : Reliability

- History
- Classical Test Score Theory
  - Models of reliability
  - Sources of error
- Estimating Reliability
- Increasing Reliability
- SEM and Confidence Intervals

Classical Test-Score Theory

- T= True Score
- X = Observed
- E = Error

- X = T+E
- E = X-T



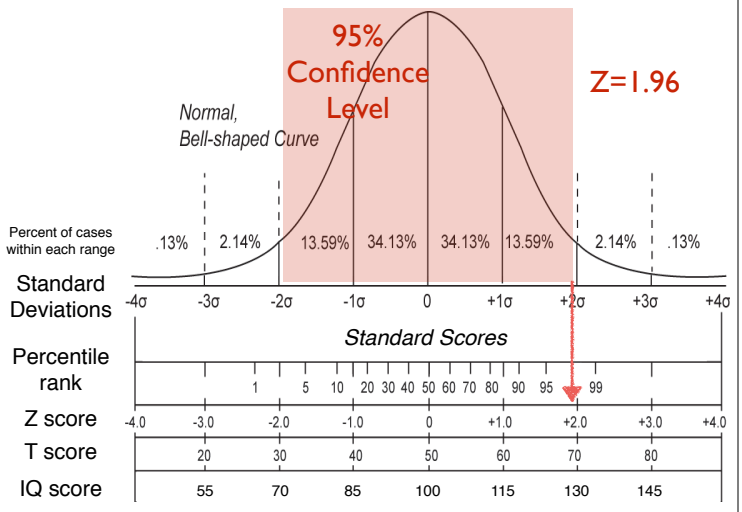
Reliability: errors & methods

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach's Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

Standard Error of Measurement

- SEM = $S\sqrt{1-r}$
- S = std dev of measured scores
- r = reliability coefficient of test

- SEM : 68% confidence interval (1 standard deviation, e.g. Z=± 1.0



Confidence Interval

- "How likely is a true score to fall within a range"
- Z = z-score associated with % range
  - **Confidence Level**
- **Confidence Interval** = Z * SEM
- Example:
  - 95% CI : Z = 1.96
  - SEM = 4.64
  - 1.96 * 4.64 = 9.1
  - 95% CI = ± 9.1 points
  - True Score Range = X ± CI
    - 106 ± 9.1 = range from 96.9 … 115.1

Increasing Reliability

- **Increase N** (number of questions, items or tests)
- **Focus** on common characteristic
  - tests are more reliable if all items measure a single characteristic
- Use **Item Analysis** ("discriminability analysis") to find items that best measure a single characteristic

Reliability Summary

- Measurement Error occurs in all fields -- Psychology focuses on it
- Kind of Reliability : *where* the error came from
- Improving Reliability: more items, focusing test, factor analysis
- Reliability is useful: calculate SEM and Confidence Intervals
- Reliability is not Validity: Unreliable tests aren't valid
- A reliable test *may* be valid

Validity

- History
  - Griggs v. Duke Power
  - Ricci v. DeStefano
- Defining Validity
- Four Kinds of Validity
- Reliability vs. Validity
- Validity as Explained Variance

Griggs v. Duke Power (1971)

- Duke Power company - Coal fired power plant
- 13 people employed as Janitors
- Wanted promotions to next level (coal handler)
- Duke required certain score on IQ test for promotion
- Of 95 employees at power station, 14 were Black, 13 of 14 were janitors
- Legal case: was the IQ test requirement valid or discriminatory?
- Supreme Court decision : "invalid"

Griggs v. Duke Power - 2

- Supreme court found

  If a test impacts different ethnic groups disparately, the business must demonstrate the test is a "reasonable measure of job performance"
- In scientific terms: tests must be *valid* predictors of specific criteria.

Disparate Impact

- aka "Adverse Impact"…
  - disproportional…
  - adverse effect…
  - on a protected class…

Impact

- A negative Effect on
  - Employment or
  - Housing

Disproportionate

- The proportion of a protected class affected by the behavior is different from a non-protected class
  - 80% rule
- Example
  - 50% of men are hired
  - 45% of women are hired
  - 45/50 = 90%

Protected Classes

- Federal Law:
  - race, color, religion, national origin, age (40+), sex, pregnancy, citizenship, familial status (kids), veterans, genetic status
- California Law:
  - …disability, sexual orientation, HIV or medical status, political party, victim of domestic violence

Definitions of Validity

- Agreement between test scores and the thing (construct) it claims to measure
- Many other definitions; some confusing or incompatible with each other
- AREA/NCME (1985, 1999, 2012) "Standards for Educational and Psychological Testing"
- One informal definition: Face Validity
- Three formal definitions: Content, Criterion, Construct

Face Validity

- Common Sense / Informal Analysis
- "I like mechanics magazines" = you like mechanics magazines.
  "I never tell a lie" = you never lie
- Question -- what factors might influence a test-taker's response?
- Face validity is not a proper type of validity
- Quizzes in magazines or on the Internet -- appear "face valid" but usually have low reliability and very low validity

Does Face Validity Matter?

- Naive view = face validity
- Tests with very little face validity...
  - what does the average test taker feel about the test?
  - motivation?
  - confusion?

Content Validity

- Do the test questions cover the construct?
- Commonly used in education (achievement/ aptitude testing)
- Example: does an Algebra test contain questions about Algebra?
- Logical, rather than statistical argument
- Fuzzy definition
- Modern theories consider Content Validity a sub-set of other types of validity

Content Validity 2

- If a test is supposed to test a specific *Construct*, problems may arise:
- Construct underrepresentation
  - test misses important information
- Construct-irrelevant variance
  - scores are influenced by outside factors
    - e.g. anxiety, reading comprehension, IQ, etc.

Content Validity

Improving on Content Validity

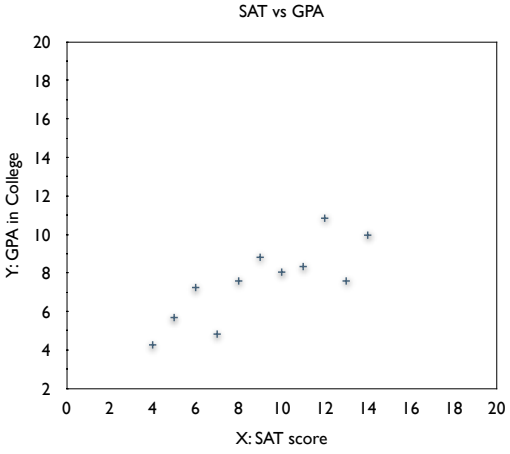
Norm vs. Criterion

- Norm-referenced test
  - "Hire the top 5% of applicants"
  - Pro - select the best
  - Con - can they do the job?
- Criterion-referenced test
  - "Hire those who can do _____"
  - Pro - they can do the job
  - Con - not selecting the best

Criterion Validity

- Criterion -- a well defined measure of performance *in the real world*
- Criterion validity -- the correlation between a test score and the specific criterion

Criterion Validity Example



- SAT vs GPA
- Validity = $R_{x,c}$
- Predictive

Predictive vs. Concurrent

- Predictive
  High School SAT score (predictor) predicts later College GPA (criterion)
- Concurrent
  Work samples from mechanics

Measuring Validity

- General: relationship between test and what it's supposed to be measuring
- Specific: Pearson product-moment correlation (r) between Test Score (X) and criterion (Y)

How much Validity?

- Validity is often lower than you'd want
- r > 0.60 is rare
- r = 0.40 or lower is common

- r^2 = % of variance explained
  - r = .60 means just 36% of variation in the criterion scores explained by the predictor score (means 64% is not explained)
  - r = .40 means 16% of variance explained (84% not)

Evaluating Validity Coefficients

- Validity depends on the cause of the correlations
- And on the meaning of the criterion
- If the reason for the relationship changes, validity may change as well.

- Examples:
  - Different *subjects*
  - Different *situation*
  - Different *criterion*

Problems with Criterion Validity

- "what is the criteria?"
- Some criteria are obvious and sensible: E.g. using SAT to predict college GPA

- For many constructs, it's not clear: E.g. intelligence. IQ test used to predict ___?

- Without a *specific* criterion, you can't calculate criterion validity!

Construct Validity

- Construct = theoretical entity. In soft sciences, usually not directly observable or measurable.
- Solution -- the world is complicated. In Psychology (as in other sciences) things can exist even if they aren't easy to measure.
- Method -- collect evidence for the construct via multiple methods, multiple sources, multiple subjects

Construct Evidence

- **Convergent Evidence** — data from multiple sources all tend to point to the same conclusion.

- **Divergent Evidence** (aka **Discriminant** Evidence) — evidence that a Construct is not the same as another

Construct Validity Example

- Construct: Insomnia
- Convergent correlations:
  - # of hours slept
  - tired feeling
  - another test of insomnia
  - R should be strong (either positive or negative)
- Divergent correlations:
  - age
  - IQ
  - R should be weak (close to zero)

Construct Validity Example

	An old test of insomnia	# of hours slept	Feeling Tired	IQ	Age
Correlation between your test of insomnia and…	✓	✓	✓	none	none
Which kind of Construct validity?	Convergent	Convergent	Convergent	Divergent	Divergent

The Love Test

- Rubin (1970)'s Love Scale
- From Literature, created 198 items on Likert scale
- Result: a "Love" scale and a "Liking" scale
- Love scale: attachment, caring, intimacy
- Convergent evidence:
  - lovers vs. friends
  - eye contact
- Divergent evidence:
  - possible to love someone w/o liking them

All Validity is Construct Validity?

- Modern theory: only one type of validity -- Construct validity
- Other types of validity are just sub-types of Construct validity.

Ricci v. DeStefano (2009)

- Eighteen firefighters (17 white, 2 Hispanic) in New Haven, CT filed suit against the city
- Background:
  - All scored above the cutoff on a test for promotion to management
  - None of the African Americans had scored above the cutoff (though they passed)
  - City vacated the test results, fearing lawsuit -- promotions were denied -- nobody was promoted
- Question: was this the right action to take?

Ricci v. DeStefano 3

- Supreme court decision:
- Found City *did the wrong thing*
- Race-based action can be taken only if "demonstrate a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute"
- Tests are discriminatory:
  - if they are *not valid for the job*.
  - Not just because *protected classes* get different results.

4 Kinds of Validity

	Description	Notes	Statistic(s)
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific outcome?	requires a well-defined criterion	Pearson's R (correlation) between Test and Criterion
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson's R)

Reliability vs. Validity

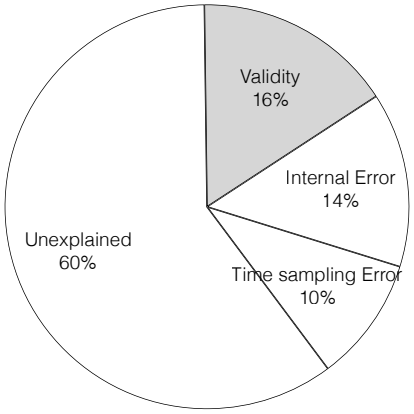
- Validity coefficient is the correlation between a test and the criterion
- We know that *Test Measurements* and *Criterion Measurements* are unreliable
- The maximum validity is the square root of the product of their individual reliabilities. $r_{12max} = sqrt(r_{11}r_{22})$
- If the tests have low reliability, validity can be hidden

Reliability vs. Validity : Example

Reliability of Test	Reliability of Criterion	
CES-D	**DSM-V Depression**	**Maximum Validity (r)**
1	1	1
0.8	1	0.89
0.6	1	0.77
0.4	1	0.63
0.2	1	0.45
1	0.5	0.71
0.8	0.5	0.63
0.6	0.5	0.55
0.4	**0.5**	**0.45**
0.2	0.5	0.32

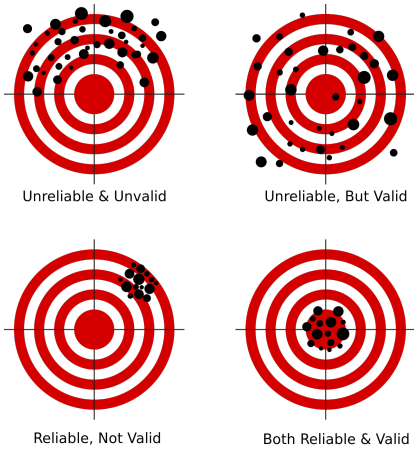
Variance: Reliability & Validity

- Variance in test scores can be divided into different portions
- In this example, only 16% is useful (validly predicts criterion)
- Other sources of error are known or unknown



Reliability vs. Validity

- Shotgun vs. rifle analogy
- Similar to accuracy vs. precision



Reliability or Validity Part 1

Procedure	Reliability or Validity	What kind?
Correlation of IQ scores with ability to handle 30 tons of coal per day	Validity	Criterion
Correlation between SAT scores taken in Junior vs. Senior year of High School	Reliability	Test-Retest
Having a committee review a high school history exam to make sure the questions cover all required topics	Validity	Content
Correlation between scores on two different versions of a Final Exam	Reliability	Parallel Forms

Reliability or Validity Part 2

Procedure	Reliability or Validity	What kind?
Showing that your new measure of Anxiety correlates with an old measure of Anxiety, and does not correlate with IQ	Validity	Construct (Convergent, Divergent)
Showing that everyone in your research lab can rate "grumpy" expressions the same way	Reliability	Inter-Observer
Looking at a question about depression and deciding that it measures depression.	Validity	Face
Showing that questions on your test all correlate strongly with each other	Reliability	Internal Consistency

Reliability or Validity Part 1

Procedure	Reliability or Validity	What kind?
Correlation of IQ scores with ability to handle 30 tons of coal per day		
Correlation between SAT scores taken in Junior vs. Senior year of High School		
Having a committee review a high school history exam to make sure the questions cover all required topics		
Correlation between scores on two different versions of a Final Exam		

Reliability or Validity Part 2

Procedure	Reliability or Validity	What kind?
Showing that your new measure of Anxiety correlates with an old measure of Anxiety, and does not correlate with IQ		
Showing that everyone in your research lab can rate "grumpy" expressions the same way		
Looking at a question about depression and deciding that it measures depression.		
Showing that questions on your test all correlate strongly with each other		

Ch. 5 - Part 2

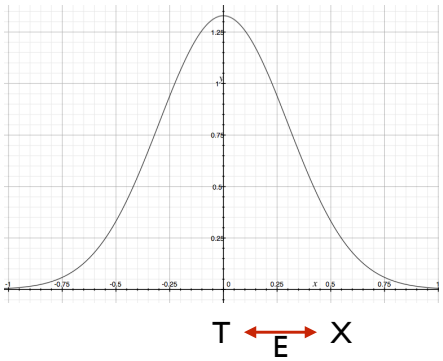
Review

- Reliability : easier to define and calculate. A property of the Test itself.
- Validity : harder to define, not inherent to the test, depends on the *way the test results are used*.

Classical Test-Score Theory

- T= True Score
- X = Observed
- E = Error

- X = T+E
- E = X-T



4 Kinds of Reliability

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach's Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

4 Kinds of Validity

	Description	Notes	Statistic(s)
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific outcome?	requires a well-defined criterion	Pearson's R (correlation) between Test and Criterion
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson's R)

Disparate Impact

- aka "Adverse Impact"…
  - disproportional…
  - adverse effect…
  - on a protected class…

How much validity do you need?

- Clinical Significance
  - Effect Size of Benefits
  - vs Cost and Harm
- Examples:
  - an inexpensive test which can detect 1% of cancers (validity r^2=1%)
  - a written test for airline pilots to detect who is unsafe (validity r^2=93%)
- Answer: it depends

Interactive Correlation Example

- http://rpsychologist.com/d3/correlation/

4 Kinds of Validity

	Description	Notes	Statistic(s)
Face	do items "look" valid?	informal, improper, non-scientific	none
Content	do test questions cover the topic?	logic & judgement - there are no stats to calculate	none
Criterion	does the test predict a specific outcome?	requires a well-defined criterion	Pearson's R (correlation) between Test and Criterion
Construct	does the test measure what it claims	modern theory: all validity is Construct validity	Convergent and Divergent correlations (Pearson's R)

Constructs, Factors, Facets

- Constructs may be "high" or "low" (also called "top" or "bottom")
- Top-level constructs are made of smaller constructs
- aka Factors, Facets, Dimensions, Domains…

Construct: Anxiety

Construct: Anxiety

- Cognitive
- Emotional
- Physiological

Convergent Validity

- Multiple factors within a construct
  or
  multiple measures of a construct
- All correlate with each other strongly

Divergent Validity

- Other factors (not part of a construct)
- Should have low to zero correlation
- What about negative correlations?

Psyc 402 Project - Overview

- Goals
- Proposal
- Data collection
- Analysis
- Paper

Psyc 402 Project Goals

- Literature Review
  - two kinds of articles
- Creating new test measures
  - writing two test questions
- Reliability of your measures
  - which kind?
- Validity of your measure
  - convergent
  - divergent (discriminant)
- APA writing

Project - Proposal

- Pick a sub-field
  - get 1 *Review* journal article
- Pick a construct to measure
  - get 1 *Original Research* article
- Design two single item survey questions to measure your construct
- Write Proposal: about 1000 words long, 2 questions, 2 articles
- **Due Date: see syllabus**
- Note: late submissions may not be included in survey

Psyc 402 - Survey

- Your 2 questions must be format that can be done on website (e.g. multiple choice, Likert, etc)
- All students will take the test answering all questions
- Anonymity
- Additional Data
  - Gender
  - Age, GPA
  - NEO-FFI-3 (5 factors: NEOAC)

Psyc 402 - Limitations

- Construct: can not choose
  - Anxiety
  - Any of the NEO-FFI-3 scores (Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness)
- Convergent & Divergent factors
  - Gender: not allowed
  - Age, GPA: OK
  - May use another student's factors

Psyc 402 Project - FAQ

- APA Format
- Length
  - Proposal : about 4 pages (1000 words)
  - Paper : about 8 pages (2000 words)
- How to find *Original Research* that includes a test? Use references or citation index.
- Survey Question Format:
  - Ordinal or Interval/Ratio Scale (numerical) answers : much easier to analyze
  - Avoid Nominal scales

Exercise 3: Literature Review

- Types of articles
- Searching in PsycInfo
- Other Sources

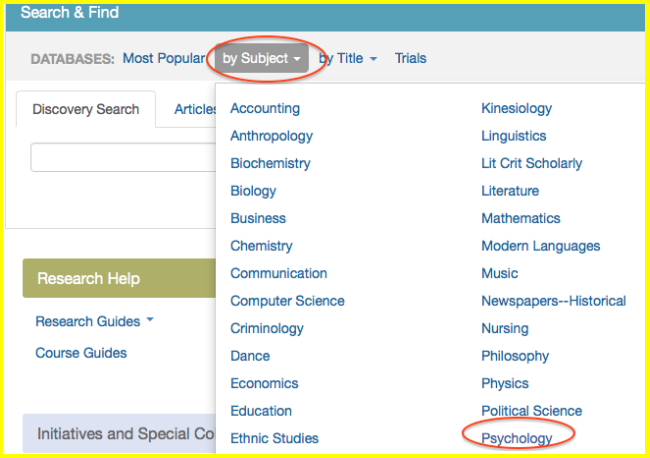
Types of Articles

- Original Research (aka "Empirical Study")
  - the most common type: the author(s) run an experiment, collect data, analyze it, and write up their conclusions
- Review (aka "Literature Review")
  - not to confused with a "Book Review" or "Peer Review"
  - provides a summary of a topic or field of research
  - reads other articles
  - does NOT do any original research

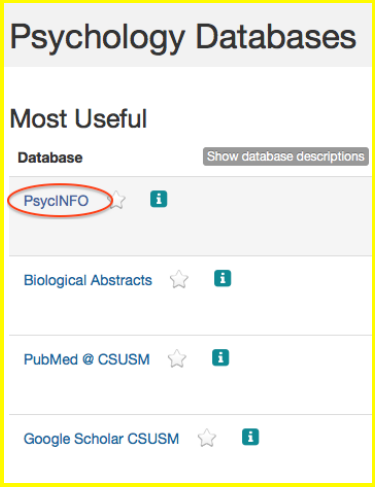
Types of Articles

- Meta Analysis
  - a type of literature review that also does some statistical research
  - re-analyzes other research

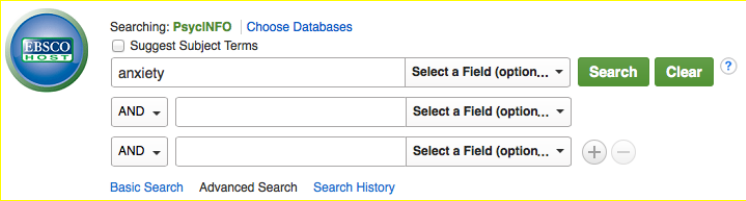
Searching PsycInfo



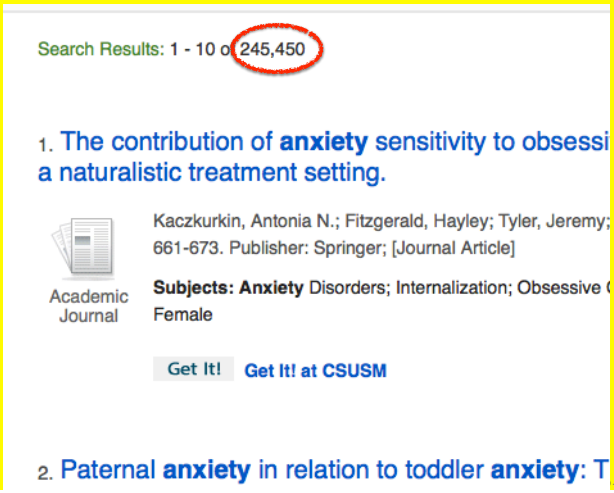
Searching PsycInfo



Search : Anxiety

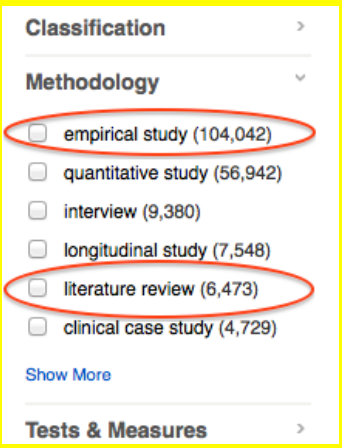


Anxiety Results



Type of Publication

- Methodology
- Empirical Study
  - aka Original Study
- Literature Review
  - aka Review

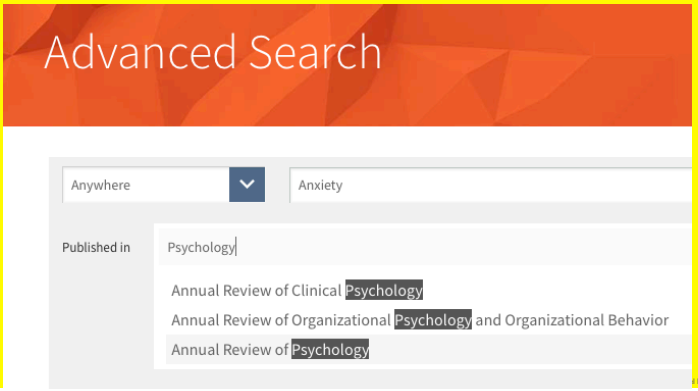


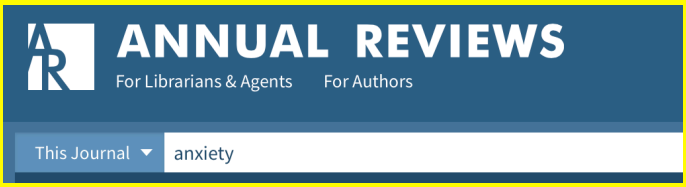
Other Sources

- Some journals specialize in *Review* articles
- https://annualreviews.org

Searching AnnualReviews.org

- Advanced Search - search only in **Annual Review of Psychology**



Searching AnnualReviews.org

- Choose the journal first, then search only in **This Journal**