

# Survey Methods & Design in Psychology

Lecture 11 (2007)

Significance Testing,  
Power,  
Effect Sizes,  
Confidence Intervals,  
Publication Bias, &  
Scientific Integrity

Lecturer: James Neill

## Overview

- Significance testing
- Inferential decision making
- Power
- Effect Sizes
- Confidence intervals
- Publication Bias
- Scientific Integrity

## Readings

Howell Statistical Methods:

- Ch8 Power

Concepts rely upon:

- Ch3 The Normal Distribution
- Ch4 Sampling Distributions and Hypothesis Testing
- Ch7 Hypothesis Tests Applied to Means

## Significance Testing

## Significance Testing

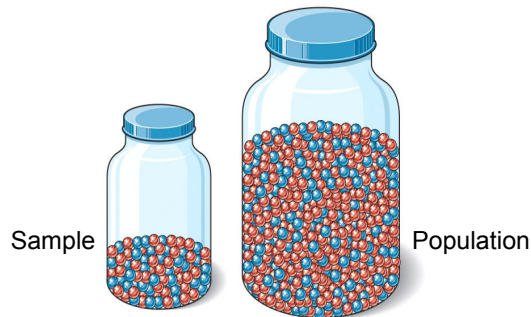
- Logic
- History
- Criticisms
- Hypotheses
- Inferential decision making table
  - Type I & II errors
  - Power
  - Effect Size (ES)
  - Sample Size ( $N$ )

## The Logic of Significance Testing

In a betting game, how many straight heads would I need to throw until you cried foul?

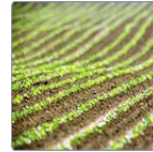


## The Logic of Significance Testing



## History of Significance Testing

- A 20<sup>th</sup> C phenomenon.
- Developed by Ronald Fisher for testing the variation in produce per acre for agriculture crop (1920's-1930's)



## History of Significance Testing

- To help determine what agricultural methods (IVs) yielded greater output (plant growth) (DVs)
- Designs couldn't be fully experimental, therefore, needn't to determine whether variations in the DV were due to chance or the IV(s).

## History of Significance Testing

- Proposed  $H_0$  to reflect expected ES in the population
- Then get  $p$ -value from data about the likelihood of  $H_0$  being true &, depending of level of false positives the researcher is prepared to tolerate (critical alpha), make decision about  $H_0$

## History of Significance Testing

- ST spread to other fields, including social science
- Spread in use aided by the development of computers and training.
- In the latter decades of the 20<sup>th</sup> C, widespread use of ST attracted critique for its over-use and mis-use.

## Criticisms of Significance Testing

- Critiqued as early as 1930
- Cohen (1980's-1990's) critiqued
- During the late 1990's a critical mass of awareness developed and there are currently changes underway in publication criteria and teaching with regard to over-reliance on ST

## Criticisms of Significance Testing

- Null hypothesis is rarely true

## Criticisms of Significance Testing

- NHT only provides a binary decision (yes or no) and indicates the direction
- Mostly we are interested in the size of the effect – i.e., *how much* of an effect?

## Criticisms of Significance Testing

- Whether a result is significant is a function of:
  - ES
  - $N$
  - critical  $\alpha$  level
- Sig. can be manipulated by tweaking any of the three (as each of them increase, so does the likelihood of a significant result)

## Criticisms of Significance Testing

ears. For example, Frank Yates (1951), a contemporary of Fisher, observed that the use of the null hypothesis significance test

has caused scientific research workers to pay undue attention to the results of the tests of significance that they perform on their data and too little attention to the estimates of the magnitude of the effects they are investigating. . . . The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers often have regarded the execution of a test of significance on an experiment as the ultimate objective. (pp. 32-33)

## Criticisms of Significance Testing

A more strongly worded criticism of null hypothesis significance testing was written by Paul Meehl (1978):

I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (p. 817)

GILL, CALIFORNIA POLYTECHNIC STATE UNIVERSITY

The current method of hypothesis testing in the social sciences is under intense criticism, yet most political scientists are unaware of the important issues being raised. Criticisms focus on the construction and interpretation of a procedure that has dominated the reporting of empirical results for over fifty years. There is evidence that null hypothesis significance testing as practiced in political science is deeply flawed and widely misunderstood. This is important since most empirical work argues the value of findings through the use of the null hypothesis significance test. In this article I review the history of the null hypothesis significance testing paradigm in the social sciences and discuss major problems, some of which are logical inconsistencies while others are more interpretive in nature. I suggest alternative techniques to convey effectively the importance of data-analytic findings. These recommendations are illustrated with examples using empirical political science publications.

## Statistical vs Practical Significance

- Statistical significance means that the observed mean differences are not likely due to sampling error
  - Can get statistical significance, even with very small population differences, if  $N$  is large enough
- Practical significance looks at whether the difference is large enough to be of value in a practical sense
  - Is it an effect worth being concerned about – does it have any noticeable or worthwhile effects?

### Recommendations

- Learn traditional Fisherian logic methodology (inferential testing)
- Learn alternative techniques (ESs and CIs)
- -> Use ESs and CIs as alternatives or complements to STs.

## Inferential Decision Making

## Significance Testing - Summary

- Logic: Sample data examined to determine likelihood it represents a population of no effect or some effect.
- History: Developed by Fisher for agricultural experiments in early 20<sup>th</sup> C
- Spread aided by computers to social science
- In recent decades, ST has been criticised for over-use and mis-application.

### Recommendations

- APA 5<sup>th</sup> edition recommends reporting of ESs, power, etc.
- Recognise merits and shortcomings of each approach
- Look for *practical* significance

## Hypotheses in Inferential Testing

Null Hypothesis ( $H_0$ ):  
No differences

Alternative Hypothesis ( $H_1$ ):  
Differences

## Inferential Decisions

When we test a hypothesis we draw a conclusion; either

**Accept  $H_0$**

$p$  is not significant (i.e. not below the critical  $\alpha$ )

**Reject  $H_0$ :**

$p$  is significant (i.e., below the critical  $\alpha$ )

## Type I & II Errors

When we accept or do not accept  $H_0$ , we risk making one of two possible errors:

**Type I error:**

Reject  $H_0$  when it is actually correct

**Type II error:**

Retain  $H_0$  when it is actually false

## Correct Decisions

When we accept or do not accept  $H_0$ , we are hoping to make one of two possible correct decisions:

**Correct rejection of  $H_0$  (Power):**

Reject  $H_0$  when there is a real difference

**Correct acceptance of  $H_0$ :**

Retain  $H_0$  when there is no real difference

## Inferential Decision Making Table

		Reality	
		$H_0$ False	$H_0$ True
Test	Reject $H_0$	Correct rejection $H_0$ = Power = $1 - \beta$	Type I error = $\alpha$
	Accept $H_0$	Type II error	Correct acceptance of $H_0$

## Significance Testing - Summary

- Type I error (false rejection of  $H_0$ ) =  $\alpha$
- Type II error (false acceptance of  $H_0$ ) =  $\beta$
- Power (false rejection of  $H_0$ ) =  $1 - \beta$
- Correct acceptance of  $H_0$  =  $1 - \beta$

Power

## Power

- The probability of rejection of a false null-hypothesis
- Depends on the:
  - Critical alpha ( $\alpha$ )
  - Sample size (N)
  - Effect size ( $\Delta$ )

## Power

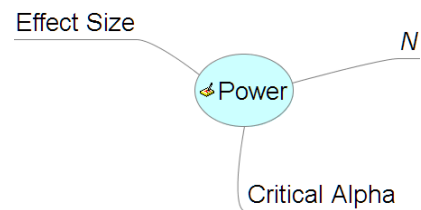
		Reality	
		$H_0$ False	$H_0$ True
Test	Reject $H_0$	POWER	Type I error = $\alpha$
	Accept $H_0$	Type II error	Correct acceptance of $H_0$

## Power

- = Likelihood that an inferential test will return a sig. result when there is a real difference
- = Probability of correctly rejecting  $H_0$
- = 1 - likelihood that an inferential test won't return a sig. result when there is a real difference ( $1 - \beta$ )
- Desirable power > .80
- Typical power ~ .60

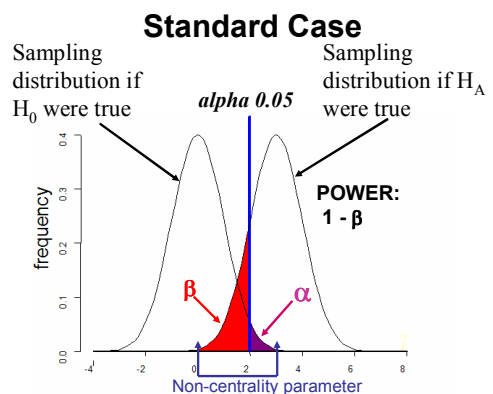
## Power

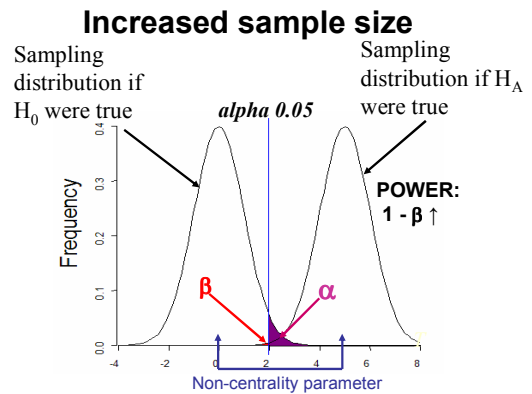
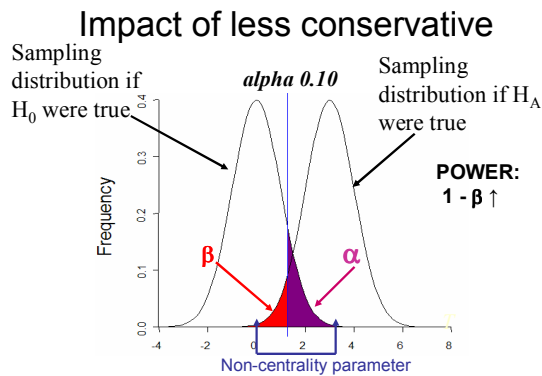
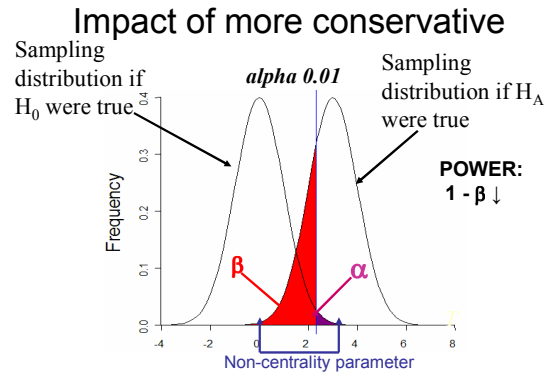
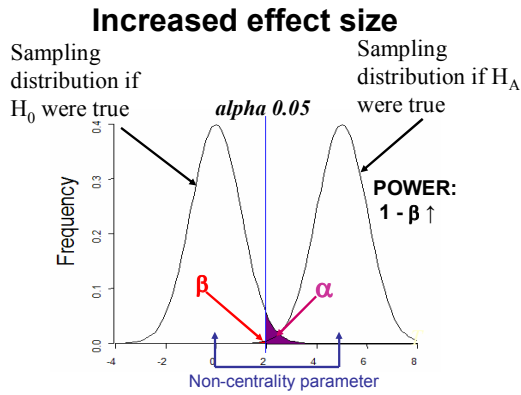
An inferential test is more 'powerful' (i.e. more likely to get a significant result) when any of these 3 increase:



## Power Analysis

- If possible, calculate expected power beforehand, based on:
  - Estimated N,
  - Critical  $\alpha$ ,
  - Expected or minimum ES (e.g., from related research)
- Also report actual power in the results.
- Ideally, power ~ .80 for detecting small effect sizes





## Power Summary

- Power is the likelihood of detecting an effect as statistically significant
- Power can be increased by:
  - $\uparrow N$
  - $\uparrow$  critical  $\alpha$
  - $\uparrow ES$
- Power over .8 “desirable”
- Power of  $\sim .6$  is more typical
- Can be calculated prospectively and retrospectively

## Effect Sizes

## Effect Sizes

- ESs express the degree or strength of relationship or effect
- Not influenced the  $N$
- ESs can be applied to any inferential test, e.g.,
  - $r$  for correlational effects
  - $R$  for multiple correlation effects
  - $d$  for difference between group means
  - *eta-squared* ( $\eta^2$ ) for multivariate differences between group means

## Commonly Used Effect Sizes

### Standardised Mean difference

- Cohen's  $d$
- $F / \eta^2$

### Correlational

- $r, r^2$
- $R, R^2$

## Cohen's $d$

- A standardised measure of the difference between two  $M$ s
- $d = M_2 - M_1 / \sigma$

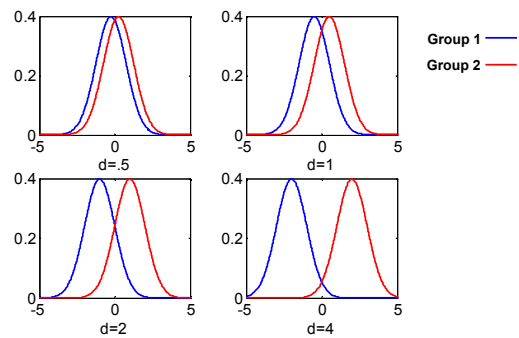
## Cohen's $d$

- Cohen's  $d$ 
  - ve = negative change
  - 0 = no change
  - +ve = positive change

## Effect sizes – Cohen's $d$

- Not readily available in SPSS
- Cohen's  $d$  is the standardized difference between two means

## Example Effect Sizes



## Interpreting Standardised Mean Differences

- Cohen (1977): .2 = small  
.5 = moderate  
.8 = large
- Wolf (1986): .25 = educationally significant  
.50 = practically significant (therapeutic)
- Standardised Mean ESs are proportional, e.g., .40 is twice as much change as .20

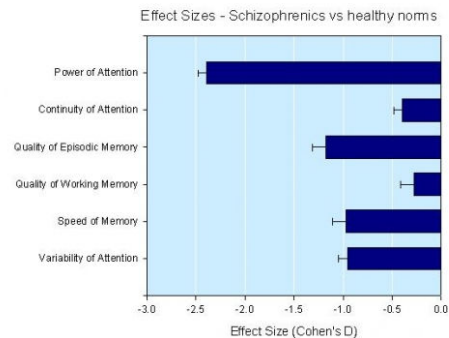
## A Small Effect Size Can be Impressive...

- In practice, a small ES can be very impressive if, for example:
  - the outcome is difficult to change (e.g. a personality construct) or if
  - the outcome is very valuable (e.g. an increase in life expectancy).
- On the other hand, a large ES doesn't necessarily mean that there is any practical value if it isn't related to the aims of the investigation (e.g. religious orientation).

## Interpreting Effect Size

- No agreed standards for how to interpret an ES
- Interpretation is ultimately subjective
- Best approach is to compare with other studies

## Graphing Effect Size - Example



## Effect Size Table - Example

	Group	Mean	Std. error	95% Confidence interval		P level	Effect size
				Lower	Upper		
Knowledge about oral cancer	no leaflet	26.11	0.19	25.73	26.48	0.001	1.29
	leaflet	30.97	0.18	30.51	31.24		
Attitudes about negative consequences	no leaflet	3.97	0.08	3.81	4.13	0.038	0.15
	leaflet	3.73	0.08	3.57	3.88		
Attitudes about lack of control	no leaflet	7.91	0.09	7.72	8.10	0.078	0.13
	leaflet	7.67	0.09	7.49	7.86		
Normative beliefs	no leaflet	13.34	0.25	12.84	13.83	0.019	0.17
	leaflet	12.51	0.24	12.03	12.99		
Anxiety about screening procedure	no leaflet	5.58	0.13	5.31	5.85	0.069	0.13
	leaflet	5.23	0.13	4.97	5.50		
Intention to accept screen	no leaflet	11.61	0.12	11.36	11.86	0.003	0.22
	leaflet	12.15	0.12	11.91	12.39		

## Effect sizes – Exercise

- 20 athletes rate their personal playing ability, with M = 3.4 (SD = .6) (on a scale of 1 to 5)
- After an intensive training program, the players rate their personal playing ability, with M = 3.8 (SD = .6)
- What is the ES and how good was the intervention?
- What is the 95% CI and what does it indicate?

## Effect sizes - Answer

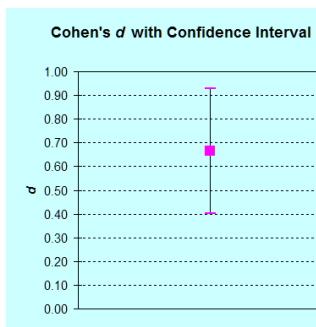
Cohen's  $d$

- $= (M_2 - M_1) / SD$
- $= (3.8 - 3.4) / .6$
- $= .4 / .6$
- $= .67$
- = a moderate-large change over time

## Effect sizes - Answer

Mean 1	Std. Dev.1	N1	Confid. Level	Mean Difference	$d$ lower limit
3.8	0.6	20	0.95	0.40	0.40
Mean 2	Std. Dev.2	N2	Pooled Variance	Cohen's $d$	$d$ upper limit
3.4	0.6	20	0.60	0.67	0.93

## Effect sizes - Answer



## Effect sizes - Summary

- ES indicates amount of difference or strength of relationship - underutilised
- Inferential tests should be accompanied by ESs and CIs
- Most common ESs are Cohen's  $d$  and  $r$
- Cohen's  $d$ :
  - .2 = small
  - .5 = moderate
  - .8 = large
- Cohen's  $d$  is not provided in SPSS – can use a spreadsheet calculator

## Power & Effect sizes in Psychology

Ward (2002) examined articles in 3 psych. journals to assess the current status of statistical power and effect size measures.

- *Journal of Personality and Social Psychology*,
- *Journal of Consulting and Clinical Psychology*
- *Journal of Abnormal Psychology*

## Power & Effect sizes in Psychology

- 7% of studies estimate or discuss statistical power.
- 30% calculate ES measures.
- A medium ES was discovered as the average ES across studies
- Current research designs do not have sufficient power to detect such an ES.

## Confidence Intervals

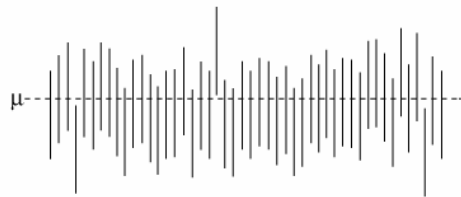
## Confidence Intervals

- Very useful, underutilised
- Gives 'range of certainty' or 'area of confidence' e.g., true  $M$  is 95% likely to lie between  $-1.96$  SD and  $+1.96$  of the sample  $M$
- Based on the  $M$ ,  $SD$ ,  $N$ , and critical  $\alpha$ , it is possible to calculate for a  $M$  or ES:
  - Lower-limit
  - Upper-limit

## Confidence Intervals

- Confidence intervals can be reported for:
  - Ms
  - Mean differences ( $M_2 - M_1$ )
  - ESs
- CIs can be examined statistically and graphically

## Confidence Intervals



## Confidence Intervals - Example

## CIs & Error Bar Graphs

### Example 1

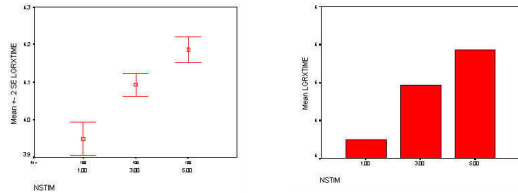
- $M = 5$ , with 95% CI of 2.5 to 7.5
- Reject  $H_0$  that the  $M$  is equal to 0.

### Example 2

- $M = 5$ , with 95% CI of  $-5$  to  $11.5$
- Accept  $H_0$  that the  $M$  is equal to 0.

- CIs around means can be presented as error bar graphs
- More informative alternatives to bar graphs or line graphs
- For representing the central tendency and distribution of continuous data for different groups

## CIs & Error Bar Graphs



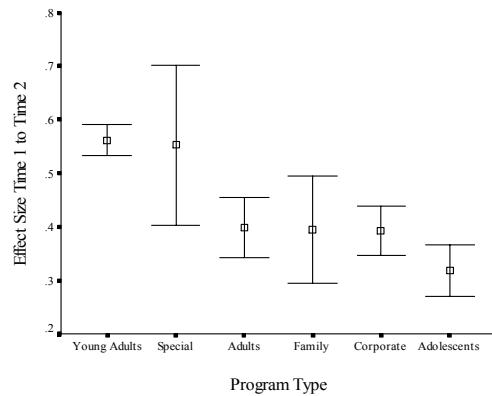
## Confidence Intervals

- In addition to getting CIs for *Ms*, we can obtain and should report CIs for *M* differences and for *ESs*.

## Confidence Interval of the Difference

Independent Samples Test

t-test for Equality of Means						
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
.764	489	.445	5.401E-02	7.067E-02	-8.48E-02	.1929
.778	355.220	.437	5.401E-02	6.944E-02	-8.26E-02	.1906



## Publication Bias, Scientific Integrity, & Cheating

## Two counter-acting biases

- Low Power:  
-> under-estimate of real effects
- Publication Bias or File-drawer effect:  
-> under-estimate of real effects

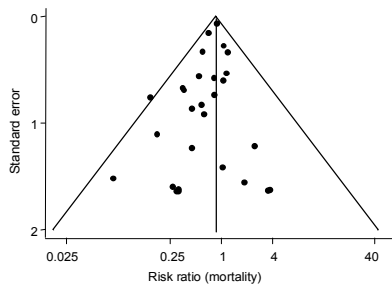
## Publication Bias

- Occurs when publication of results depends on their nature and direction.
- Studies that show a significant effect are more likely to be published.
- Type I publication errors are underestimated to the extent that they are: “frightening, even calling into question the scientific basis for much published literature.”  
(Greenwald, 1975, p. 15)

## Funnel Plots

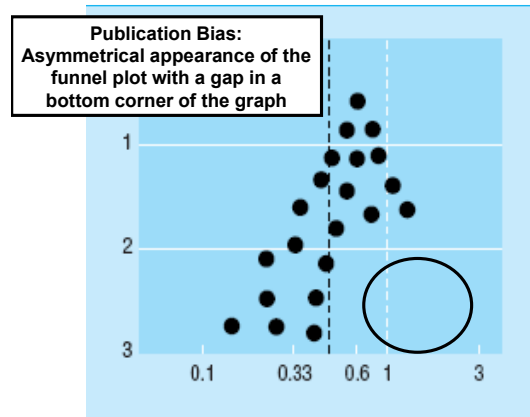
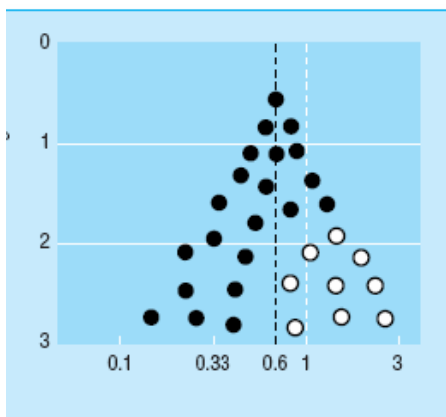
- A funnel plot is a scatter plot of treatment effect against a measure of study size.

## Funnel Plots



## Funnel Plots

- Precision in the estimation of the true treatment effect increases as  $N$  increases.
- Small studies scatter more widely at the bottom of the graph.
- In the absence of bias the plot should resemble a *symmetrical* inverted funnel.



## Publication Bias

- In this situation the effect calculated in a meta-analysis will overestimate the treatment effect
- The more pronounced the asymmetry, the more likely it is that the amount of bias will be substantial.

## File-drawer Effect

- Tendency for null results to be 'filed away' (hidden) and not published.
- No. of null studies which would have to 'filed away' in order for a body of significant published effects to be considered doubtful.

## Why Most Published Findings are False

Research results are less likely to be true:

1. The smaller the study
2. The smaller the effect sizes
3. The greater the number and the lesser the selection of tested relationships
4. The greater the flexibility in designs
5. The greater the financial and other interests
6. The hotter a scientific field (with more scientific teams involved)

## Academic Integrity: Students

- $N = 954$  students enrolled in 12 faculties of 4 Australian universities
- Self-reported:
  - Cheating (41%),
  - Plagiarism (81%)
  - Falsification (25%).

## Countering the Bias

---

### Journal of Articles in Support of the Null Hypothesis

---

[INDEX](#) [ABOUT](#) [MANUSCRIPT](#) [REVIEWER](#) [EDITORIAL](#) [LINKS](#) [CONTACT](#)  
[SUBMISSION](#) [SUBMISSION](#) [BOARD](#)

Welcome to the *Journal of Articles in Support of the Null Hypothesis*. In the past other journals and reviewers have exhibited a bias against articles that did not reject the null hypothesis. We seek to change that by offering an outlet for experiments that do not reach the traditional significance levels ( $p < .05$ ). Thus, reducing the file drawer problem, and reducing the bias in psychological literature. Without such a resource researchers could be wasting their time examining empirical questions that have already been examined. We collect these articles and provide them to the scientific community free of cost.

## Summary

- Counteracting biases in scientific publishing:
  - tendency towards low-power studies which underestimate effects
  - tendency to publish significant effects over non-significant effects
- Studies are more likely to draw false conclusions if conducted with small  $N$ ,  $ES$ , many effects, design flexibility, in hotter fields with greater financial interest
- Violations of academic integrity are prevalent, from students through researchers

## Recommendations

- Decide on  $H_0$  and  $H_1$  (1 or 2 tailed)
- Calculate power beforehand & adjust the design to detect a min. ES
- Report power, significance, ES and CIs
- Compare results with meta-analyses and/or meaningful benchmarks
- Take a balanced, critical approach, striving for objectivity and scientific integrity