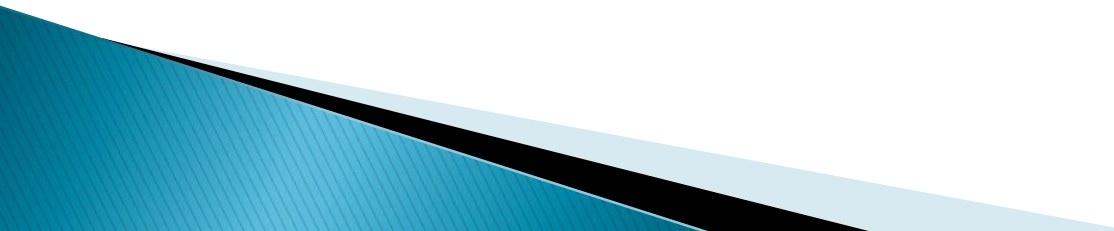# An Overview of Modern Data Analysis Strategies

Rob Cribbie
Quantitative Methods Program
Department of Psychology
York University

Part 2: Issues with Modern NHST

▸ It is well known that there are problems/controversies surrounding the use of NHST in psychology and related fields

▸ We will briefly discuss these issues in an attempt to motivate the alternatives that we will discuss over the next couple days of the course

Geoff Cumming and his "p"
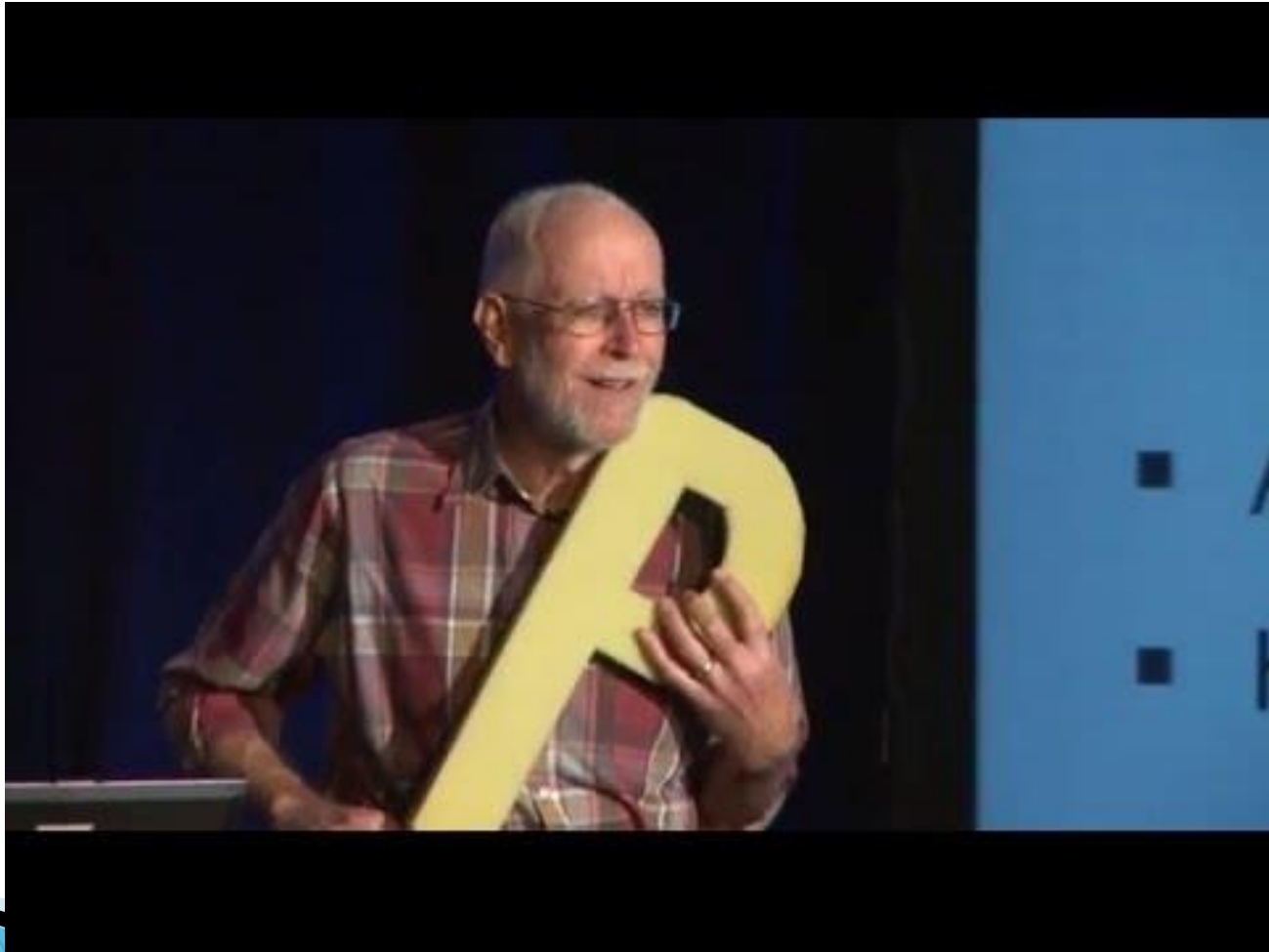
Issues with Modern NHST

- 1) Inverse Probability Error
  - NHST does not address the important question of researchers, namely: "What is the probability that H_0 is true, given the data collected?"
    - Instead, NHST answers "What is the probability of the data, assuming H_0 is true?"
      - Recall: A p–value represents the probability of obtaining a test statistic as extreme or more extreme than that found, assuming H_0 is true

  - Jacob Cohen discussed that many, many authors of methods texts (including himself) have made this error in their writing

Issues with Modern NHST

- 2) H_0 is always false
  - In almost all research settings, the proposed null hypothesis is false
    - Can you think of an effect with a true/population magnitude of exactly 0 (to many decimal places)?
      - i.e., is H_0: $\mu_1 = \mu_2$ ever true? Is H_0: $\rho = 0$ ever true?

  - Thus, with enough power we will always reject the null hypothesis, so why worry about null hypotheses and Type I errors?

Issues with Modern NHST

▸ 3) *p*-values are highly correlated with sample sizes
  ◦ As N increases, *p*-values decrease
    · Thus, NHST is invalid in small N studies (where *p*-values will generally be larger) and large N studies (where *p*-values will generally be smaller)
      · Practically significant effects can be declared not statistically significant with a small N OR practically insignificant effects can be declared not statistically significant with a large N
        · In other words, NHST is only valid with moderate levels of power, where neither β or 1-β are low
    · Most problematic is that most researchers are not aware of how strongly *p*-values and N are related
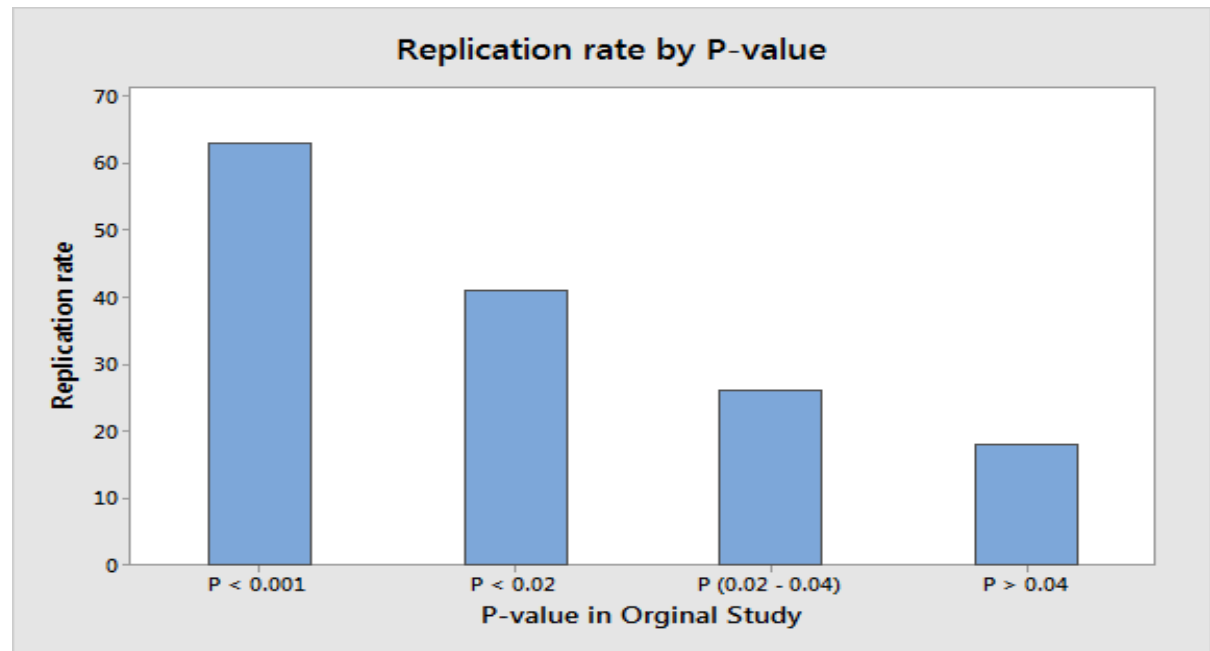
Issues with Modern NHST

- 4) NHST encourages researchers to make dichotomous (yes/no) decisions
  - For example, while adopting NHST (reject/do not reject H_o), researchers are encouraged to report only that there *was* or *was not* a relationship
    - Would we ever encourage a researcher to categorize a continuous variable?
  - Dichotomous decisions provide little information regarding the strength of the relationship
    - However, some argue that humans NEED to make dichotomous decisions
  - This criticism is more directly aimed at the Neyman-Pearson approach than the Fisher approach to NHST

Issues with Modern NHST

- 5) Researchers assume that p-values relate to the probability of successful replication
  - p-values are generally NOT a good metric for measuring replicability
  - Note that here replicability is defined in terms of statistical significance at both instances
    - This is NOT the best way to measure replicability, but relevant for this discussion
  - Assuming the null hypothesis is false, replication ability relates more to power than to p-values If two studies are conducted, each with power = .9, then the probability that both are statistically significant is $.9^2 = .81$
  - Confidence intervals also relate to replication; i.e., a 95% CI is able to capture a fairly large percentage of future means
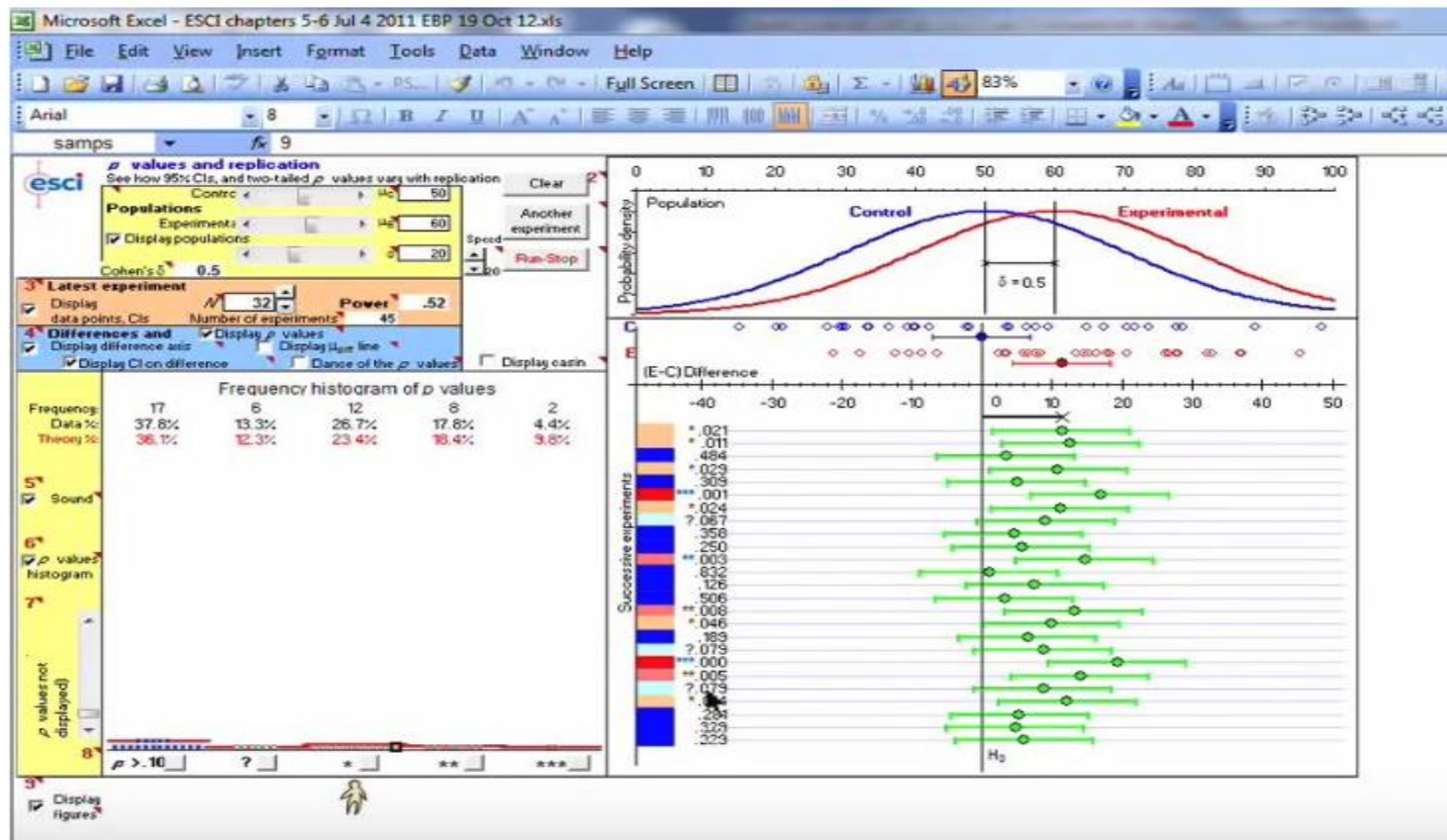
Issues with Modern NHST

▸ However, since *p*-values are related to sample size, and sample size relates to power, *p*-values are often found to relate to replicability

Replication rate by P-value

Issues with Modern NHST

- Cumming's *Dance of p-values*
  - https://www.youtube.com/watch?v=5OL1RqHrZQ8
  - *1:45–6:45*

Issues with Modern NHST

▸ 6) The nominal Type I error rate is often set at .05, regardless of the nature of the research
  ◦ Over time, the repeated adoption of α = .05 has led researchers to blindly use α = .05 without regard for how exploratory or confirmatory the study might be
  ◦ In fact, many researchers do not even specify their selected α level, presumably since it is implied that it is α = .05
    · Hard to imagine, when you consider how important the α level is to NHST

Issues with Modern NHST

- 7) Small p-values are thought to imply large effects or practical significance
  - If sample sizes are held constant, p-values correlate with effect sizes
  - However, as discussed earlier, small N studies can lead to large p-values even though the effect is practically meaningful
    - Conversely, large N studies can lead to small p-values even though the effect is not practically meaningful
  - It is important to not associate p-values with clinical significance, practical significance, etc.

Issues with Modern NHST

- 8) A non-significant effect does not allow us to conclude that 'H_0 is true'
  - How often do you read a study with a non-significant effect where the author states that therefore "the means are equal", "there is no difference in the means" or "there is no relationship among the variables"?
  - A student of mine explored this issue in clinical research comparing treatments
    - She explored 270 studies that compared treatments for various psychological issues
    - About half of the studies that found no statistically significant difference made claims related to equivalence (e.g., "same", "equal", "equally effective")

Issues with Modern NHST

▸ Providing evidence of a lack of relationship can be handled through two means:

◦ Negligible Effect/Equivalence Testing

　• An NHST procedure that essentially reverses the traditional NHST hypotheses

　　• E.g.,

　　　• H_0: $\mu_1 - \mu_2 \leq -\delta \mid \mu_1 - \mu_2 \geq \delta$

　　　• H_a: $-\delta \leq \mu_1 - \mu_2 \leq \delta$

◦ Bayesian Analysis

　• Bayesian approaches (e.g., Bayes Factors) allow us to quantify relative evidence for a hypothesis (e.g., H_0: $\mu_1 = \mu_2$)

　　• More on this to come

A Survey of Psychology Students and Faculty Regarding (Mis)Interpretations of p-values

- Haller & Krause (2002) surveyed three groups of individuals from psychology departments:
  - Students
  - Faculty not teaching statistics
  - Faculty teaching statistics
- All individuals were asked to respond True/False to a series of questions regarding the interpretation of p-values

The Survey Instructions

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is ($t = 2.7$, d.f. = 18, $p = 0.01$). Please mark each of the statements below as "true" or "false". "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

The Survey

1) You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).   [ ] true / false [ ]

2) You have found the probability of the null hypothesis being true.   [ ] true / false [ ]

3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).   [ ] true / false [ ]

4) You can deduce the probability of the experimental hypothesis being true.   [ ] true / false [ ]

5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.   [ ] true / false [ ]

6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

[ ] true / false [ ]

The Survey (all answers are false)

1) You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).   [ ] true / false [ ]

2) You have found the probability of the null hypothesis being true.   [ ] true / false [ ]

3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).   [ ] true / false [ ]

4) You can deduce the probability of the experimental hypothesis being true.   [ ] true / false [ ]

5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.   Toughest question (IMO), it is really tricky… we don't know if Ho is true   [ ] true / false [ ]

6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

[ ] true / false [ ]

The Survey Results

Figure 1: Percentages of participants in each group who made at least one mistake, in comparison to Oakes' original study (1986).
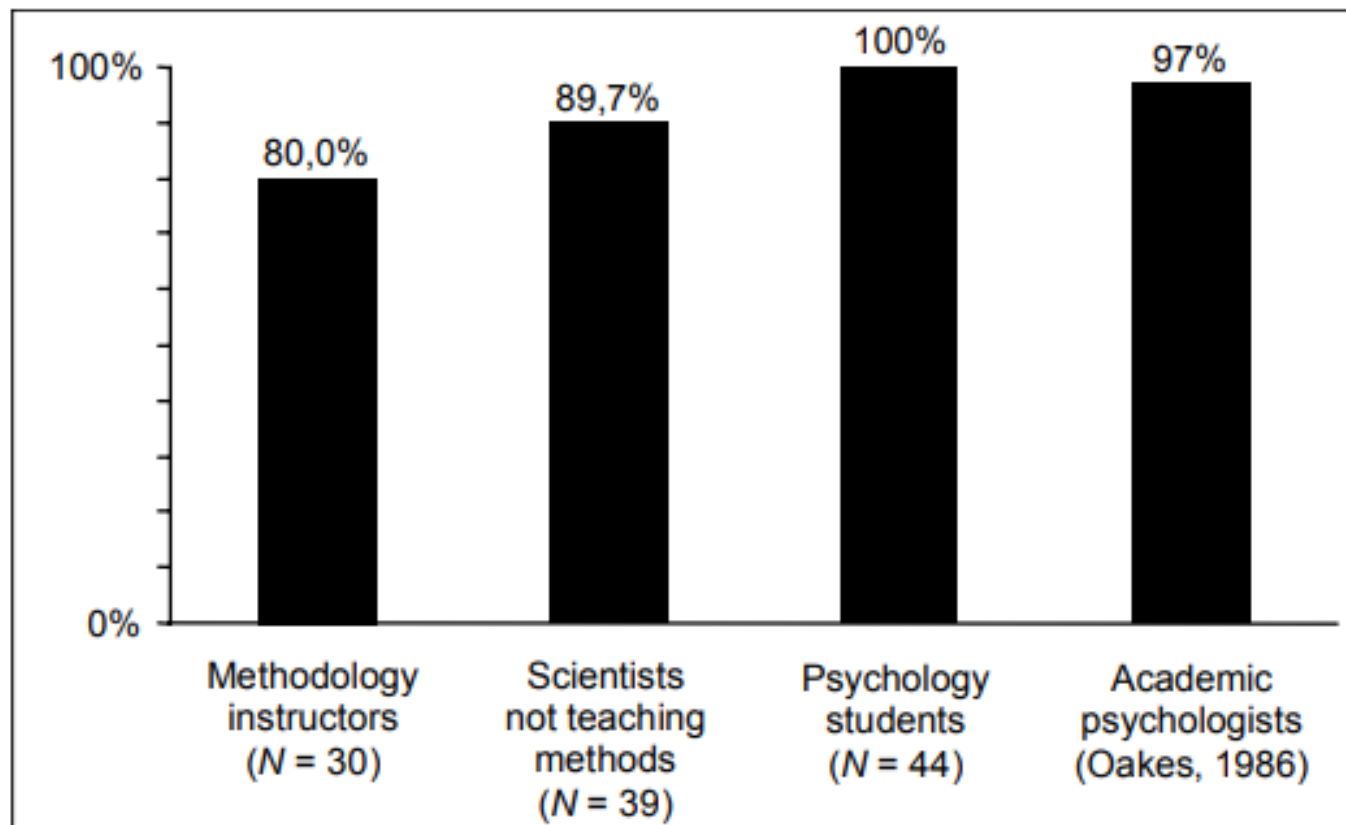
Failing Grade: 89% of Introduction to Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly

- Cassidy et al. (2019, *Advances in the Methods and Practices of Psychological Science*)
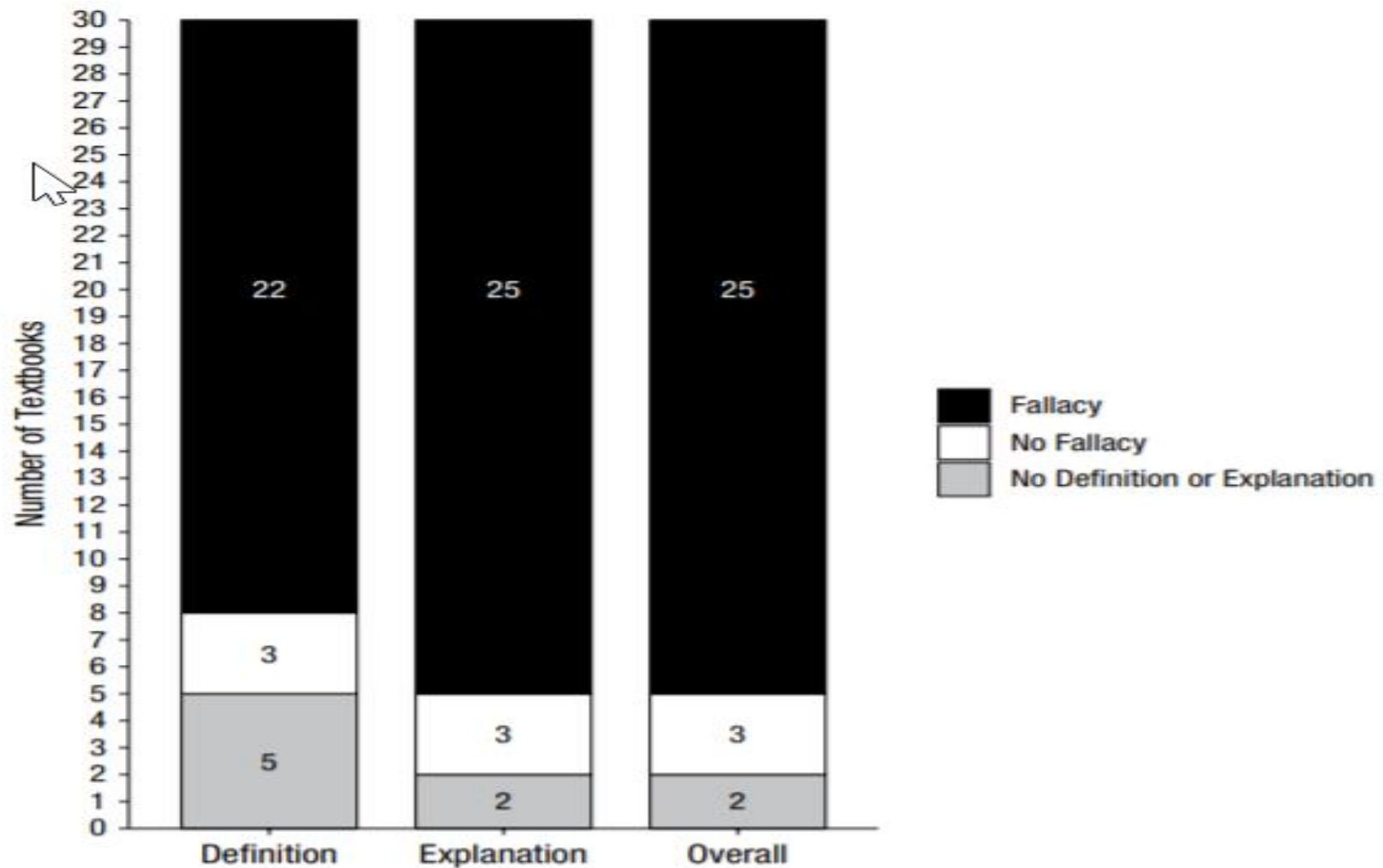
Results

Fig. 1. Frequency of definitions and explanations of statistical significance in the 30 textbooks. The "Definition" and "Explanation" bars show the number of textbooks without a definition or explanation, the number with a correct definition or explanation, and the number with a fallacious definition or explanation. The "Overall" bar indicates the number of textbooks lacking either a definition or explanation, the number with a definition or explanation (or both) with no fallacies, and the number with at least one fallacy.

Neo–Fisherian Solution to NHST Issues (Haig, 2017)

- Nominal Type I error rate is not specified (i.e., no α)
  ◦ *p*–value magnitude is of primary importance
- No significant/not significant distinction
- High *p*–values do not necessarily mean accepting the null
  ◦ Factors like N, effect magnitude, etc. play a role
- Test three hypotheses (null + each direction)
  ◦ Instead of just null + alternate
- Effect sizes (and CIs on effect sizes) are included as complementary information
- Clear distinction between statistical and substantive significance

Journal 'Basic and Applied Social Psychology' bans NHST

Editorial

David Trafimow and Michael Marks
New Mexico State University

The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

... the $p < .05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research. We hope and anticipate that banning the NHSTP will have the effect of increasing the quality of submitted manuscripts by liberating authors from the stultified structure of NHSTP thinking thereby eliminating an important obstacle to creative thinking.

Discussion Point

▸ Should academic journals in the behavioral sciences ban NHST?