



# Effect Size Reporting in Social- Personality Research: The Good, the Bad, and the Ugly

*Rob Cribbie, Emily Panzarella, Linda Farmus, Naomi  
Martinez Gutierrez, Nataly Beribisky, & Udi Alter*

Me doing this presentation is like David Ayres playing goal for the Carolina Hurricanes ... a great opportunity but not well earned 😊



# Introduction



- ➔ Although it was a long time coming, over the past few years there has been an abrupt shift from a focus on  $p$ -values and null hypothesis significance testing (NHST) to a focus on effect sizes (ESs) and meta-analytic thinking

The reign of the  $p$ -value is over

**The  $P$  value is dead**

EDITORIAL · 20 MARCH 2019

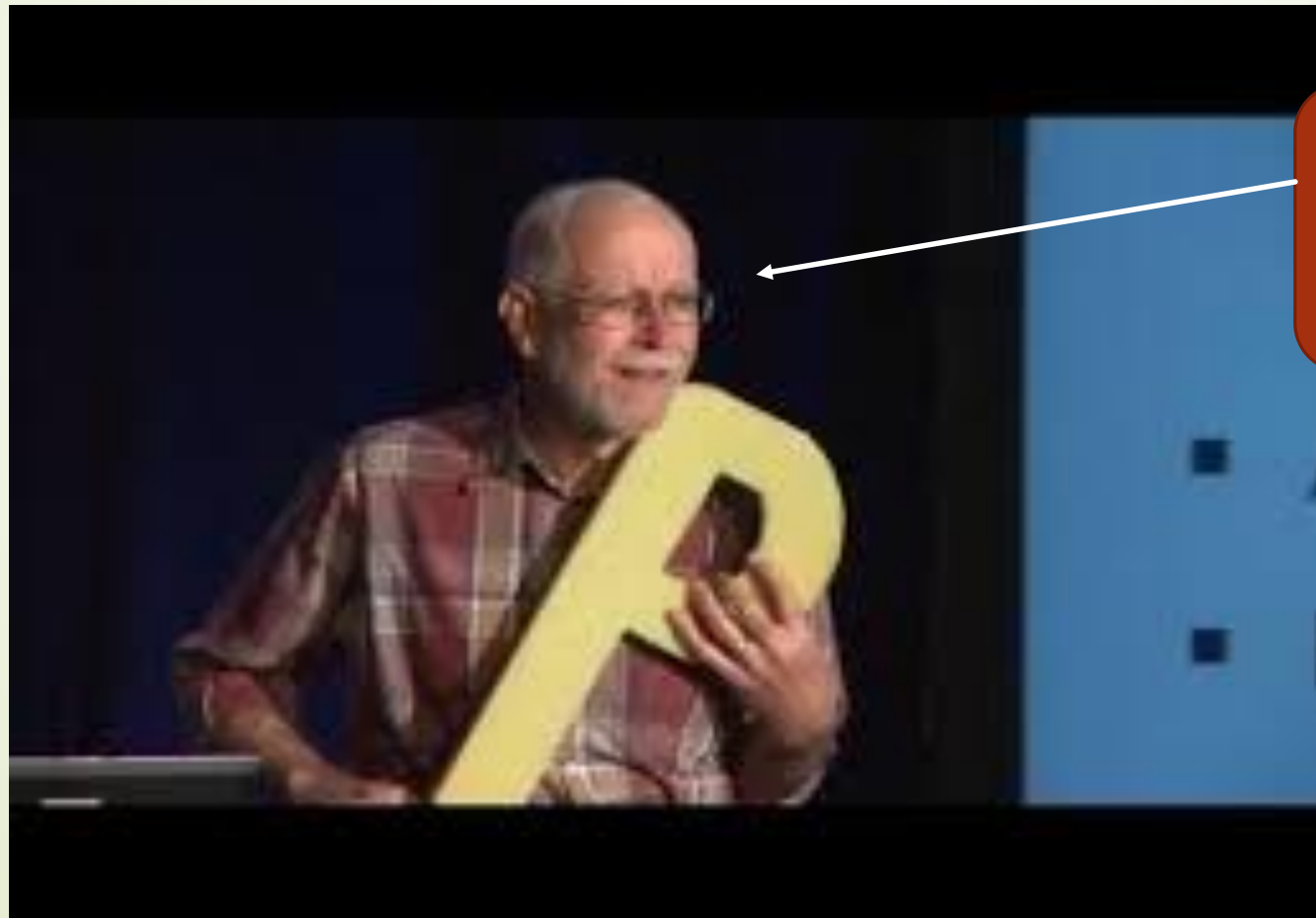
**It's time to talk about ditching statistical significance**

**Abandon Statistical Significance**

Blakeley B. McShane<sup>a</sup>, David Gal<sup>b</sup>, Andrew Gelman




However, methodologists (in favor of abandoning NHST) often forget how much love there is for  $p$ -values ...



Geoff Cumming  
sarcastically cuddling  
his  $p$ -value ...

Or, how comfortable researchers are with the use of  $p$ -values





# From $p$ -Values to Effect Sizes: Why Would the Shift be Difficult?

- $p$ -values come in one form
  - There are no standardized vs unstandardized  $p$ -values, there are no Hedges  $g$  corrections to  $p$ -values, etc.
- $p$ -values are compared against popular  $\alpha$  levels (.05)
  - Dichotomous decisions are easy and intuitive
- $p$ -values are reported routinely by statistical software
- $p$ -values are easy to interpret
  - Probabilities are one of the easiest statistics to interpret (at least superficially)

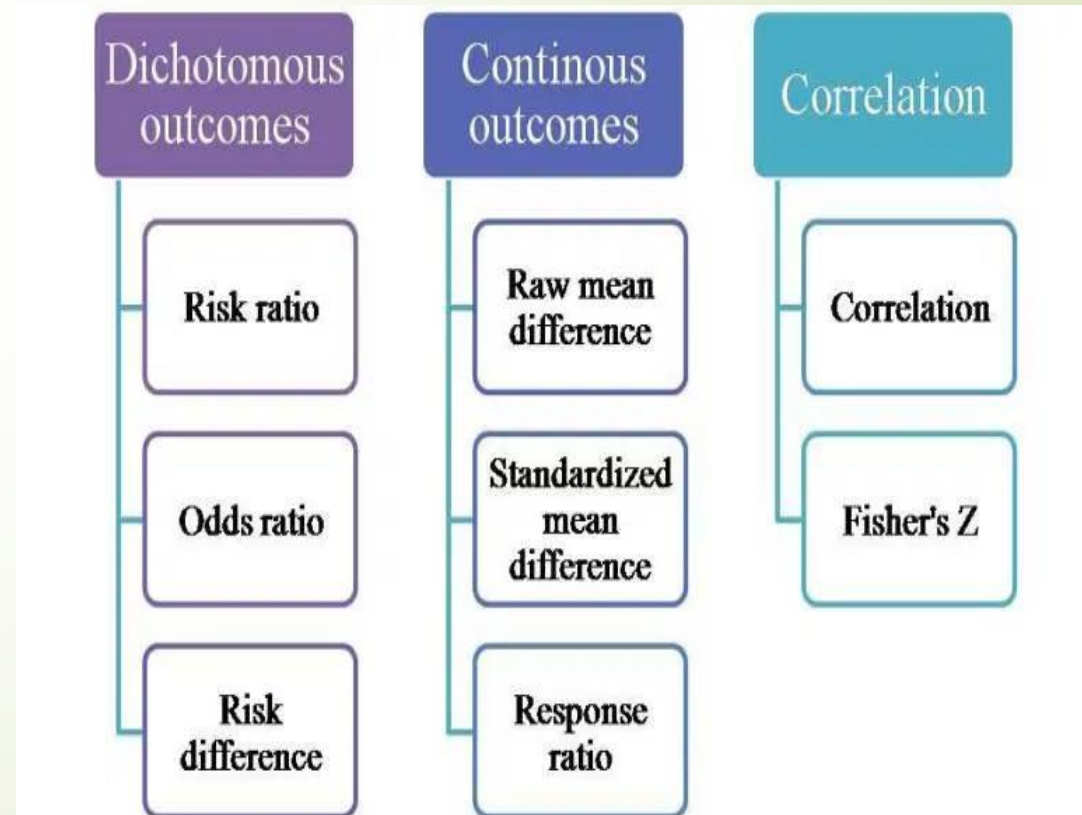
# Understanding Effect Sizes

- Relative to  $p$ -values, ESs come in many forms, are often more difficult to interpret, and are often difficult to find/produce in software

- Forms of Effect Size

- Unstandardized/standardized

- $b, \beta, M_1 - M_2, d, g, \Delta, \eta^2, \omega^2, r, r_P, r_{SP}, R^2, OR, RR, \text{etc.}$





# Understanding Effect Sizes

## ► Interpreting Effect Sizes

- *t*-shirt sizes
  - S, M, L, XL
  - Different cutoffs for every statistic
- Field specific ES magnitude interpretation based on the distribution of ESs
- Context-dependent interpretations
  - Recommended, but usually not straightforward

**Table 1**

*The Distribution of Correlation Coefficients Found Among Studies Included in Meta-Analytic Reviews, and Empirical Guidelines for Interpreting the Magnitude of Correlation Coefficients*

Distribution of correlation coefficients	Assessment review <sup>a</sup>	Treatment review <sup>b</sup>	Combined reviews <sup>c</sup>	Empirical guidelines <sup>d</sup>
Lower third	.02 to .21	-.08 to .17	-.08 to .17	< .20
Middle third	.21 to .33	.17 to .28	.18 to .29	20 to .30
Upper third	.35 to .78	.29 to .60	.30 to .78	> .30

## What does effect size tell you?

By [Saul McLeod](#), published 2019



# More factors contributing to confusion around effect sizes ...

- Most researchers completed their graduate studies before ESs became part of the curriculum
- Textbooks rarely include information regarding ESs and when they do it is limited in nature

Time to Overcome the Neglect of Effect Sizes in Teaching Psychological Research Findings



Johannes Hönekopp and Joanna Greer  
Northumbria University

**Misunderstandings and omissions in textbook accounts of effect sizes**

Paul H. Morris\* 

Department of Psychology, University of Portsmouth, UK

# QM researchers to the rescue ....

## Reporting Effect Sizes in Original Psychological Research: A Discussion and Tutorial

Jolynn Pek and David B. Flora  
York University

## It's the Effect Size, Stupid

What effect size is and why it is important

Robert Coe

of Education, University of Durham, email [r.j.coe@dur.ac.uk](mailto:r.j.coe@dur.ac.uk)

## Exploring perceptions of meaningfulness in visual representations of bivariate relationships

Nataly Beribisky, Heather Davidson

## How to Select, Calculate, and Interpret Effect Sizes

Joseph A. Durlak  
Loyola University Chicago



Okay, effect sizes are complicated, but do researchers really understand  $p$ -values?

- ▶ None of this assumes that substantive area researchers (or methodologists) can accurately interpret  $p$ -values, it only says that researchers are more comfortable with  $p$ -values and that the way in which researchers utilize and report  $p$ -values is relatively straightforward

---

## **A Dirty Dozen: Twelve $P$ -Value Misconceptions**

Steven Goodman

---

Why Are  $P$  Values Misinterpreted So Frequently?

# Blah, blah, blah, ... What's the point?

- ▶ The fact that  $p$ -values are more straightforward to adopt, use, and find than ESs means that there might be a drastic difference in the reporting practices and interpretation of  $p$ -values and effect sizes
  - ▶ Prior reviews have found that effect size reporting in different disciplines has varied anywhere from 1% to 87% (Sun, Pan, & Wang, 2010)

## Effect Size Estimates: Current Use, Calculations, and Interpretation

Catherine O. Fritz and Peter E. Morris  
Lancaster University

Jennifer J. Richler  
Vanderbilt University

### *Number (and Percentage) of Articles Reporting Effect Size Estimates Associated With ANOVA*

Year	Articles with ANOVA	Any ES measure	$\eta^2$	$\eta_p^2$
2009	27	18 (67)	1 (6)	17 (94)
2010	32	15 (47)	5 (33)	9 (60)
Overall	59	33 (56)	6 (18)	26 (79)





# Effect Size Reporting

- An ES measure should always be reported
- An ES can be reported in an unstandardized (units of the variables) or standardized (generic units) metric
- Confidence intervals should always accompany ESs
- More important to report ESs for specific tests than global tests (ES for an omnibus ANOVA? Why?)
- ESs should always be interpreted and should consider not generic cutoffs or the distribution of ESs in the discipline, but instead the magnitude of the ES within the context of the study

# Studies on Effect Size Reporting in Social-Personality Psychology



- At least we didn't find any ....



# Current Study: Effect Size Reporting in Social-Personality Research

- ▶ We examined ES reporting and interpreting practices within Social-Personality Psychology
- ▶ We reviewed high impact journals
  - ▶ *Journal of Personality and Social Psychology* (5.733)
  - ▶ *European Journal of Personality* (3.494)
  - ▶ *Journal of Experimental Social Psychology* (2.870)
  - ▶ *Journal of Research in Personality* (2.850)
  - ▶ *Personality and Social Psychology Bulletin* (2.498)
  - ▶ *Social Psychological and Personality Science* (2.633)
- ▶ All articles published in these journals in 2018 were reviewed except for those exclusively reporting qualitative research, simulation studies, scale validation, reviews, editorials, or journal announcements



# Current Study: Effect Size Reporting in Social-Personality Research

- ▶ We only coded information related to the primary hypothesis
- ▶ Coder Training
  - ▶ There were several steps involved in training coders for this study
    - ▶ Preliminary meeting to discuss the nature of the study and brainstorm the specifics of the review (topic area, important questions, etc.)
    - ▶ Meeting to narrow down specific journals and create rough coding sheet
    - ▶ Sample coding of 3 articles to flesh out issues with the coding/coding sheet and identify other important items for the coding sheet
    - ▶ Sample coding of 5 articles with the final coding sheet, with discussion of differences in coding
    - ▶ Sample coding of 10 articles by the undergraduate coders to calculate reliability
      - ▶ 97.5% agreement across the 10 (articles) x 36 (subjective items coded for each article) = 360 items



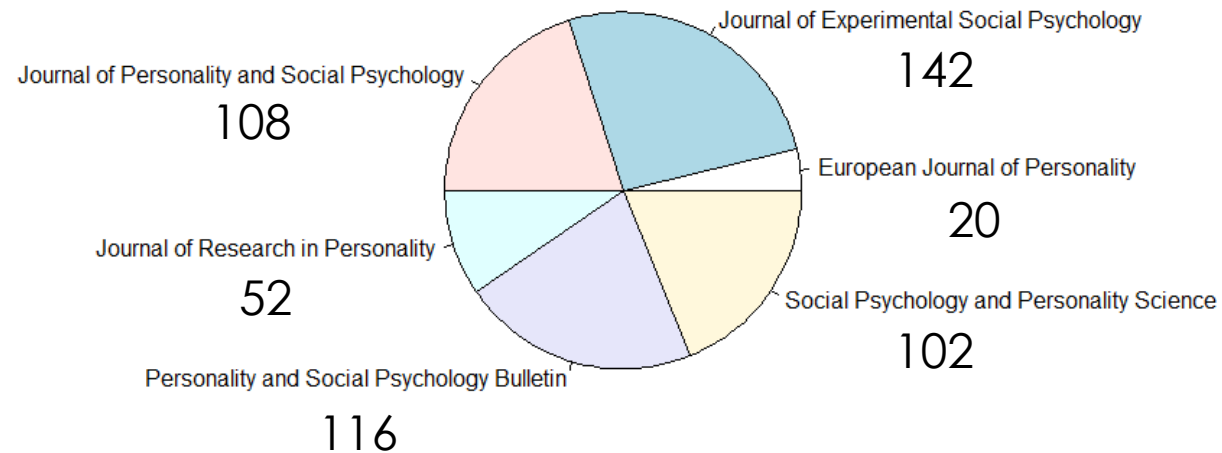


# Current Study: Research Questions

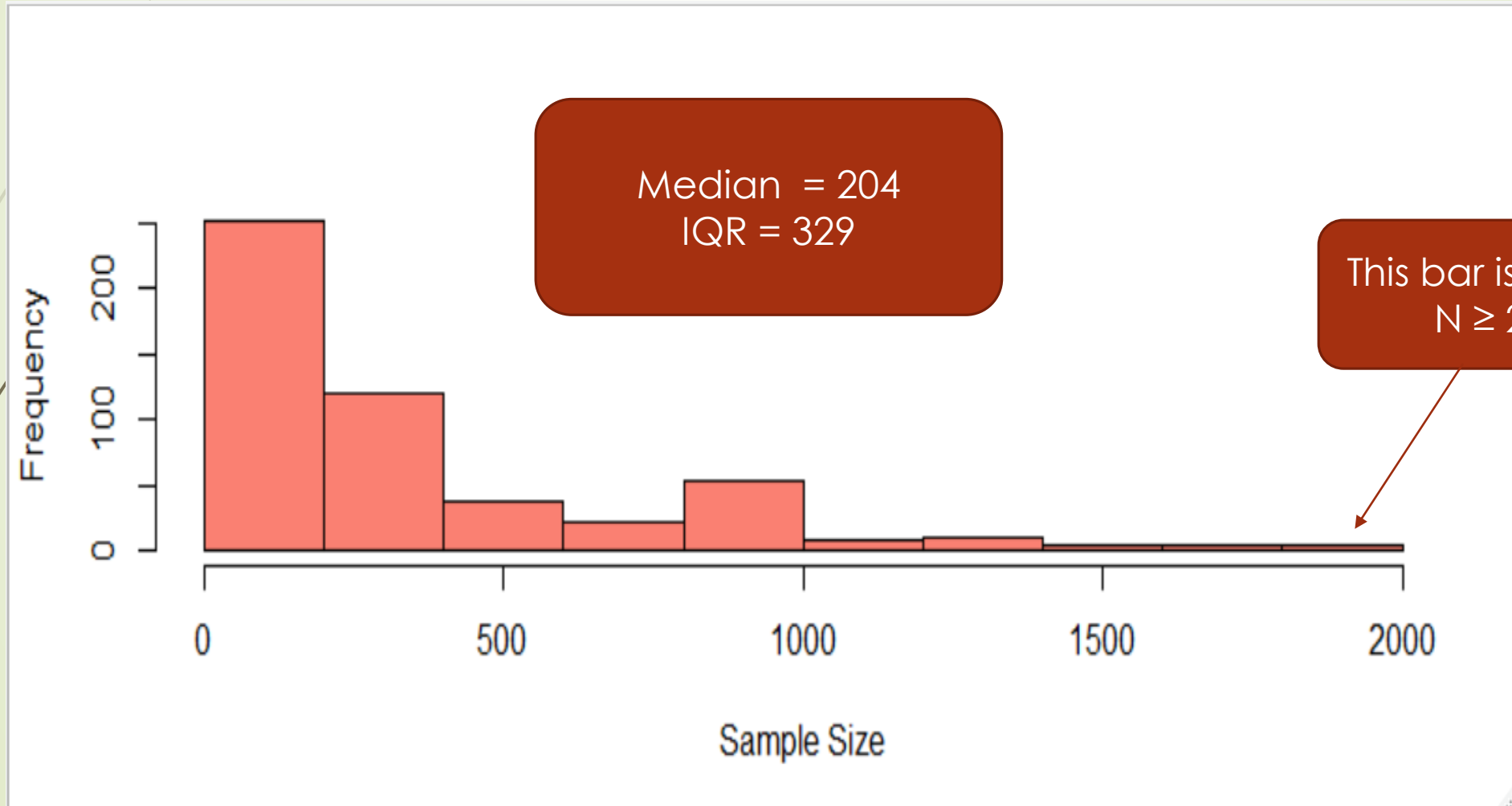
- 1) What proportion of articles provide and interpret an ES for their main hypothesis and related follow up tests?
- 2) What proportion of articles provide and interpret CIs for ESs for their main hypothesis and related follow-up tests?
- 3) Is the complexity of the statistical model used related to the reporting and interpretation of corresponding effect sizes?
- 4) Are standardized or unstandardized effect sizes reported most often?
- 5) Did researchers discuss the relationship between NHST and effect size results?

## Results: Number of Studies

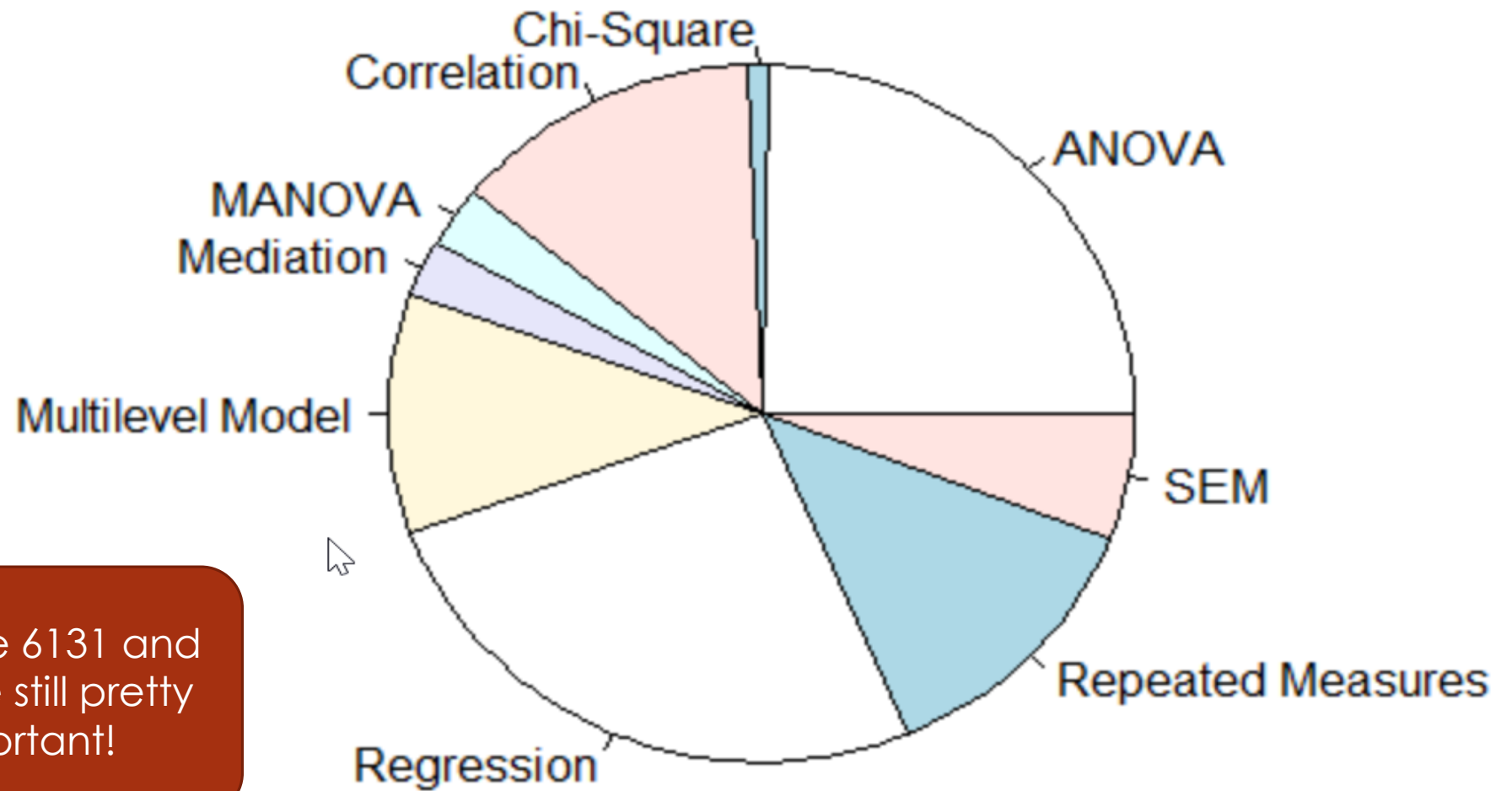
➔ 540 articles coded!!!



# Results: Sample Sizes within Articles



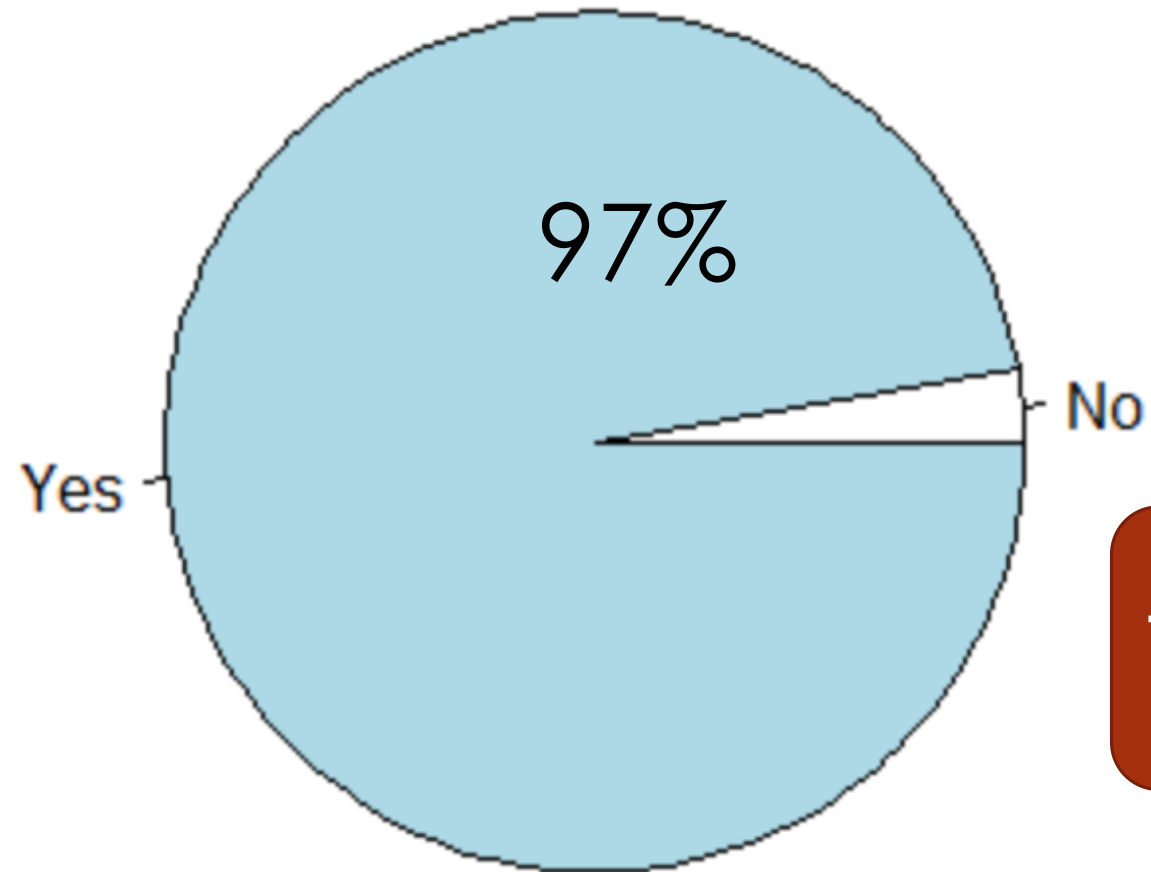
# Results: Type of Analysis



Looks like 6131 and 6132 are still pretty important!

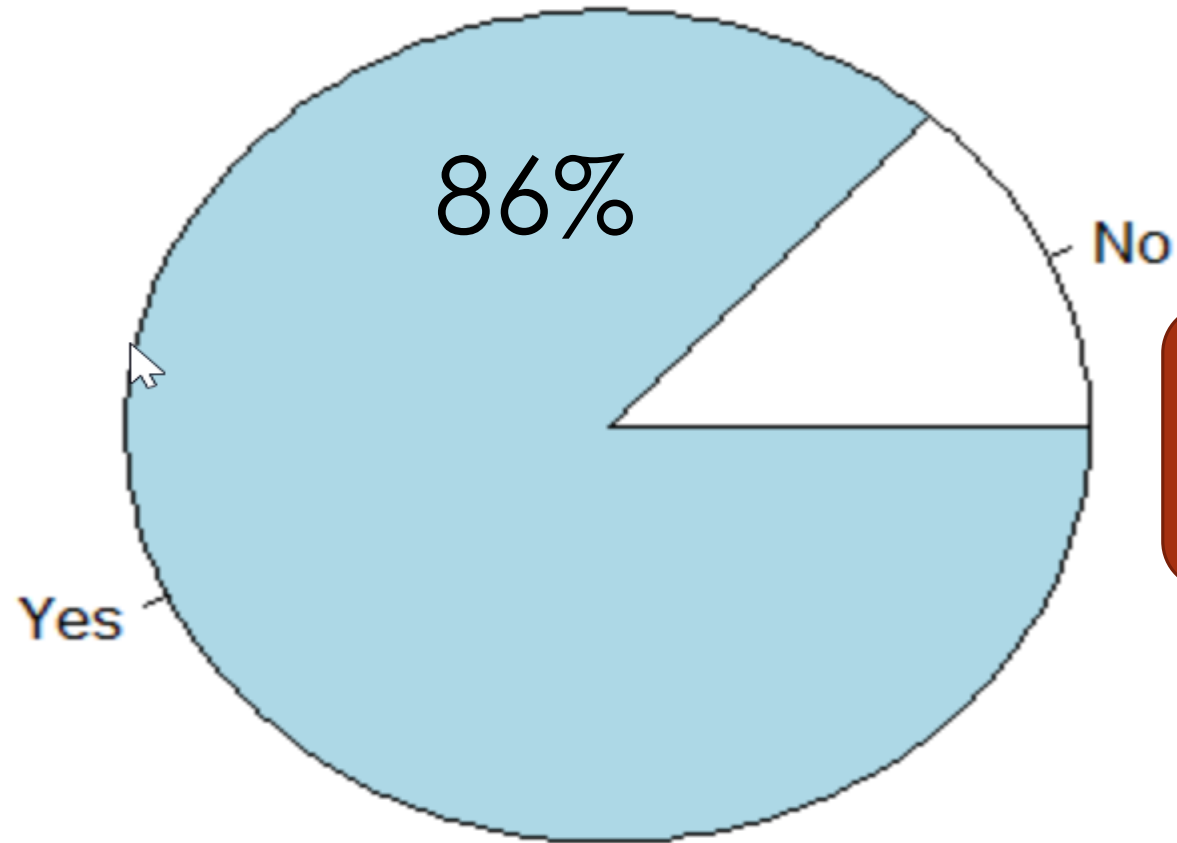


# Results: Was any ES reported for the Primary Hypothesis?



The Good!

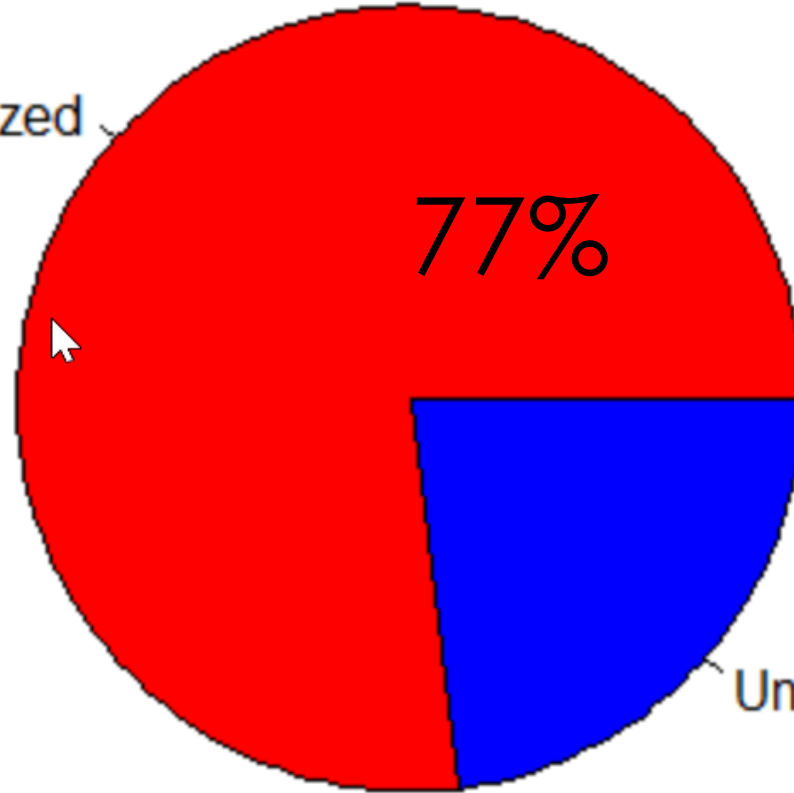
Results: Was any ES reported for follow-up tests (related to the primary hypothesis)?



The Good!

# Results: Standardized vs Unstandardized Effect Sizes

Standardized

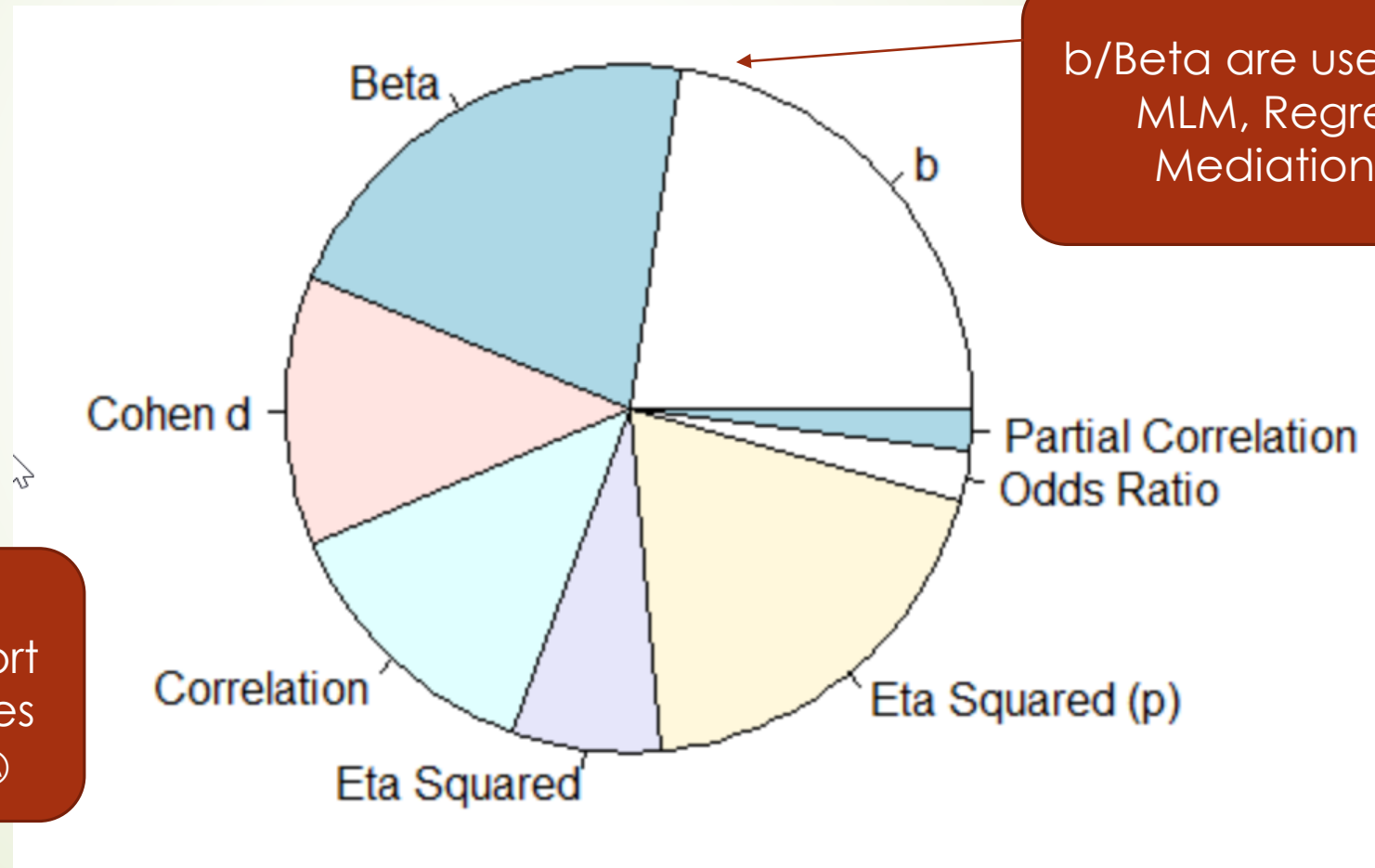


Unstandardized

Note: A table of means was not sufficient to be classified as reporting an “unstandardized effect size” ... the authors must discuss or report the mean difference

This is difficult to interpret ... e.g., some effect sizes have no unstandardized version (e.g.,  $r$ )

# Results: Type of ES Reported

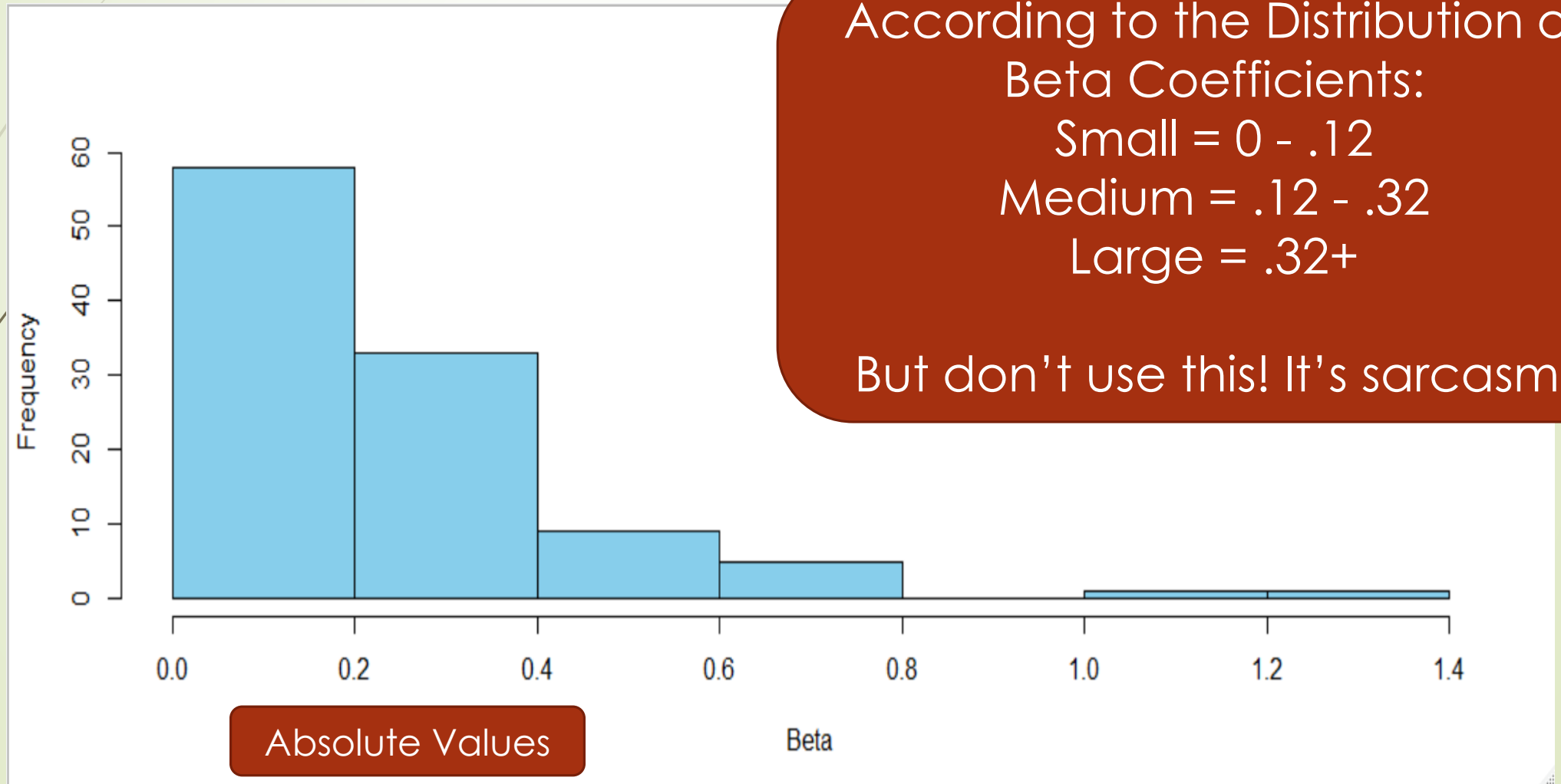


b/Beta are used in SEM, MLM, Regression, Mediation, etc.

Very few researchers report  $r_{sp}^2$  or Pratt indices for regression ☹️



# Results: Distribution of Beta Coefficients



According to the Distribution of Beta Coefficients:

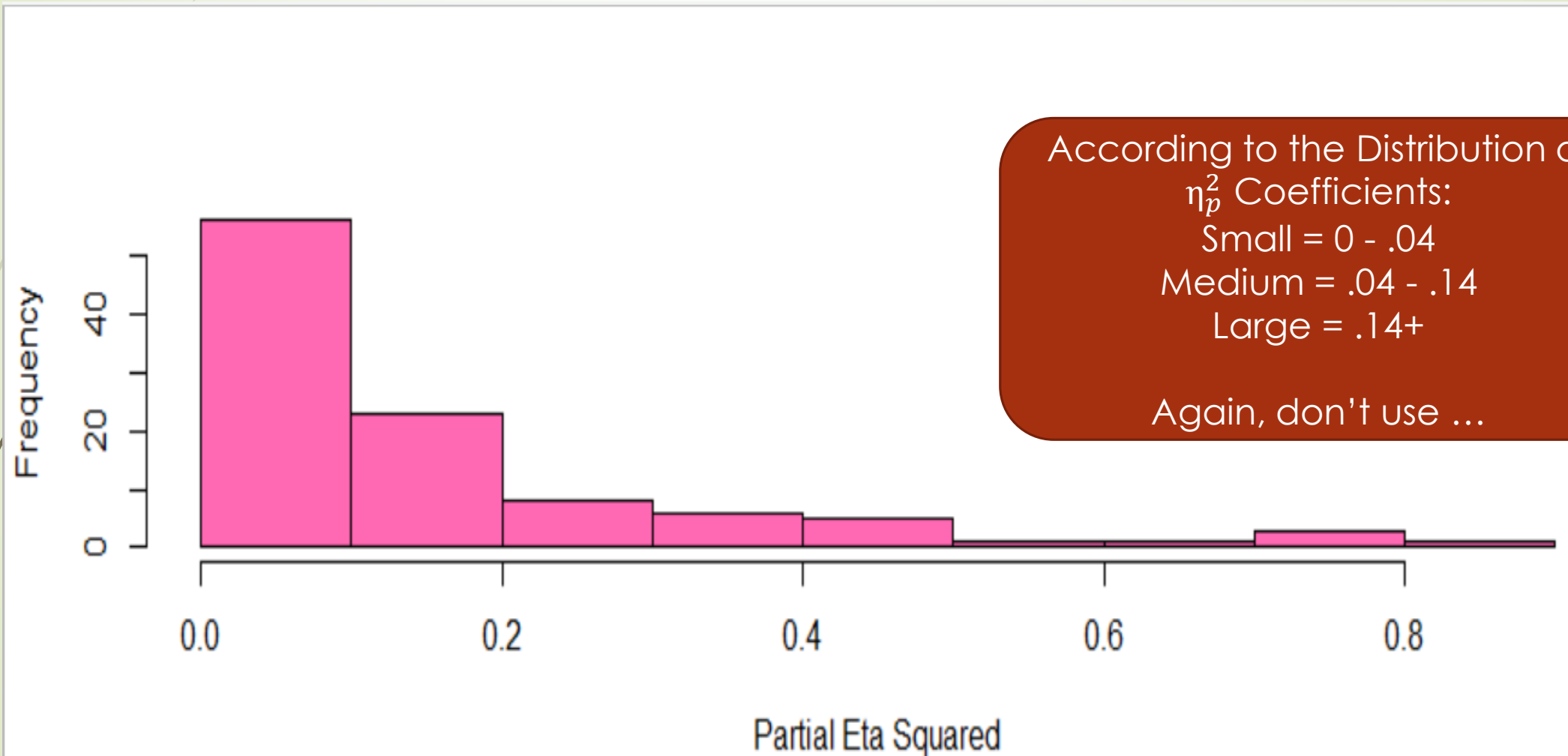
Small = 0 - .12

Medium = .12 - .32

Large = .32+

But don't use this! It's sarcasm!

# Results: Distribution of $\eta_p^2$ Coefficients



According to the Distribution of  $\eta_p^2$  Coefficients:

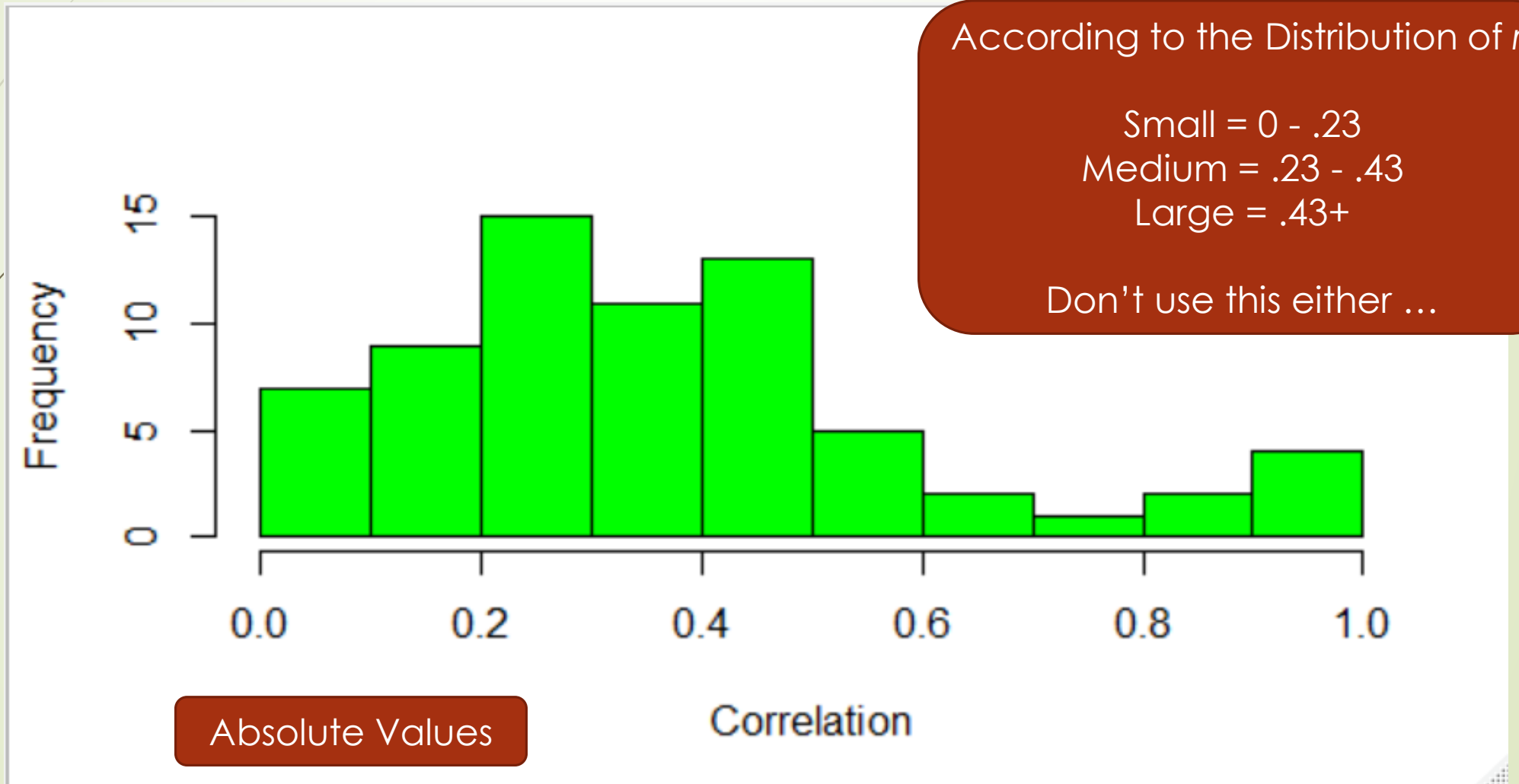
Small = 0 - .04

Medium = .04 - .14

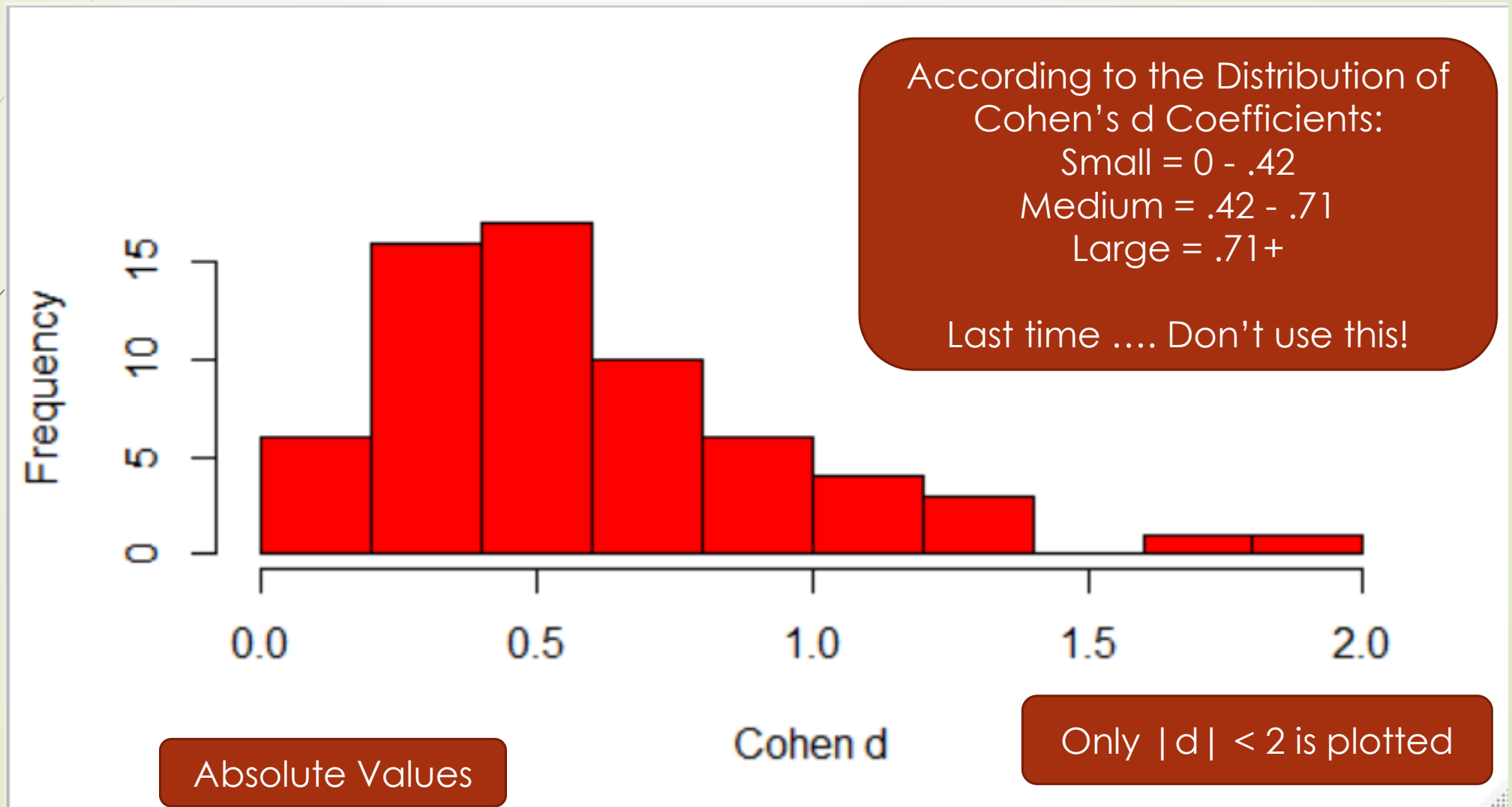
Large = .14+

Again, don't use ...

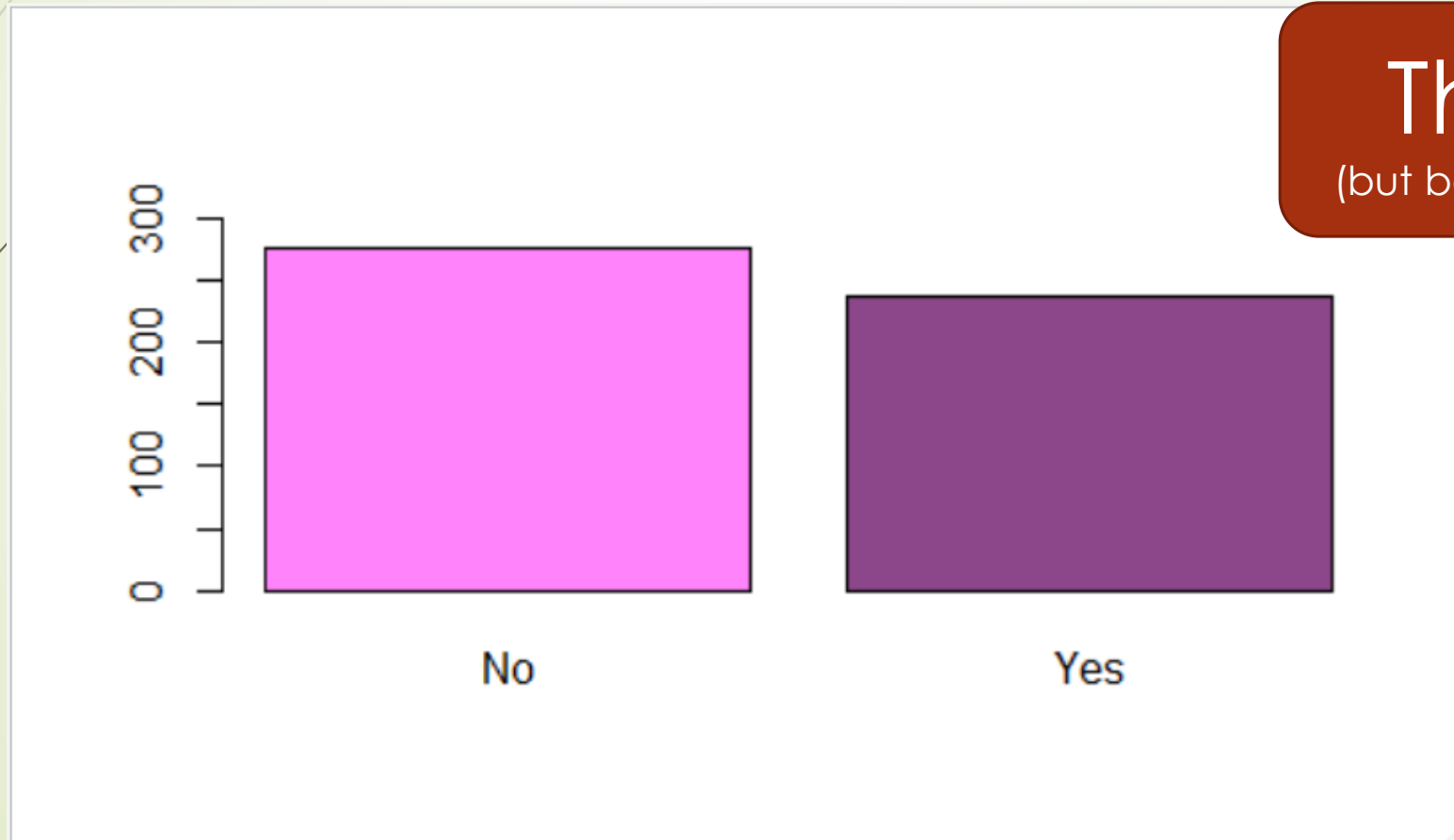
# Results: Distribution of Correlation Coefficients



# Results: Distribution of Cohen's $d$ Coefficients



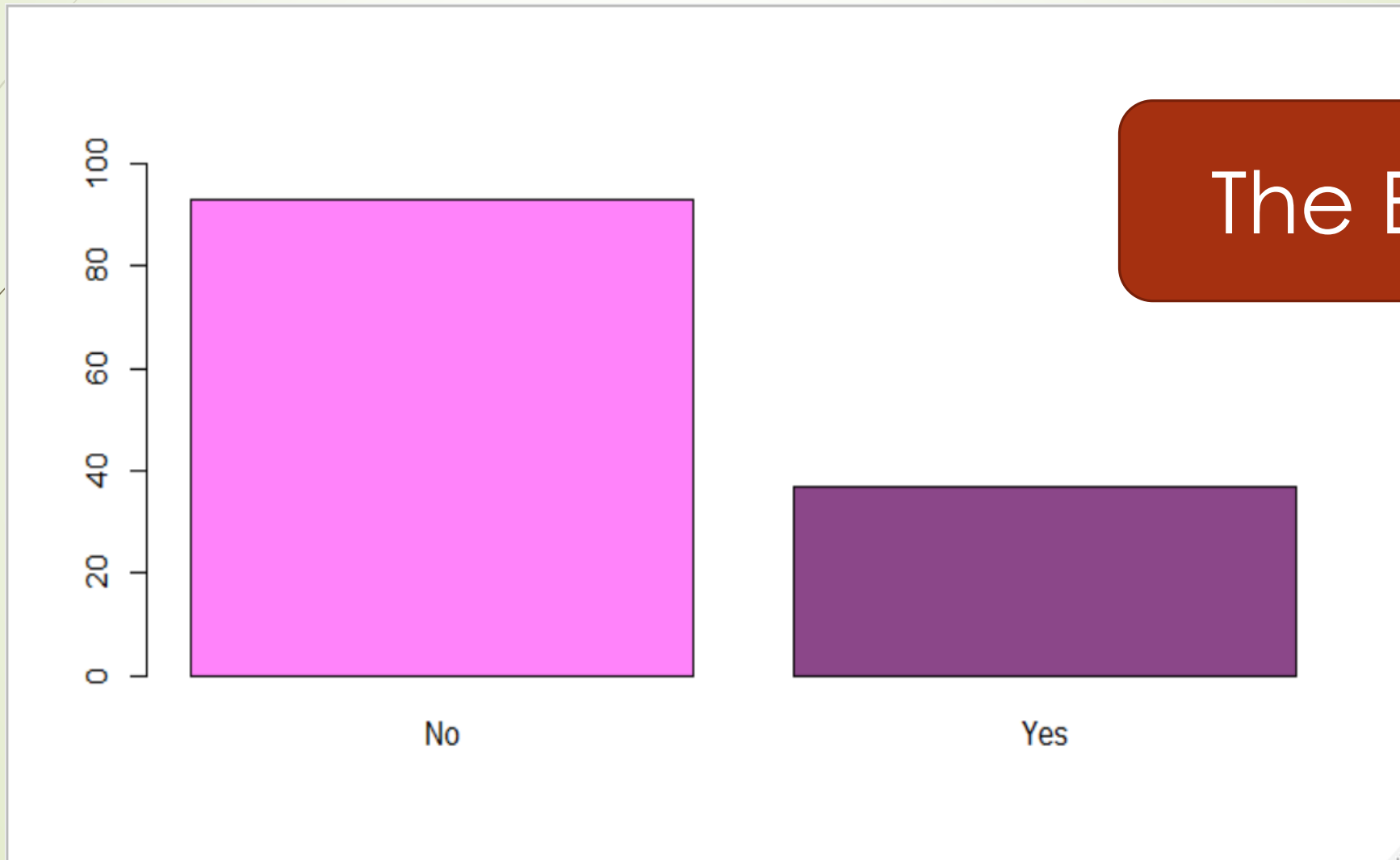
# Results: Any CI for ES Reported?



**The Bad!**  
(but better than expected)



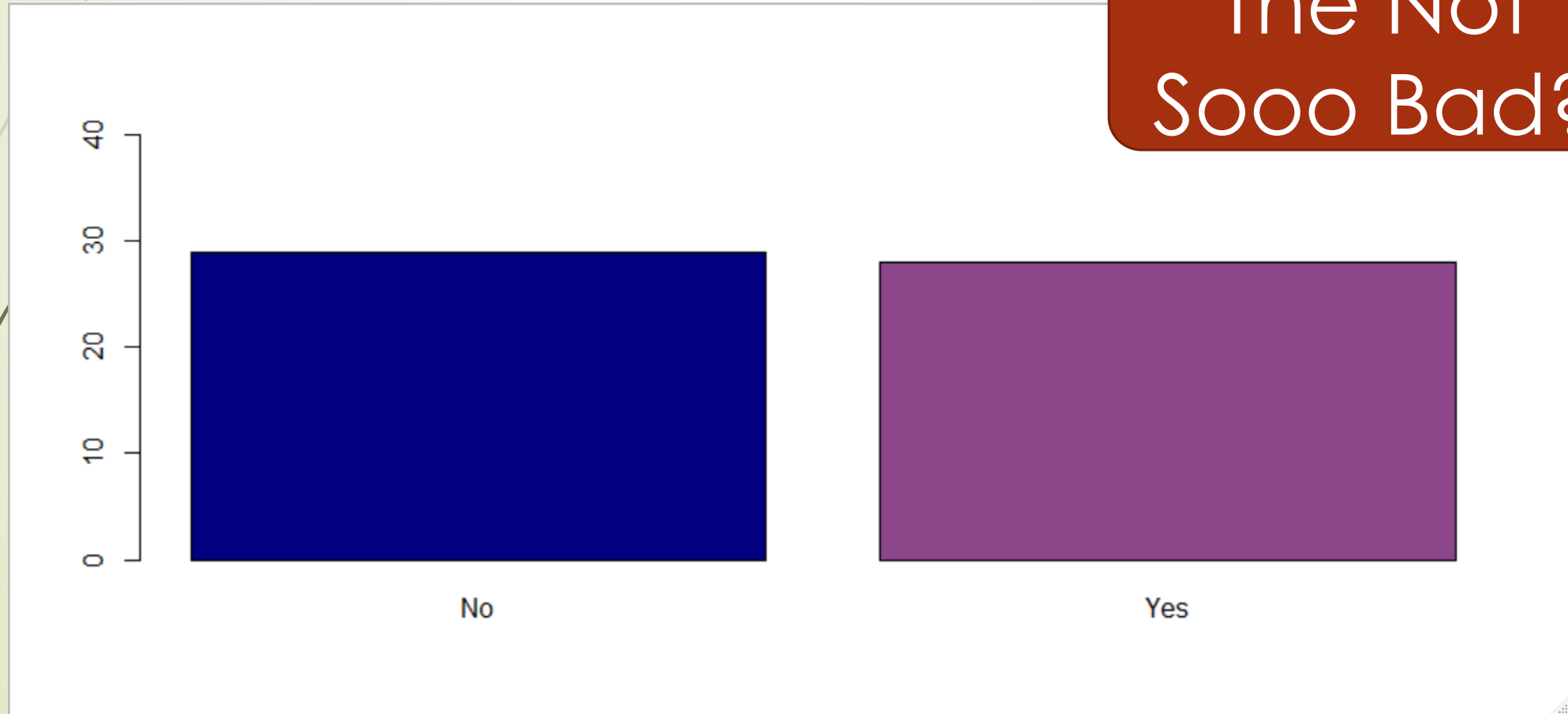
# Results: Any CI for $\eta^2/\eta_p^2$ Reported?



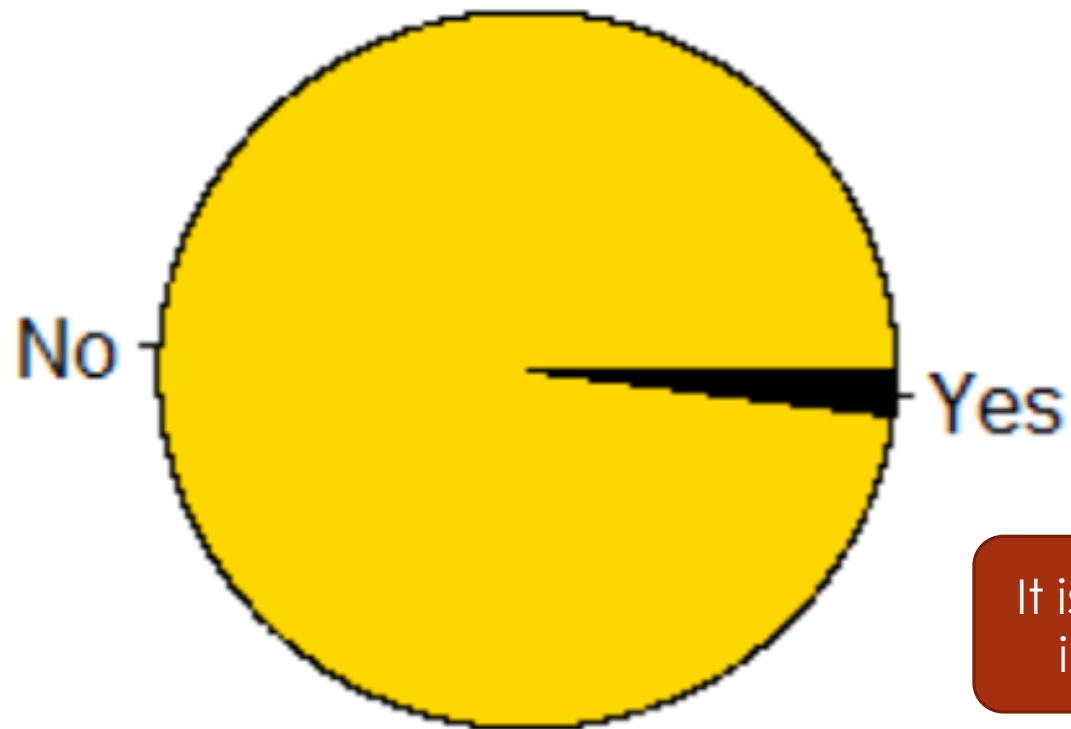
The Bad!

# Results: Any CI for Cohen's d Reported?

The Not  
Sooo Bad?



# Results: Any Interpretation of ES CI?

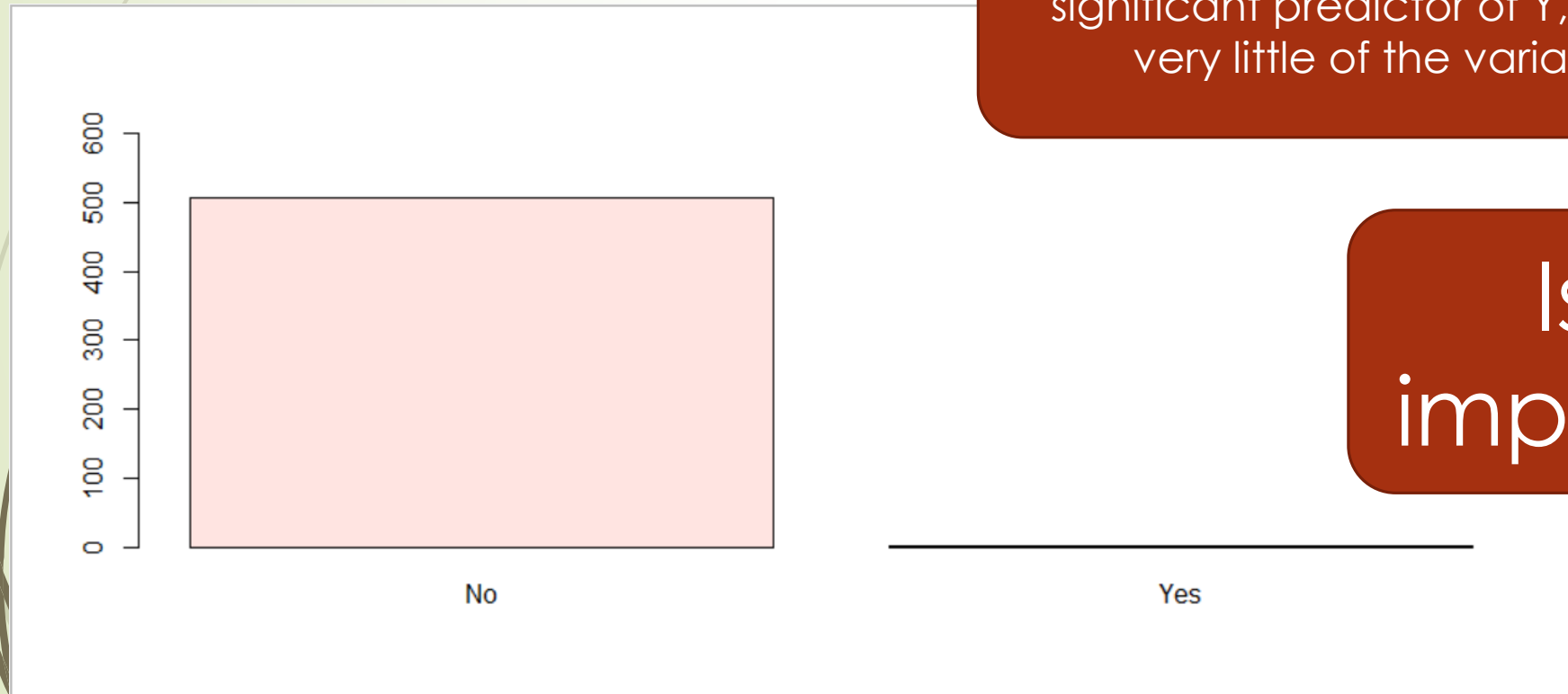


The Ugly!

It is important for researchers to interpret the width of the CI

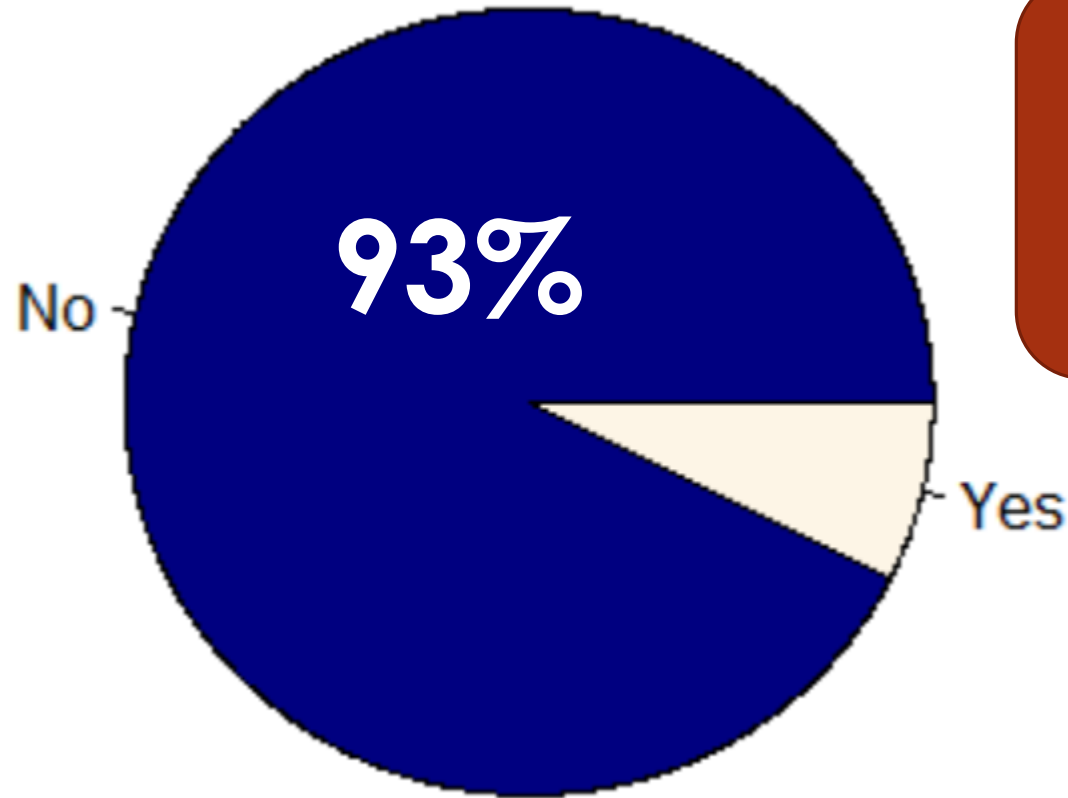
# Results: Any Discussion of the Relationship b/w the NHST and ES Results?

E.g., although X was a statistically significant predictor of Y, X explained very little of the variability in Y



Is this important?

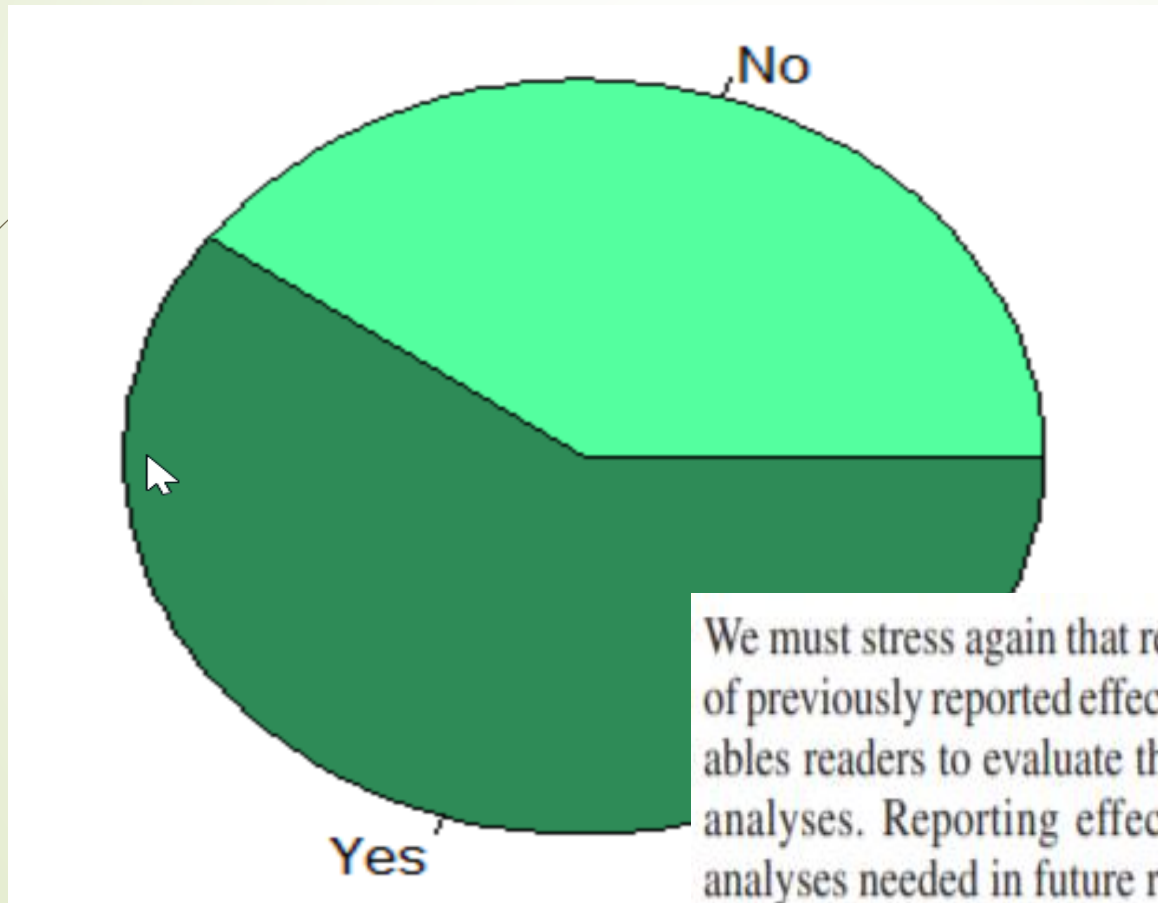
# Results: Any Interpretation of ES Magnitude via *t*-shirt Sizes?



This was very surprising!



# Results: Any Interpretation of ESs Within the Context of the Study?

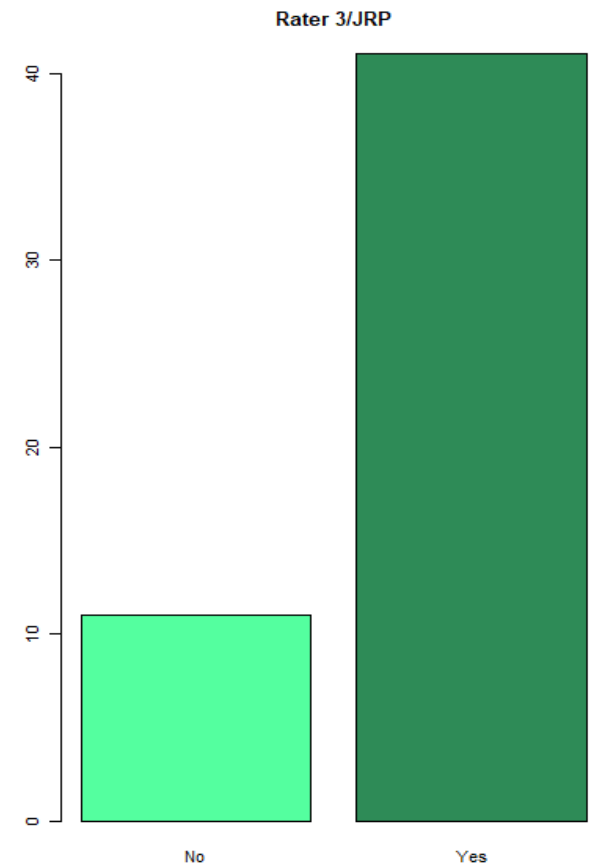
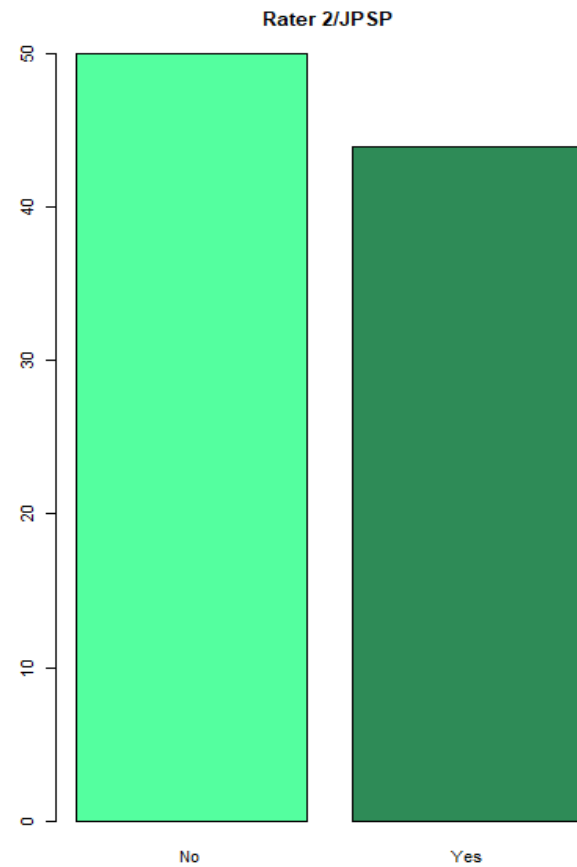
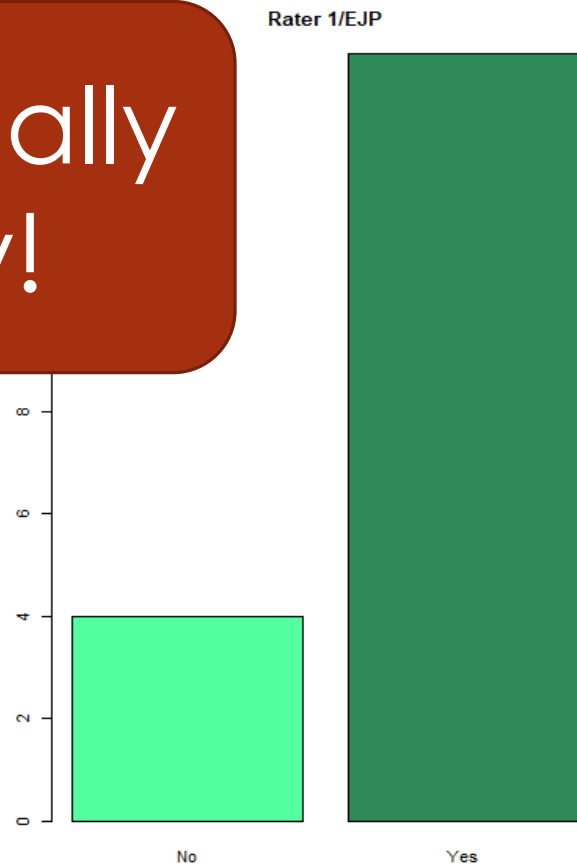


This seems pretty good, but ...

We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is *essential* [italics added] to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research. (Wilkinson & TFSL, 1999, p. 599)

Here are the results for a few journals/raters ...  
and this is only part of the issue


The Really Ugly!





# What is going on with interpretations via the context of the study?

- ▶ There are (at least) two reasons for the differences in the frequency of ‘in-context’ interpretations across journals
  - ▶ Journals/editors have different policies that result in very drastic differences in the frequency of in-context interpretations
  - ▶ Coders differ in terms of what constitutes an “in-context interpretation”



# Why would coders differ in terms of what constitutes an 'in-context' interpretation?

- Does the following description count as an 'in-context' interpretation?
- Example 1
  - *“Given the relationship between length of therapy and therapy outcome in depressed senior citizens, both in this study and past studies, it is important for therapists working with this population to continue therapy for at least 8 weeks.”*
- There are mentions of context (variables, population, psychological issue), however “the relationship” does not address the magnitude of the relationship

# Why would coders differ in terms of what constitutes an 'in-context' interpretation?

- ▶ Does the following description count as an 'in-context' interpretation?
- ▶ Example 2
  - ▶ *"The personality factor scale showing the strongest association with political orientation was Honesty-Humility ( $r = .21$ ); that is, lower Honesty-Humility was modestly associated with a more right wing political orientation. These results are consistent with recent studies showing the central role of this factor in the domains of ideology and values (e.g., Lee et al., 2010)."*
- ▶ There are mentions of context (variables, past studies, practical consequences) and magnitude, but are the authors linking the magnitude to the context (i.e., quantifying practical significance)



## Kelley & Preacher ... “On Effect Size”

- ▶ *“Translating the effect size along with the corresponding interval estimate into meaningful substantive terms is something that we see as a principal use of effect sizes. Some studies report effect sizes but interpret the results from only the perspective of a dichotomous reject or fail-to-reject outcome from a null hypothesis testing framework, perhaps with only an additional consideration of the direction of the effect size.”*
- ▶ Ok, so we need more than just a yes/no or directional interpretation



## What does Flora say ...

- ▶ *“At early stages of research, the direction instead of the magnitude of effect sizes is reasonably highlighted and interpreted; CIs of these effect sizes are likely to be wide, reflecting uncertainty in their estimation. At later stages of research, especially when consensus has been reached in terms of establishing the meaningfulness of a measure and replication is sought, the interpretation of effect sizes should focus on their magnitude and potential repercussions in terms of practical significance.”*
- ▶ Wait ... maybe it is not always necessary to link magnitude to practical significance ... but do coders have the time to classify a study as ‘early stage’ vs ‘later stage’?



# Lakens (2013) on interpreting Cohen's $d$ ...

- *“However, the best way to interpret Cohen's  $d$  is to relate it to other effects in the literature, and if possible, explain the practical consequences of the effect. **Regrettably, there are no clear recommendations of how to do so.**”*
- I think this well summarizes the current state of effect size interpretations (and the issues with coding effect size interpretations)

# Summary and Conclusions

- Effect size reporting within Psychology has increased substantially over the past decade or so
  - Further, effect size reporting for follow-up tests is also respectable
- Confidence interval reporting for ESs is improving, however interpretation of the intervals is still pretty much non-existent
- Researchers almost never link their NHST results to their effect size results
- Researchers rarely interpret effect sizes in terms of *t*-shirt sizes
- Coding whether an interpretation of ES magnitude is made “in context” is extremely difficult
  - The main issue is that there are few recommendations or good examples of how such an interpretation would be framed