Type I vs. Type II Errors in Neuropsychology

David W. Loring
Emory University

Disclosures

- Employee of Emory University
- Active Research Support
  - *NINDS, NIMH*
- Consultant for *NeuroPace, Inc*
- Book Royalties
  - *Oxford University Press*
- Editorial Stipends
  - *EPILEPSIA*
  - *NEUROPSYCHOLOGY REVIEW*

Learning Objectives

- Appreciate how confidence intervals affect diagnostic thresholds according to clinical history/presentation
- Understand statistical factors influencing critical threshold
- Discuss quantitative history and meaning of *p*<.05

YOU CAN'T BE A MATHEMATICIAN AND HAVE A SENSE OF HUMOR.
YOU CAN'T BE A MATHEMATICIAN AND NOT HAVE A SENSE OF HUMOR.
S. harris

Neuropsychological Testing

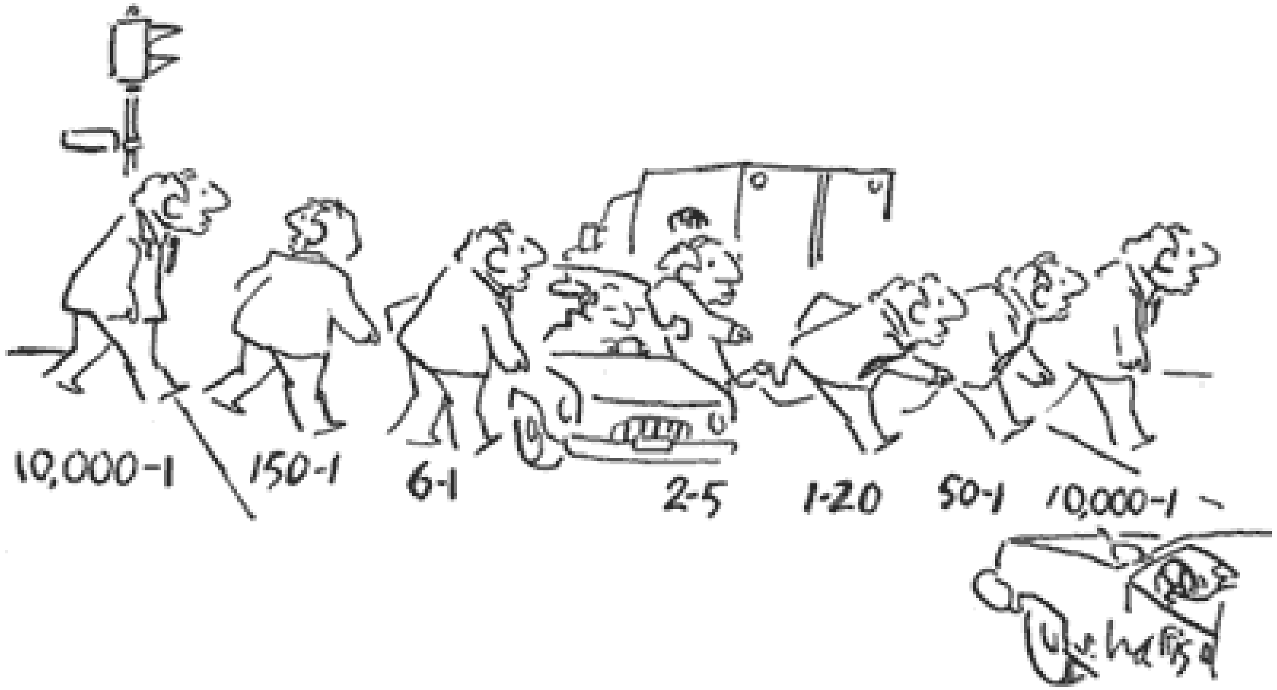
- … standardized assessment tools and integrate the findings with other data to determine whether cognitive decline has occurred:
    1. to differentiate neurologic from psychiatric conditions
    2. to identify neurocognitive etiologies, and
    3. to determine the relationship between neurologic factors and difficulties in daily functioning
    4. *(to establish the presence of interval change)*

American Family Physician (2019)

Neuropsychological Assessment & Norms

- Comparison of obtained score to reference group
  - Ideally, large cognitively healthy group of similar demographics from formally standardized sample
  - May be compared to patient groups with specific diagnoses of clinical characteristics (abnormal norms)
- Referral questions typically dichotomous: yes vs. no
  - Diagnostic
  - Functional capacity
  - Surgical candidacy
  - Interval decline
- At the test level, relies on cutpoints from continuous variable distributions

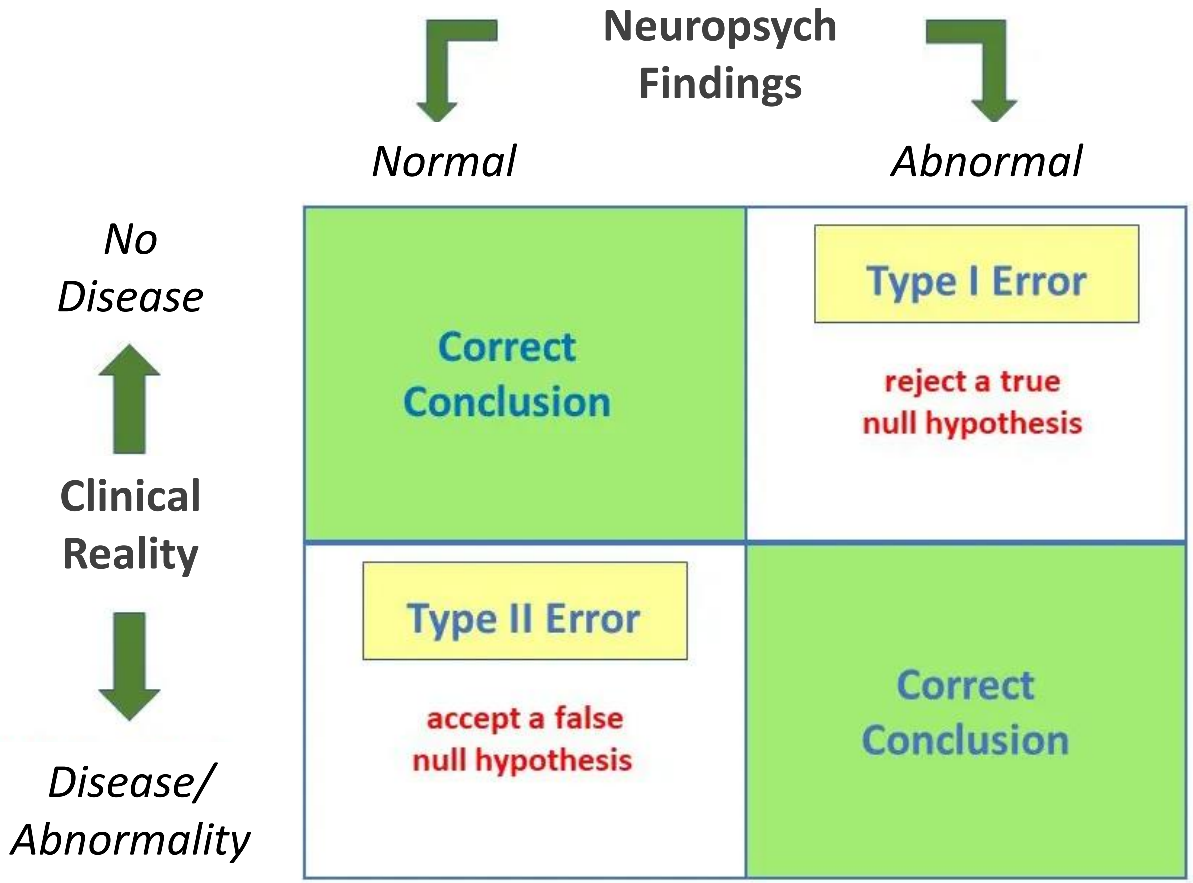
Probabilities



Superhero: Shamrock (Marvel Comics)

- Daughter of militant IRA member (1982)
- Vessel for displaced poltergeists and souls who died as innocent war victims
- Superpower: probability manipulation (good luck)

Diagnostic Possibilities



Neuropsych Findings

Clinical Reality

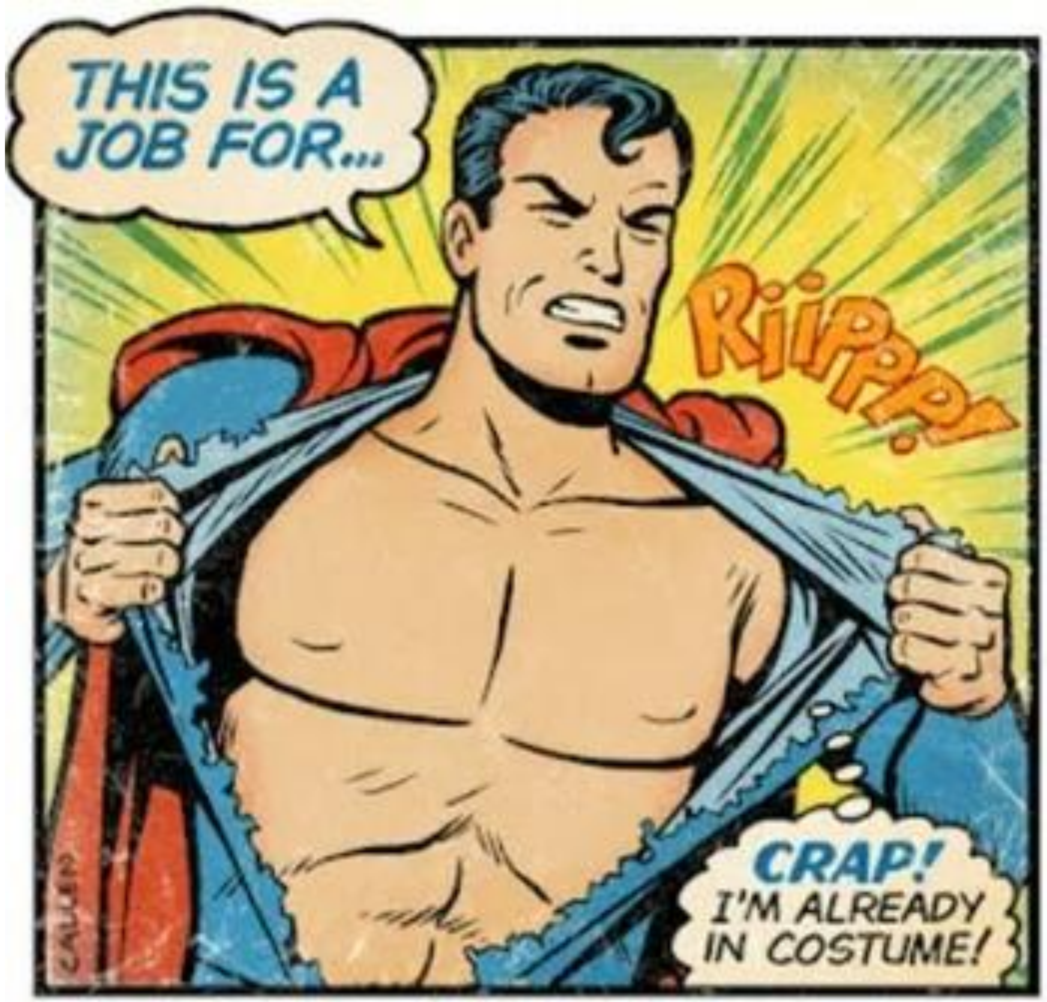
	Normal	*Abnormal*
No Disease	Correct Conclusion	Type I Error reject a true null hypothesis
Disease/ Abnormality	Type II Error accept a false null hypothesis	Correct Conclusion

https://www.simplypsychology.org/type_I_and_type_II_errors.html

Type I Superhero Readiness Error

THIS IS A JOB FOR...
CRAP! I'M NOT WEARING MY COSTUME!
CALLEN

Type II Superhero Readiness Error



Type I Diagnostic Error in Neuropsychology

- Make inference of abnormality when none exists
  - Neurologic disease, interval change, absence of capacity
  - Malingering, response exaggeration
- May lead to unnecessary diagnostic evaluations
  - Waste of resources and cost, time off work, medical risk (e.g., LP)
- May lead to patient worry regarding prognosis associated with abnormality
- Denial of benefits with Performance Validity Measures

Type II Diagnostic Error in Neuropsychology

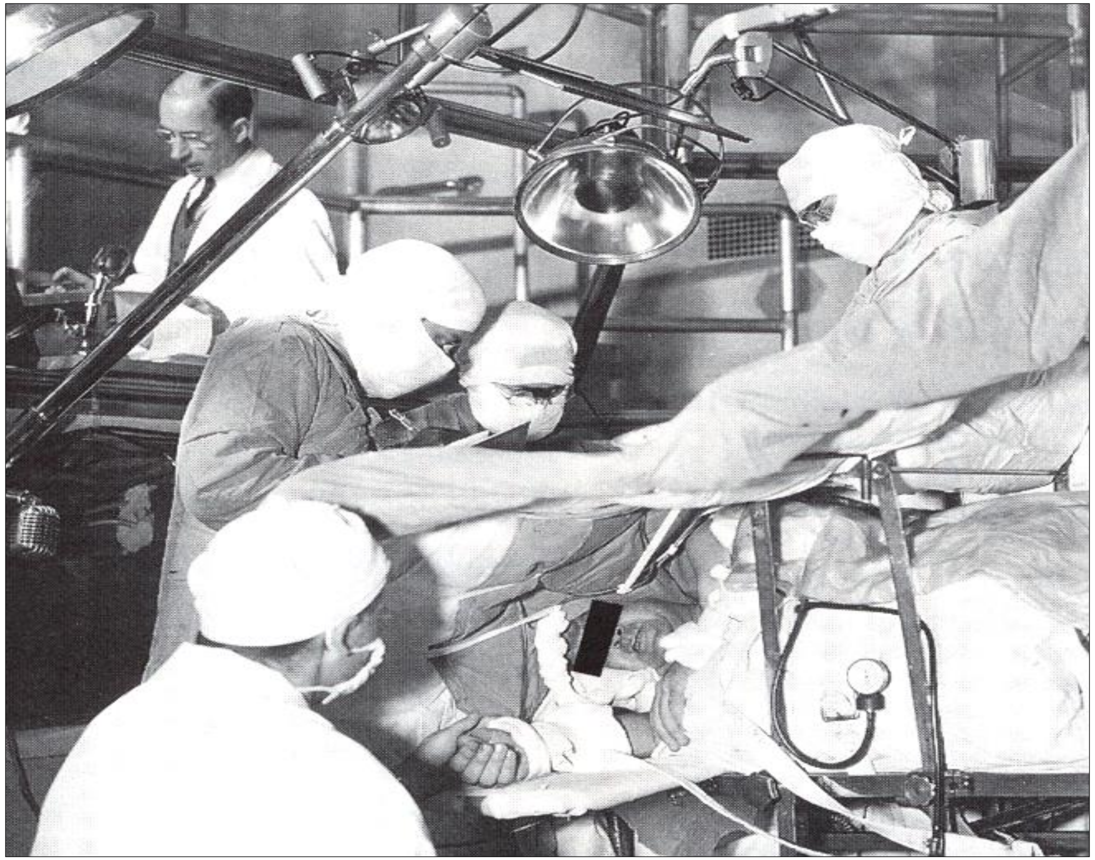
- Makes diagnostic inference of normality when underlying pathology/clinical condition is present
- May lead to treatment or intervention delay
  - By clinician
  - By patient
- May lead to safety or other risks (e.g., driving)

Type I vs. Type II Risk Tradeoff

- Test level
  - Method/threshold of inferring abnormality
  - Normative dataset
    - Representativeness to individual patient
    - Quality of normative samples
- Patient level
  - Diagnostic criteria applied
  - Quality of published research
  - Clinical skill of neuropsychologist

Case Example – Possible interval decline?

- 27-year-old woman with prior left temporal resection
- Nothing unusual in preoperative workup diagnostically (i.e., EEG, PET, MRI all consistent with left TLE)
- No red flags with neuropsych testing (verbal abilities including naming lower than other skills)
- Reports decline in memory during FU clinical interview

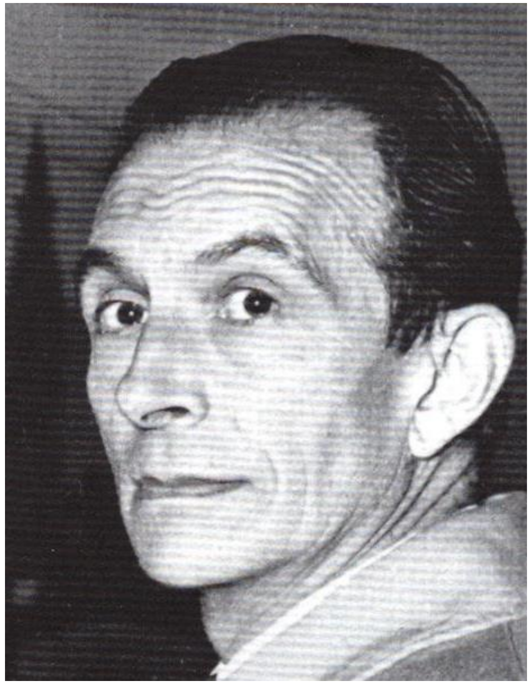


AVLT Interval Change Scenarios

- Pre = 65/75, Post = 45/75
- Pre = 65/75, Post = 40/75

- Pre = 50/75, Post = 45/75
- Pre = 50/75, Post = 40/75

- Pre = 35/75, Post = 45/75
- Pre = 35/75, Post = 40/75



André Rey (1906-1965)

Which of the following is true regarding RCIs?

A. RCIs are an objective measure of clinically meaningful change

B. RCIs must be empirically derived from subjects tested on multiple occasions in the absence of treatment interventions

C. A 90% RCI is preferred to an 80% RCI

D. A decline in function cannot be inferred unless the difference score is greater than an *a priori* statistical threshold

E. None of the above

Empirical RCI (epilepsy patients, 8-month FU)

- AVLT Total; max=75
  - *r* = .69, 90% RCI = ± 15
- What is the SD of AVLT Total in sample?
  - 10.4
- What is the AVLT Total Mean in sample?
  - 48.4
  - Most basic interpretation: WNL Δ = 33.4-63.4

Sawrie et al. (1996). *JINS*, *2*(6), 556-564

AVLT Confidence Intervals Estimates Derived from Mayo Normative Studies

Measure	Pearson Coefficient	Standard Deviation	$SE_{Estimation}$	Single Score 90% CI	$SE_{Predication}$	Interval Change 90% CI
Trial 1-5 Sum	.798	10.0	4.01	13.2	6.03	19.82
List B	.507	1.7	0.85	2.80	1.47	4.81
Delayed Recall	.761	3.5	1.49	4.92	2.27	7.46

Loring et al. (2022), *JINS*

Sawrie: *r* = .69, SD=10.4; 90% RCI = ± 15

Common Error in CI application

- If 90% CI = 20, then scores is reported as *x* (90% CI=*x*-10 to *x*-20) regardless of position in distribution
  - e.g., 50 (90% CI=40-50)
- What then for lower score?
  - 20 (90% CI=10 - 30)
  - 10 (90% CI= 0 - 20)
  - 5 (90% CI= -5 - 10)

CI Differences with Individual Test Scores

- Regression to the mean
- Scores at distribution tails are likely to be “extreme” due in part to chance fluctuation
- Upon retesting, score will likely be closer to the distribution mean reflecting chance contribution with original perforamnce

Confidence Interval Midpoint

- Appropriate “midpoint” anchor for Confidence Intervals is *Predicted True Score*.
- *Predicted True Score* always falls between the observed score and the population mean
- Calculated by reliability [*predicted true score* = (observed score * reliability) + ((population mean * (1-reliability))].

Midpoint Example

- AVLT Total reliability = .8
- Observed *T*=40 associated with predicted *T*=42 ([40 * .8]+ [50 * .2], or [32 + 10])
- NOT, 40 (90% CI = 33-47)
- BUT RATHER, 40 (90% CI = 35-49)
- Lower reliabilities results in larger adjustments from observed to predicated true score

WAIS-IV Performance Discrepancy and Conditional Probabilities

- Patient obtains VCI = 92, PRI = 101
- According to WAIS-IV manual, 9-point discrepancy occurred in 24% of normative sample (Table B.2); average discrepancy=10.5 points (SD=8.0)
- Non-lateralizing: occurs with some frequency in healthy subjects (p>.05)

Abnormal norms, base rates

- Epilepsy surgery evaluation
  - Is VCI vs. PRI discrepancy of 9 points more likely to be left brain epilepsy or right brain epilepsy? Or normal?
- Post-stroke evaluation
  - Is VCI vs. PRI of 9 points more likely to be left brain epilepsy or right brain epilepsy? Or normal?
- Memory evaluation
  - Is PRI vs. VCI discrepancy of 9 points more likely to be AD or DLB?

Which of the following is/are true?

A. $p > .05$ is the probability that the null hypothesis is true

B. 1 – p is the probability that the alternative hypothesis is true

C. $p < .05$ means that the tested hypothesis is false and should be rejected

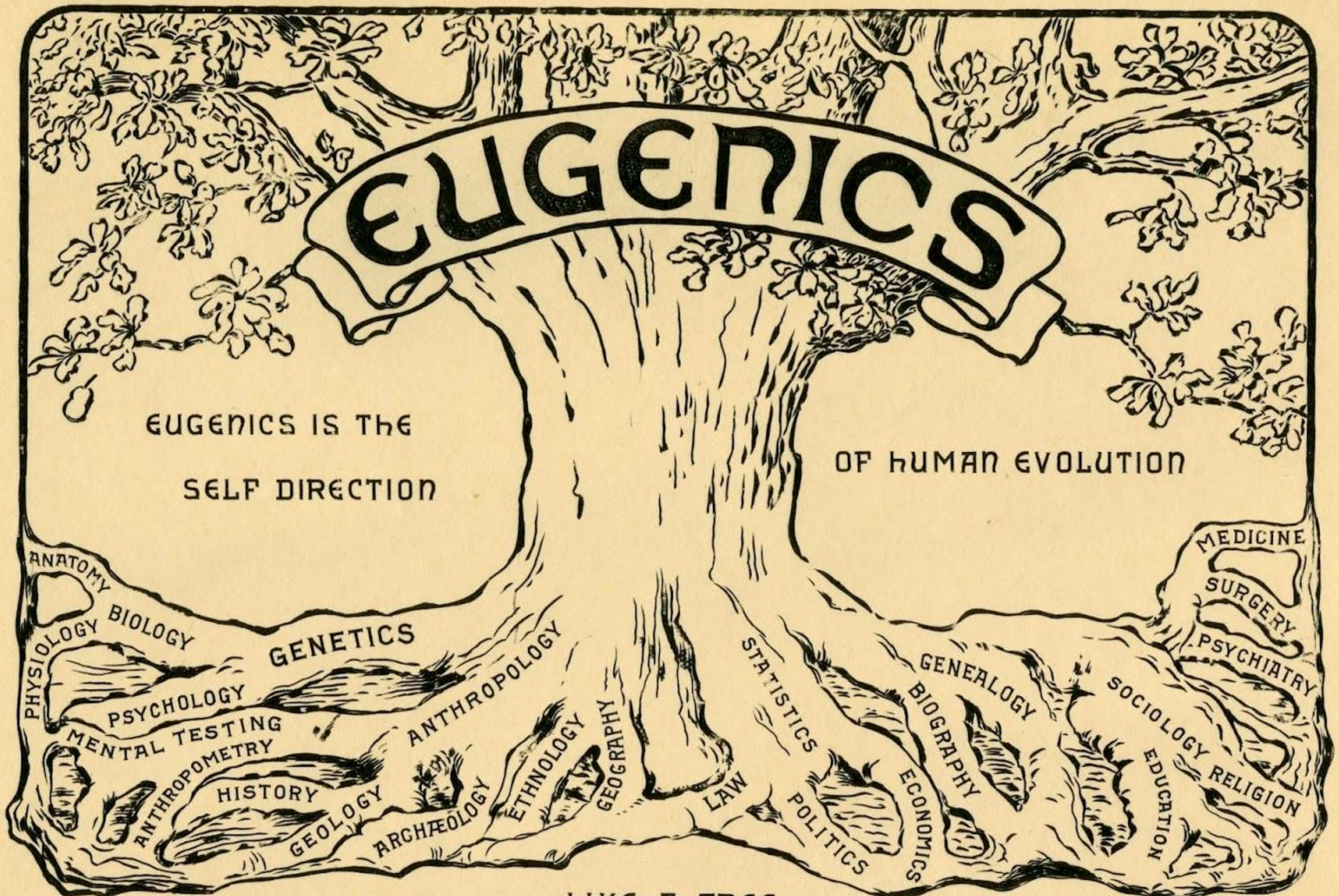
D. $p > .05$ means no effect was observed

E. None of the Above

Why p<.05?

Sir Ronald A. Fisher (1890-1962)

- Galton Professor of Eugenics, University College London
- Editor, *Annals of Eugenics*



LIKE A TREE
EUGENICS DRAWS ITS MATERIALS FROM MANY SOURCES AND ORGANIZES
THEM INTO AN HARMONIOUS ENTITY.

PLATE V.

KEY TO HEREDITY CHART.

	Male.	Female.	
	□	○	No Data.
Red	E	E	Epileptic.
Black	F	F	Feeble-minded.
Green	I	I	Insane.
Violet	C	C	Criminalistic.

Other letters used in or around the squares or circles are:

A	Alcoholic.
B	Blind.
D	Deaf.
M	Migraneous.
N	Normal.
Ne.	Neurotic.
P	Paralytic.
Sx.	Sexually immoral.
S	Syphilitic.
T	Tubercular.
W	Wanderer or confirmed runaway.

FIGURES

Epilepsia, 51(7):1107–1119, 2010
doi: 10.1111/j.1528-1167.2009.02486.x

HISTORICAL REVIEW

100 years of *Epilepsia*: Landmark papers and their influence in neuropsychology and neuropsychiatry

Bruce Hermann

Department of Neurology, Matthews Neuropsychology Lab, University of Wisconsin School of Medicine and Public Health,
Madison, Wisconsin, U.S.A.

Summary

As part of the 2009 International League Against Epilepsy (ILAE) Centenary Celebration, a special symposium was dedicated to *Epilepsia* (*100 Years of Epilepsia: Landmark Papers and Their Influence*). The Associate Editors were asked to identify a particularly salient and meaningful paper in their areas of expertise. From the content areas of neuropsychology and neuropsychiatry two very interesting papers were identified using quite different ascertainment techniques. One paper addressed the problem of psychosis in temporal lobe epilepsy, whereas the other represents the first paper to appear in *Epilepsia* presenting quantitative assessment of cognitive status in epilepsy. These two papers are reviewed in detail and placed in historical context.

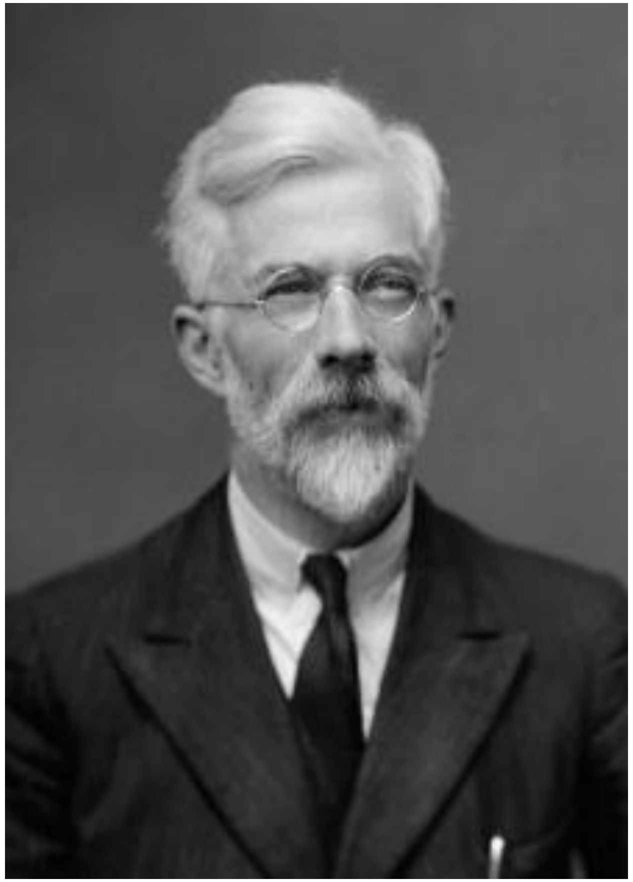
KEY WORDS: Epilepsy, Neuropsychiatry, Neuropsychology, Epilepsy colony, Eugenics.

Other Famous Statisticians/Eugenicists

- *Karl Pearson* (1857-1936) – First Galton Chair of Eugenics “superior and inferior races cannot coexist; if the former are to make effective use of global resources; the latter must be extirpated”
- *Charles Spearman* (1863-1945). “The general conclusion emphasized by near every investigator is that as regards “intelligence”, the Germanic stock has on the average a marked advantage over the South European. And this result would seem to have had vitally important practical consequences in shaping the recent very stringent American laws as to admission of immigrants”
- *Raymond B. Cattell* (1905-1998) - "intelligence tests point to significant differences between races“ I think it might not be a bad idea to remove the inscription from the Statue of Liberty which calls for the “wretched refuse” of the other countries to migrate here. This is not what you want to build a nation of. If we have immigration, we ought to have it from the best sources.”

Why *p* < .05?

- “The value for which P = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.”



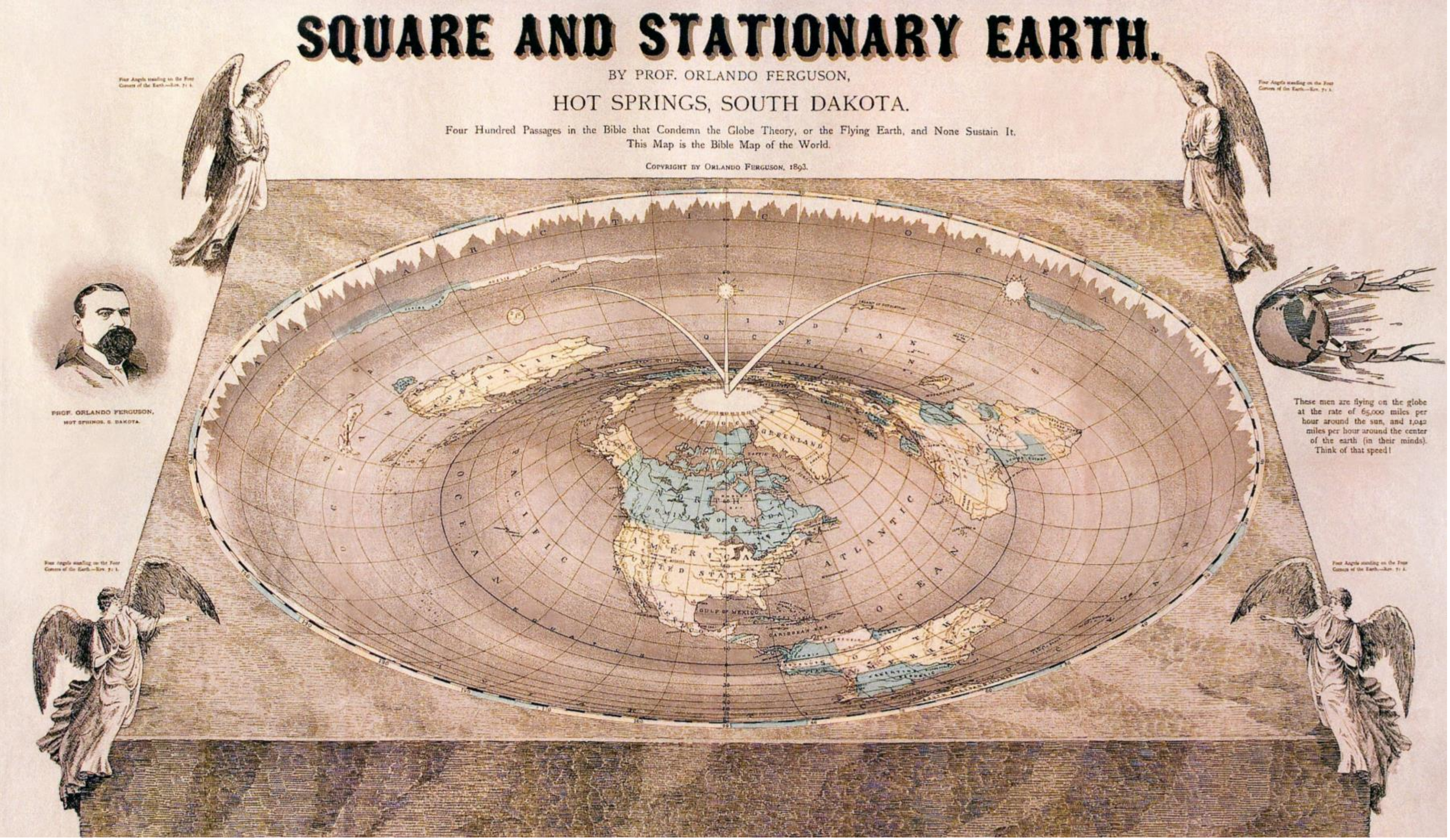
Fisher, R.A. (1925). *Statistical Methods for Research Workers.*

Why *p* < .05?

- “...If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred … A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance...”

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503-513.

The Earth is Round (*p* < .05)


SQUARE AND STATIONARY EARTH.
BY PROF. ORLANDO FERGUSON,
HOT SPRINGS, SOUTH DAKOTA.
Four Hundred Passages in the Bible that Condemn the Globe Theory, or the Flying Earth, and None Sustain It.
This Map is the Bible Map of the World.
COPYRIGHT BY ORLANDO FERGUSON, 1893.
Four Angels standing on the Four Corners of the Earth.—Rev. 7: 1.
PROF. ORLANDO FERGUSON,
HOT SPRINGS, S. DAKOTA.
These men are flying on the globe at the rate of 65,000 miles per hour around the sun, and 1,042 miles per hour around the center of the earth (in their minds). Think of that speed!
ARCTIC OCEAN
INDIAN OCEAN
PACIFIC OCEAN
ATLANTIC OCEAN
GREENLAND
DOMINION OF CANADA
AMERICA
UNITED STATES
GULF OF MEXICO

Cohen, J. (1994). *American Psychologist, 49*(12), 997-1003.

“Significant” Results

- “The primary product of a research inquiry is one or more measures of effect size, not *p* values.”

  Jacob Cohen (1990)

- “Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them.”

  Gene Glass (2004)

American Statistical Society Statement on *p*-Values

- “The statistical community has been deeply concerned about issues of reproducibility and replicability of scientific conclusions … much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices such as … to ban *p*-values …”

Mayo, D.G. London School of Economics Karl Popper talk, May 10, 2016

Matrixx Initiatives, Inc. v. Siracusano
U.S. 27, 131 S. Ct. 1309 (2011)

- Issue: Is statistical significance required to establish materiality and scienter for a plaintiff to state a claim for securities fraud where a manufacturer withheld damaging information about its product?
- Supreme Court 9-0: NO
  - *Justice Sotomayor: "Given that medical professionals and regulators act on the basis of evidence of causation that is not statistically significant, it stands to reason that in certain cases reasonable investors would as well."*

12 Angry Men

Civil Litigation Evidentiary Standards

- … the burden of persuasion that applies is called “a preponderance of the evidence.” This standard requires the jury to return a judgment in favor of the plaintiff if the plaintiff is able to show that a particular fact or event was more likely than not to have occurred. Some scholars define the preponderance of the evidence standard as requiring a finding that at least 51 percent of the evidence favors the plaintiff’s outcome.

https://www.justia.com/trials-litigation/lawsuits-and-the-court-process/evidentiary-standards-and-burdens-of-proof/

Reliable Change

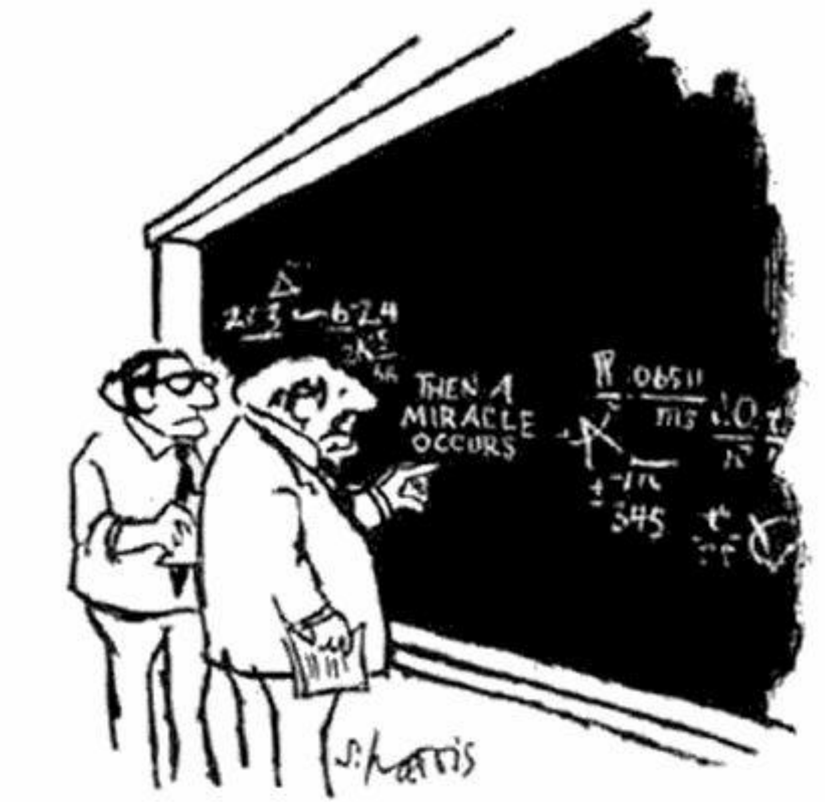
- Should RCIs should be 2-tailed?
- What is the difference between 80% and 90% RCI?
- Why not 75% or 70%?
- Why not “more likely than not?”

Need for Evidence Based Medicine:
Type I vs. Type II in Clinical Research Reports

Law of Large Numbers vs. Law of Small Numbers

- *Law of Large Numbers*
  - As sample size increases, sample becomes more representative of the whole population (Bernouli, 1713)
  - Small sample sizes may not be representative of the population of interest
- *Law of Small Numbers*
  - The belief that any randomly obtained sample, particularly those with small sample sizes, will be representative of the population of interest with similarity on all critical features

Published Research Predicting Reality



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

- Less likely to be true:
  - Smaller effect size
  - Number of published studies is smaller
  - Greater variability in designs, definitions, outcomes, and analytic modes
- More likely to be true:
  - Larger number of studies and fewer tested relationships (confirmatory designs, meta-analyses)

Ioannidis, J. P. (2005). *PLoS Med.*

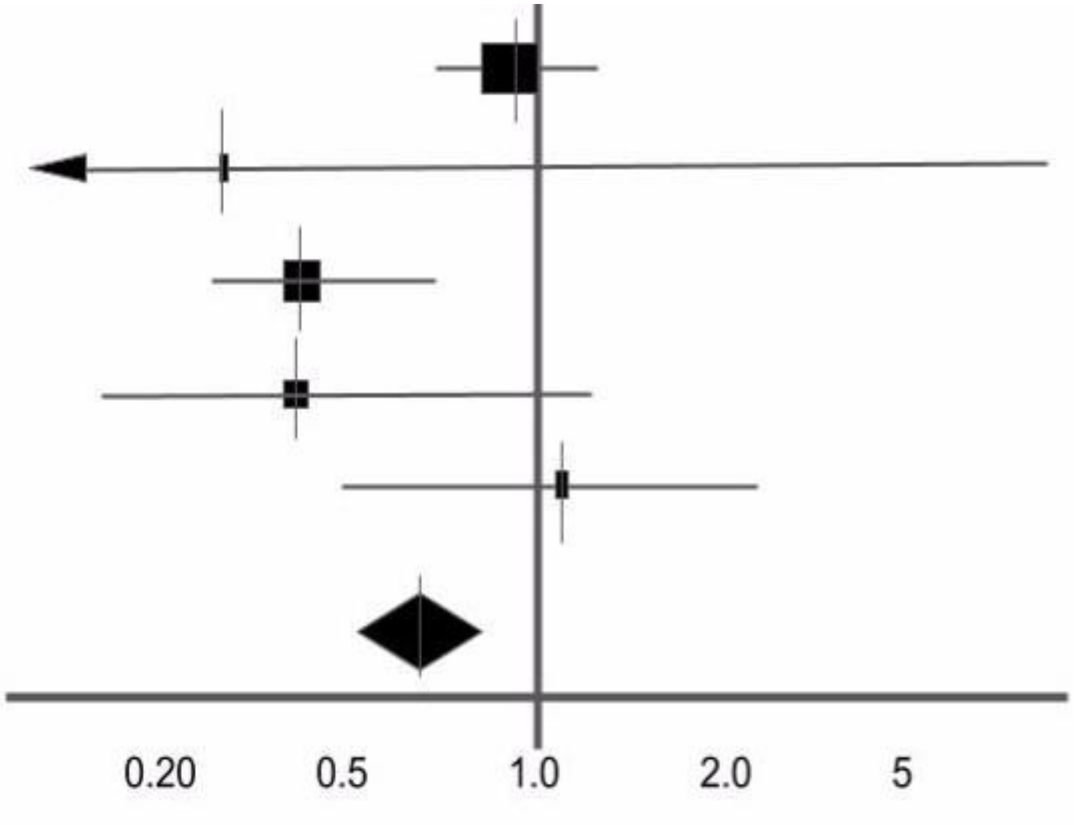
Type of Heterogeneity

- Clinical heterogeneity
  - Variability in participants, interventions and outcomes
- Methodological heterogeneity
  - Variability in study design and risk of bias
- Statistical heterogeneity
  - Variability in treatment effects, resulting from clinical or methodological variability

Meta Analysis

- What is the direction of the effect?
- What is the size of the effect?
- Is the effect consistent across studies?
- What is the strength of evidence for the effect
  - Relies additional on judgments based on assessments of study designs and risk of bias as well as statistical measures of uncertainty.

Forest Plots



Cochrane

Summary

- Type I vs. Type II errors occur in all neuropsychological decision-making contexts
- Statistics apply to the test level, and not to construct level
- Type I vs. Type II relative risks will vary based upon assessment context and patient characteristics

Thank you!

“Actually, I was hoping for a more inspiring mission statement.”