Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane
Northwestern University

David Gal
University of Illinois at Chicago

The Null Hypothesis Significance Testing (NHST) Paradigm

- NHST is the dominant statistical paradigm in academic training and reporting in the biomedical and social sciences.

- Yet, it has been widely and long criticized by statisticians, psychologists, and other social scientists:
  - Misinterpretation of *p*-values.
  - The *p*-value is a poor measure of the evidence for or against a statistical hypothesis.
  - The dichotomization of results into "statistically significant" and "not statistically significant" has "no ontological basis."
    - "Surely, God loves the 0.06 nearly as much as the 0.05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*?"
  - The "nil hypothesis" of zero effect is "always false."
  - The $\alpha = 0.05$ threshold (or for that matter any other threshold) is arbitrary.
  - Statistical significance and practical importance are often confused.
  - ...

ASA Statement on Statistical Significance and *p*-values

P1. *p*-values can indicate how incompatible the data are with a specified statistical model.

P2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

P3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

P4. Proper inference requires full reporting and transparency.

P5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

P6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Does the Dichotomization of Evidence Intrinsic to the NHST Paradigm Cause Errors in Reasoning?

- Assigning things to different categories naturally leads to the conclusion that the things thusly assigned are categorically different.

- **Hypothesis**: The dichotomization of results into the different categories "statistically significant" and "not statistically significant" causes researchers to think of evidence in dichotomous terms:
  - Evidence that reaches the conventionally defined threshold of statistical significance (i.e., $p < 0.05$) is interpreted as a demonstration of a difference whereas evidence that fails to reach this threshold is disregarded.

- **Example** (Messori et al. 1993):
  - The result of our calculation was an odds ratio of 0.61 (95% CI 0.298 – 1.251; $p > 0.05$); this figure differs greatly from the value reported by Hommes *et al*. (odds ratio, 0.62; 95% CI, 0.39 – 0.98; $p < 0.05$)...we concluded that subcutaneous heparin is not more effective than intravenous heparin, exactly the opposite to that of Hommes and colleagues.

Does the Dichotomization of Evidence Intrinsic to the NHST Paradigm Cause Errors in Reasoning?

- Assigning things to different categories naturally leads to the conclusion that the things thusly assigned are categorically different.

- **Hypothesis**: The dichotomization of results into the different categories "statistically significant" and "not statistically significant" causes researchers to think of evidence in dichotomous terms:
  - Evidence that reaches the conventionally defined threshold of statistical significance (i.e., $p < 0.05$) is interpreted as a demonstration of a difference whereas evidence that fails to reach this threshold is disregarded.

- **Example** (Messori et al. 1993):
  - The result of our calculation was an odds ratio of **0.61** (95% CI 0.298 – 1.251; $p > 0.05$); this figure **differs greatly** from the value reported by Hommes *et al*. (odds ratio, **0.62**; 95% CI, 0.39 – 0.98; $p < 0.05$)...we concluded that subcutaneous heparin is not more effective than intravenous heparin, **exactly the opposite** to that of Hommes and colleagues.

Does the Dichotomization of Evidence Intrinsic to the NHST Paradigm Cause Errors in Reasoning?

- Assigning things to different categories naturally leads to the conclusion that the things thusly assigned are categorically different.

- **Hypothesis**: The dichotomization of results into the different categories "statistically significant" and "not statistically significant" causes researchers to think of evidence in dichotomous terms:
  - Evidence that reaches the conventionally defined threshold of statistical significance (i.e., $p < 0.05$) is interpreted as a demonstration of a difference whereas evidence that fails to reach this threshold is disregarded.

- **Example** (Messori et al. 1993):
  - The result of our calculation was an odds ratio of 0.61 (95% CI 0.298 – 1.251; ***p* > 0.05**); this figure **differs greatly** from the value reported by Hommes *et al*. (odds ratio, 0.62; 95% CI, 0.39 – 0.98; ***p* < 0.05**)...we concluded that subcutaneous heparin is not more effective than intravenous heparin, **exactly the opposite** to that of Hommes and colleagues.

Does the Dichotomization of Evidence Intrinsic to the NHST Paradigm Cause Errors in Reasoning?

- **More examples**:
  - Stoehr 1999 reports on Merilaita & Jormalainen 1997: They report in their results section that "substrate choice correlated significantly with food choice in **males (r =0.32; *p* < 0.05**; N=44) but not in **females (r = 0.32; NS**; N = 24)." Later in the paper, **the authors discuss the biological reasons why there is a relationship between these variables in males, but not in females**. The correlation coefficients, however, are identical; substrate choice and food choice had exactly the same relationship in males and females.
  - Greenland 2017 reports on Schmidt & Rothman 2014: In a sadly typical example, researchers claimed **their study conflicted with earlier** results because their estimated risk ratio (RR) was **1.20, 95% CI=(0.97,1.48)** versus a previously reported RR=**1.20, 95% CI=(1.09,1.33)**.
  - See also: Poole 1987; Rothman et al. 1993; Rothman et al. 2008; Greenland 2017.

Does the Dichotomization of Evidence Intrinsic to the NHST Paradigm Cause Errors in Reasoning?

- **More examples**:
  - Stoehr 1999 reports on Merilaita & Jormalainen 1997: They report in their results section that "substrate choice correlated significantly with food choice in **males (r =0.32; *p* < 0.05**; N=44) but not in **females (r = 0.32; NS**; N = 24)." Later in the paper, **the authors discuss the biological reasons why there is a relationship between these variables in males, but not in females**. The correlation coefficients, however, are identical; substrate choice and food choice had exactly the same relationship in males and females.
  - Greenland 2017 reports on Schmidt & Rothman 2014: In a sadly typical example, researchers claimed **their study conflicted with earlier** results because their estimated risk ratio (RR) was **1.20, 95% CI=(0.97,1.48)** versus a previously reported RR=**1.20, 95% CI=(1.09,1.33)**.
  - See also: Poole 1987; Rothman et al. 1993; Rothman et al. 2008; Greenland 2017.
- Do even expert statisticians think dichotomously and thus make similar errors?

Study 1: Descriptive Statements

- Below is a summary of a study from an academic paper:
- The study aimed to test how different interventions might affect terminal cancer patients' survival. Participants were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Participants were then tracked until all had died. Participants in Group A lived, on average, 8.2 months post-diagnosis whereas participants in Group B lived, on average, 7.5 months post-diagnosis ($p = 0.27$).
- Which statement is the most accurate summary of the results?

  A. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **greater** than that lived by the participants who were in Group B.

  B. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **less** than that lived by the participants who were in Group B.

  C. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **no different** than that lived by the participants who were in Group B.

  D. Speaking only of the subjects who took part in this particular study, it **cannot be determined** whether the average number of post-diagnosis months lived by the participants who were in Group A was greater/no different/less than that lived by the participants who were in Group B.

Study 1: Descriptive Statements

- Below is a summary of a study from an academic paper:
- The study aimed to test how different interventions might affect terminal cancer patients' survival. Participants were randomly assigned to one of two groups. Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Participants were then tracked until all had died. Participants in Group A lived, on average, 8.2 months post-diagnosis whereas participants in Group B lived, on average, 7.5 months post-diagnosis (p = 0.01 / 0.27).
- Which statement is the most accurate summary of the results?

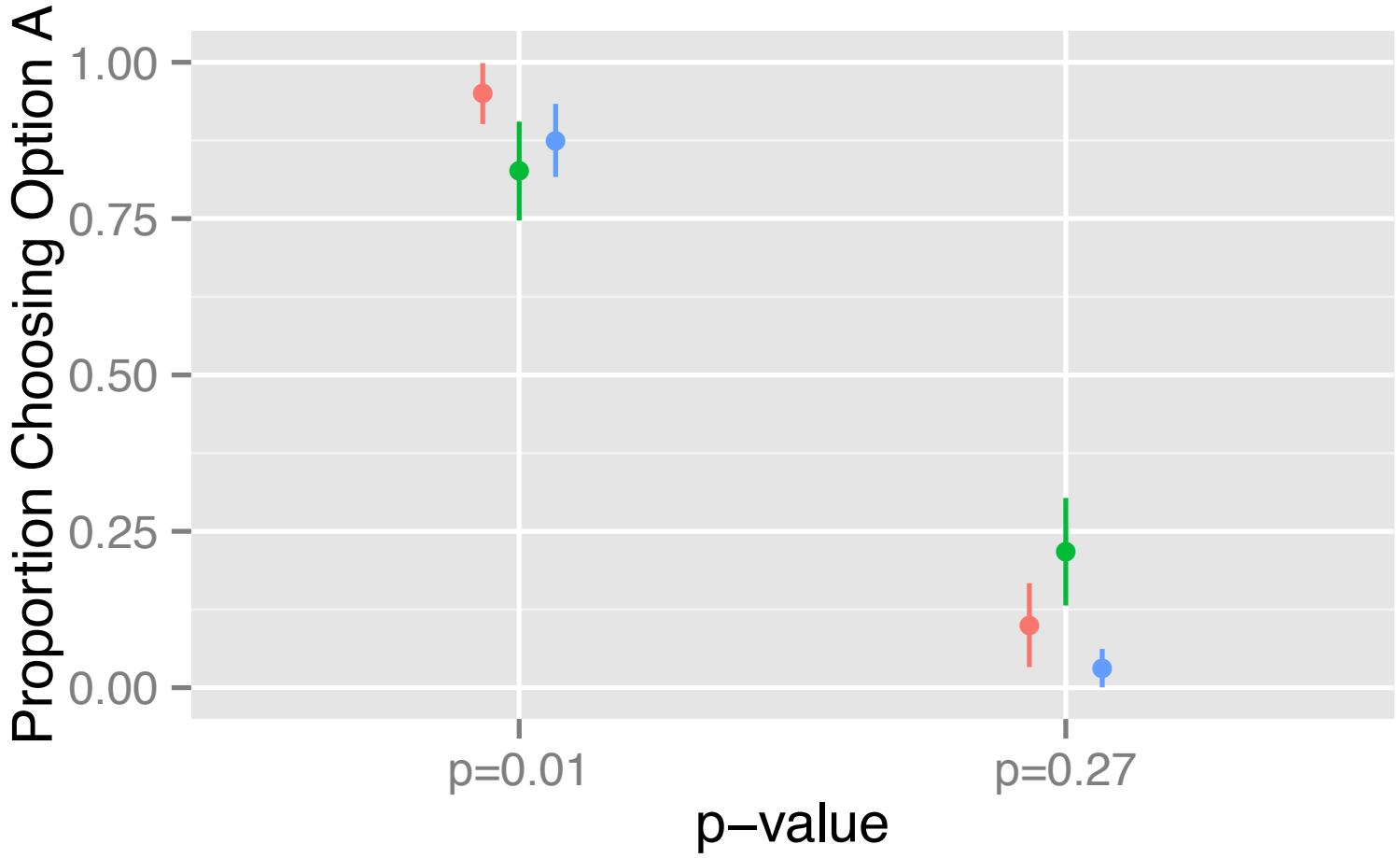
  A. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **greater** than that lived by the participants who were in Group B.

  B. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **less** than that lived by the participants who were in Group B.

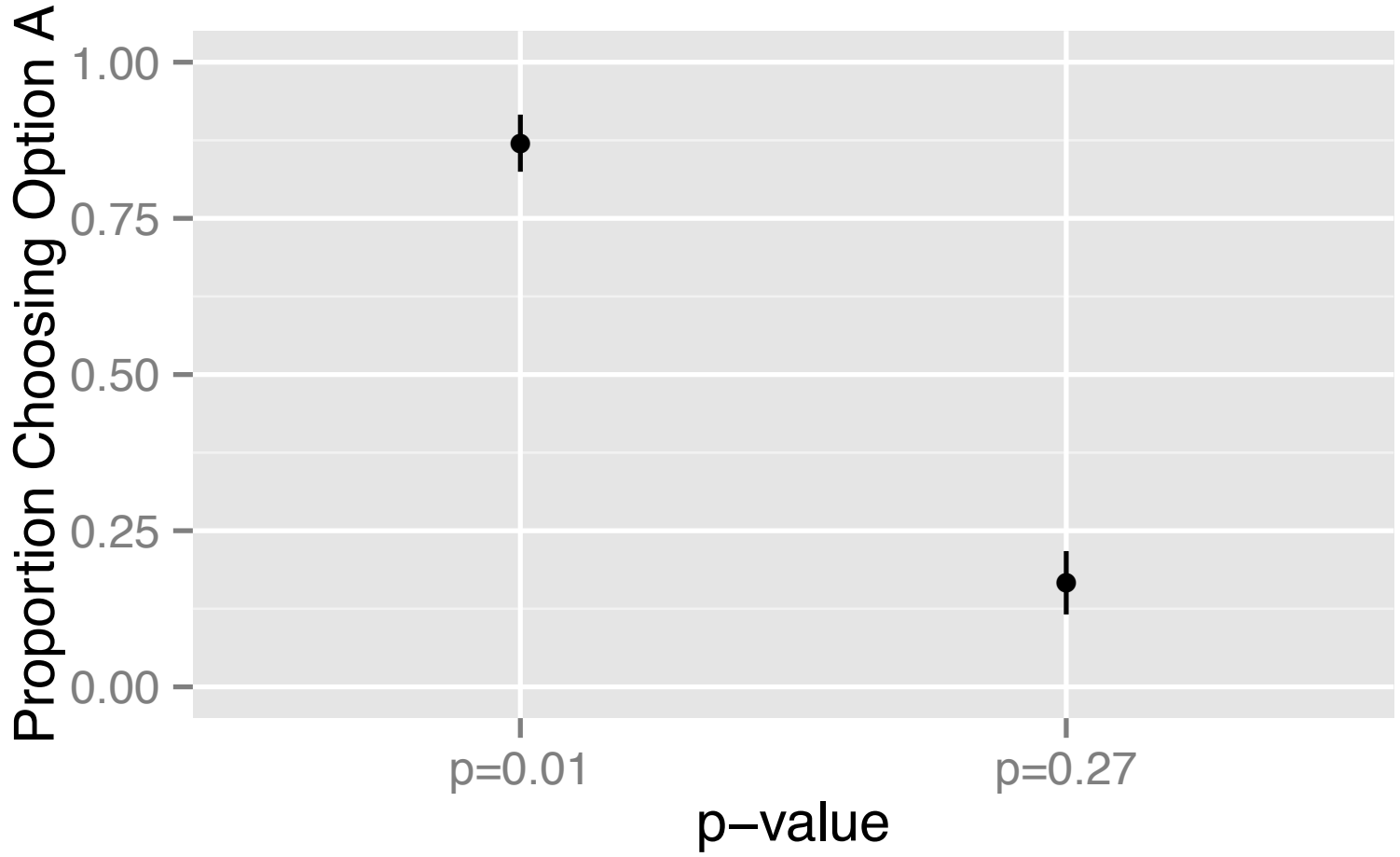
  C. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **no different** than that lived by the participants who were in Group B.

  D. Speaking only of the subjects who took part in this particular study, it **cannot be determined** whether the average number of post-diagnosis months lived by the participants who were in Group A was greater/no different/less than that lived by the participants who were in Group B.

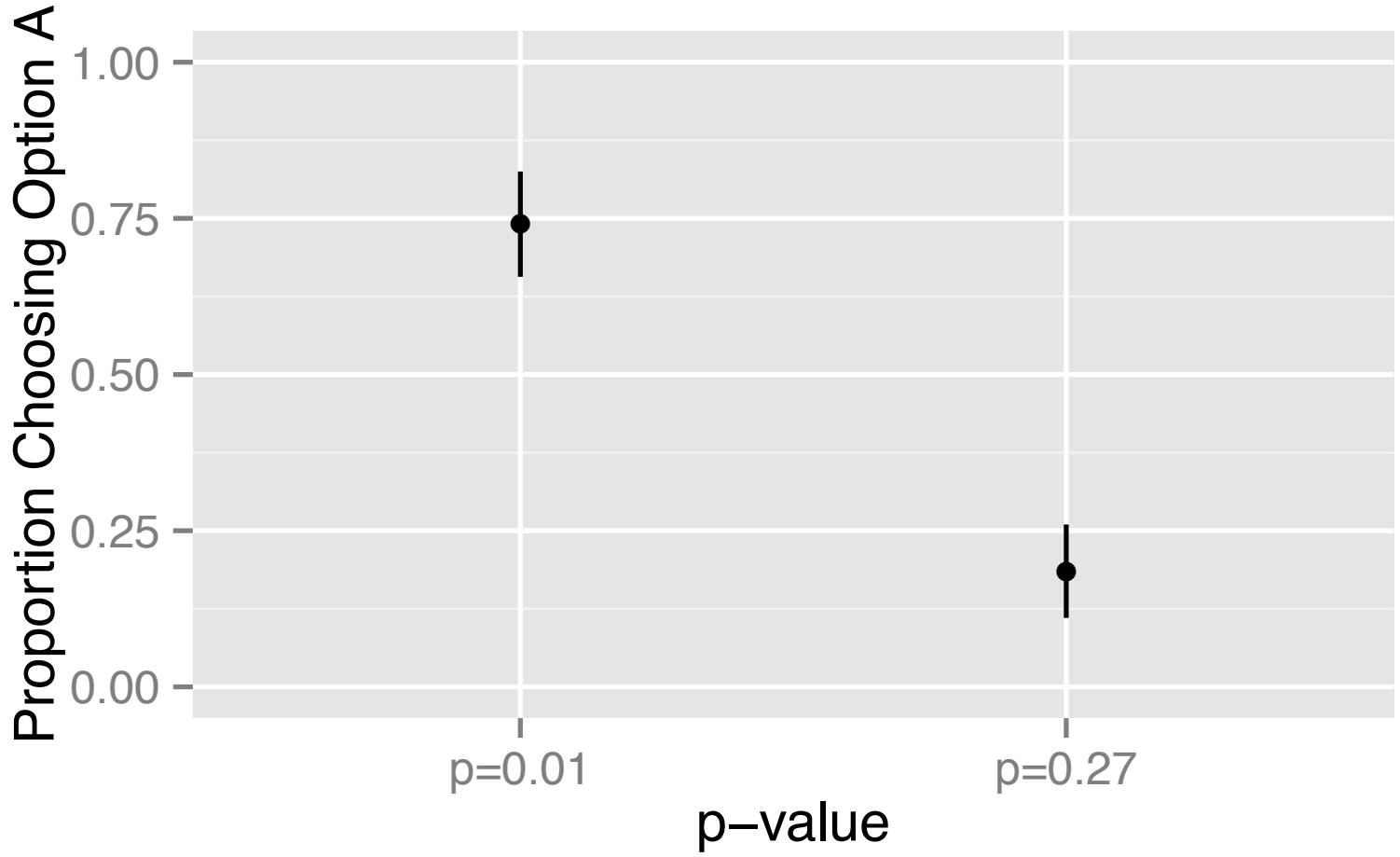
Study 1: Descriptive Statements
New England Journal of Medicine Authors



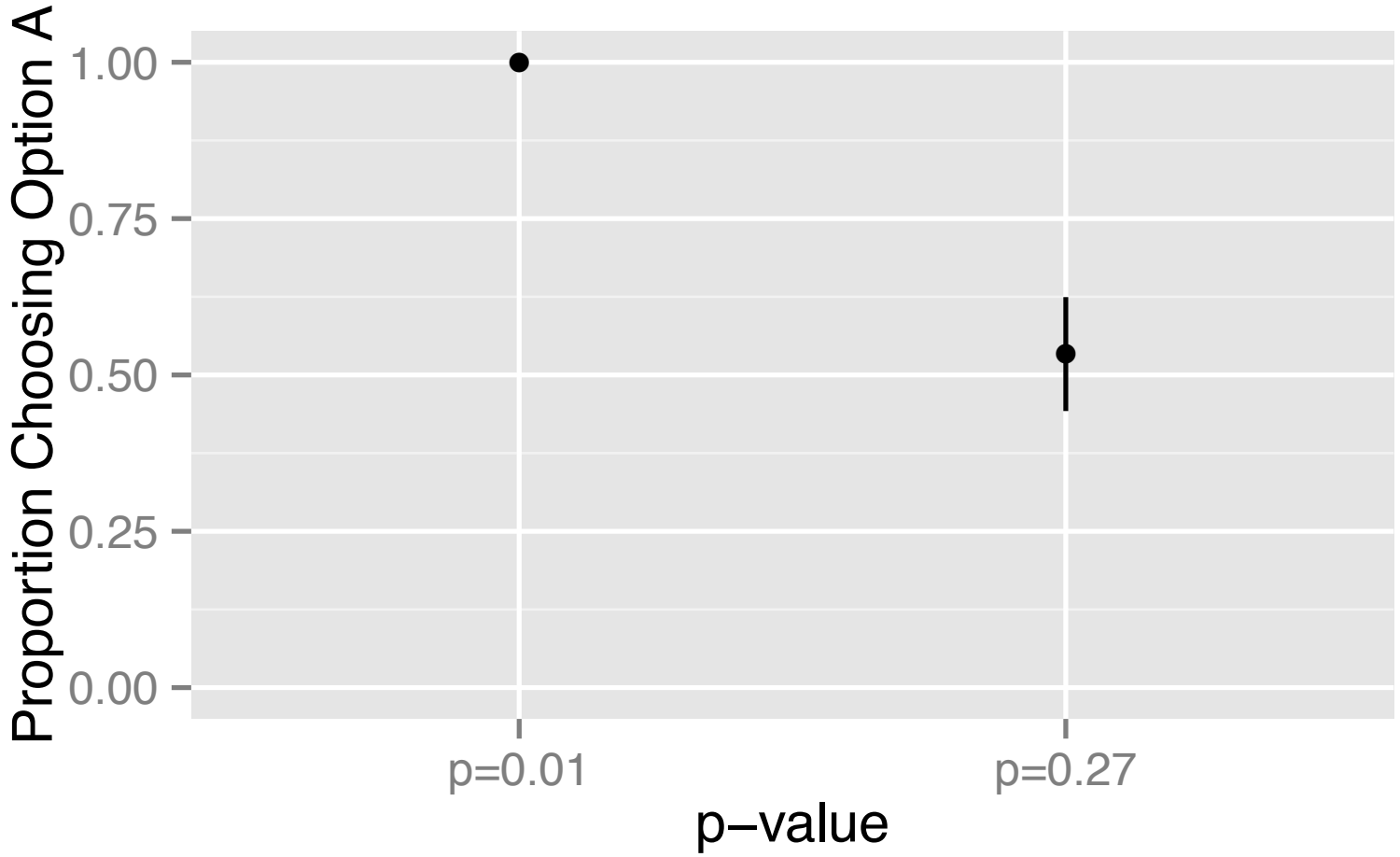
Study 1: Descriptive Statements
Psychological Science Editorial Board



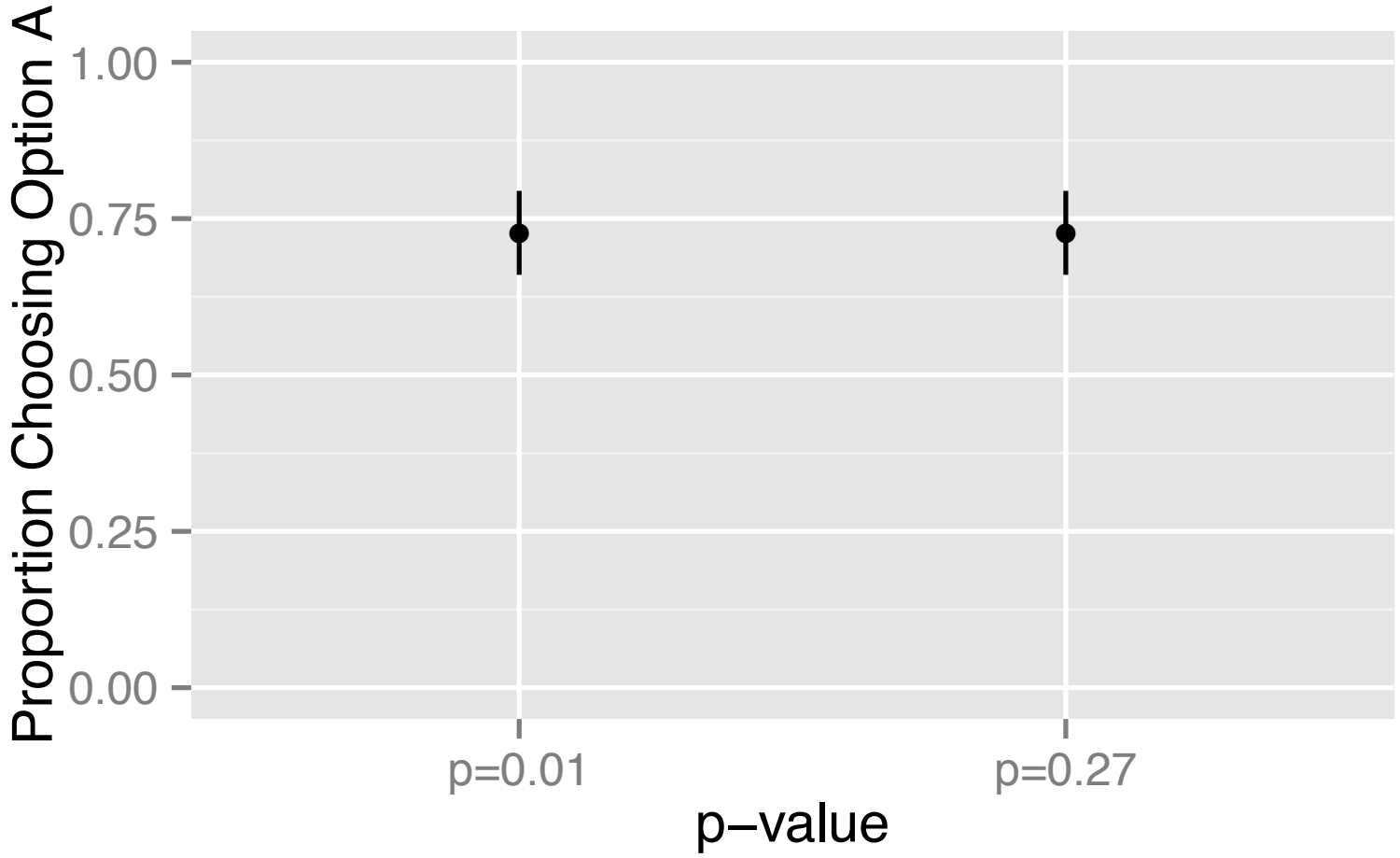
Study 1: Descriptive Statements
Marketing Science Institute Young Scholars



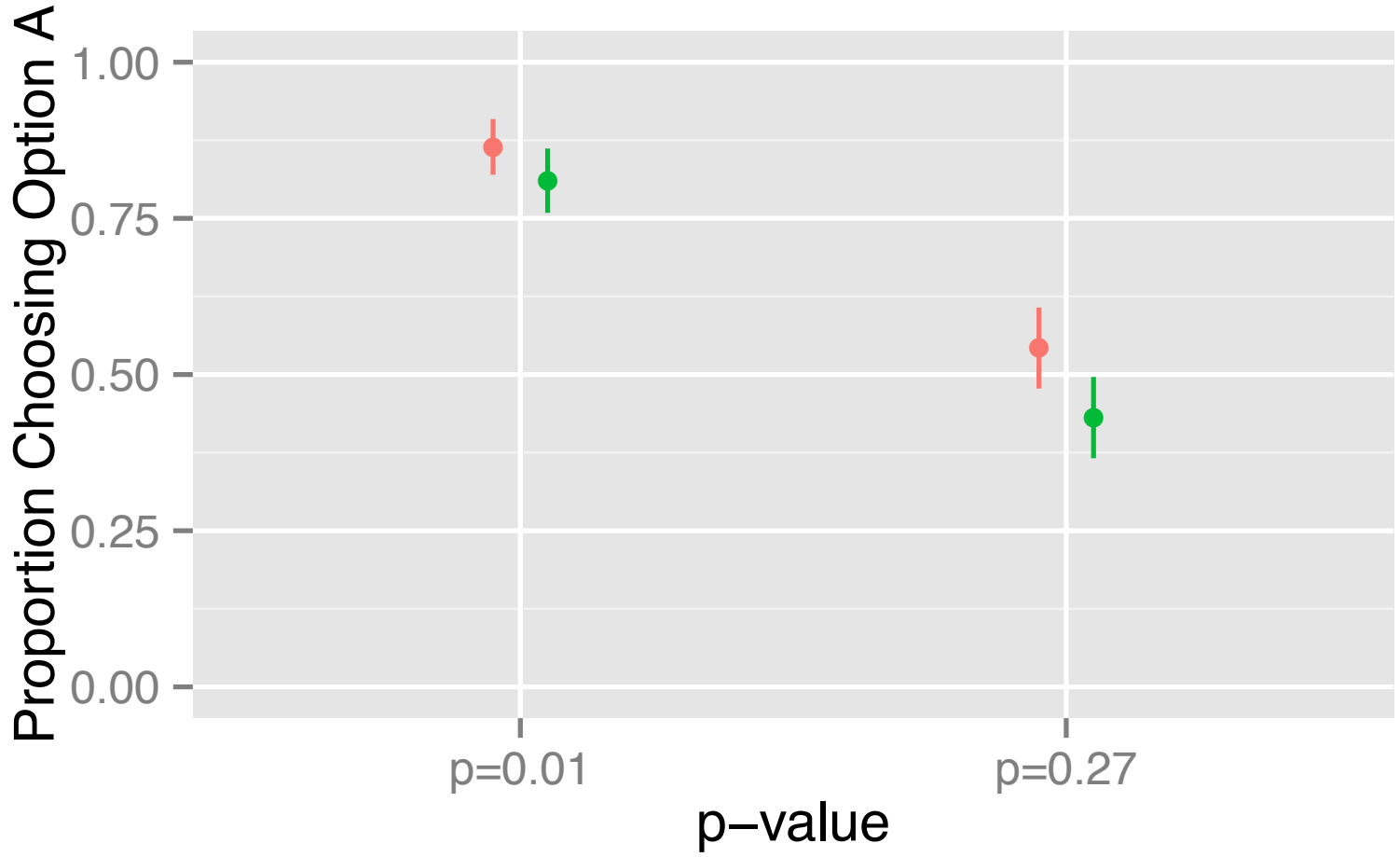
Study 1: Descriptive Statements
Undergraduates Who Took A Statistics Course



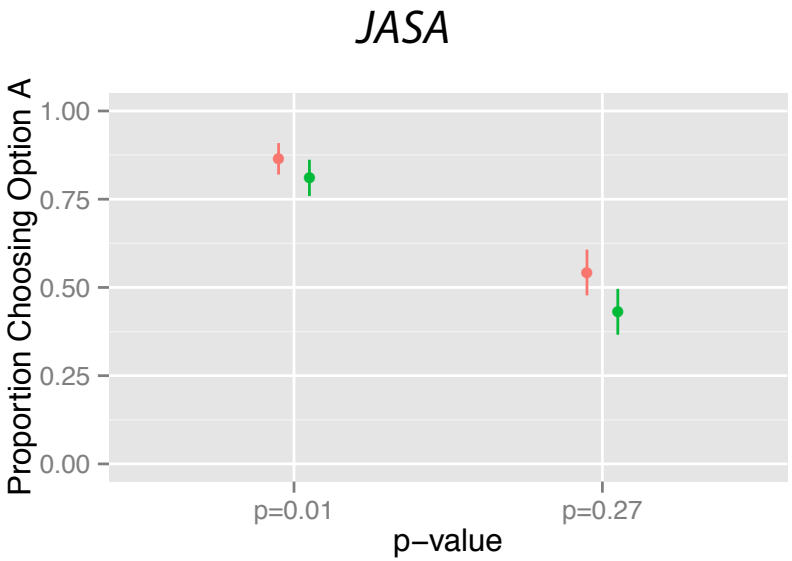
Study 1: Descriptive Statements
Undergraduates Who Did *Not* Take A Statistics Course

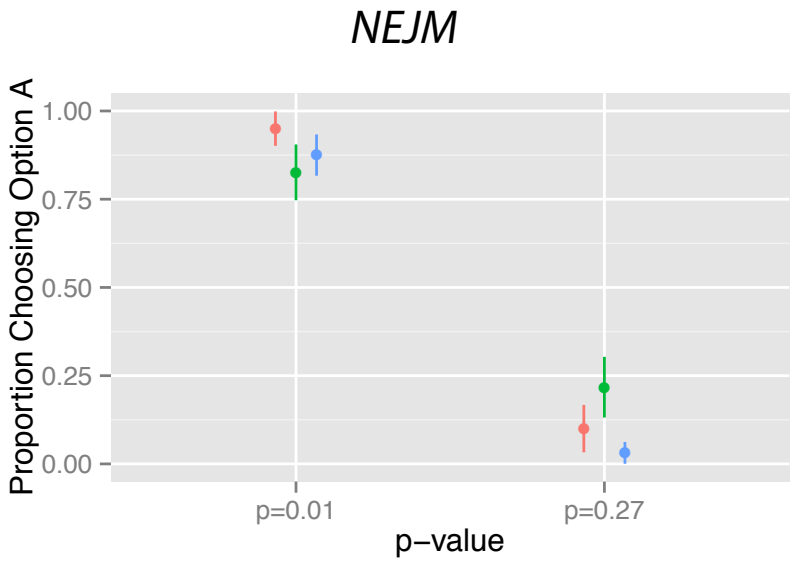


Study 1: Descriptive Statements
Journal of the American Statistical Association Authors



Study 1: Descriptive Statements
JASA & *NEJM* Authors





Study 2: Judgment and Choice

- Below is a summary of a study from an academic paper:
- The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease.
- A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a *p*-value of 0.025/0.075/0.125/0.175.
- Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?

  A. A person drawn randomly from the same patient population as the patients in the study is **more likely** to recover from the disease if given Drug A than if given Drug B.

  B. A person drawn randomly from the same patient population as the patients in the study is **less likely** to recover from the disease if given Drug A than if given Drug B.

  C. A person drawn randomly from the same patient population as the patients in the study is **equally likely** to recover from the disease if given Drug A than if given Drug B.

  D. It **cannot be determined** whether a person drawn randomly from the same patient population as the patients in the study is more/less/equally likely to recover.

Study 2: Judgment and Choice

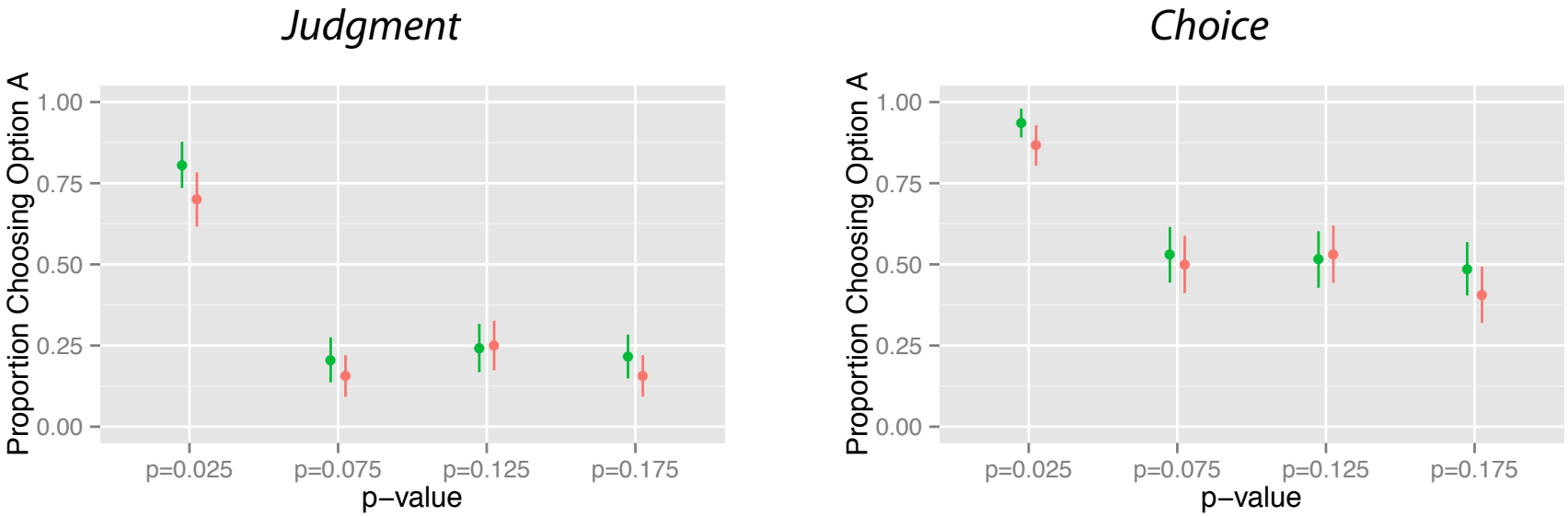
- Below is a summary of a study from an academic paper:
- The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease.
- A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a *p*-value of 0.025/0.075/0.125/0.175.
- If you were to advise a loved one who was a patient from the same population as those in the study, what drug would you advise him or her to take?

    A. I would advise Drug A.

    B. I would advise Drug B.

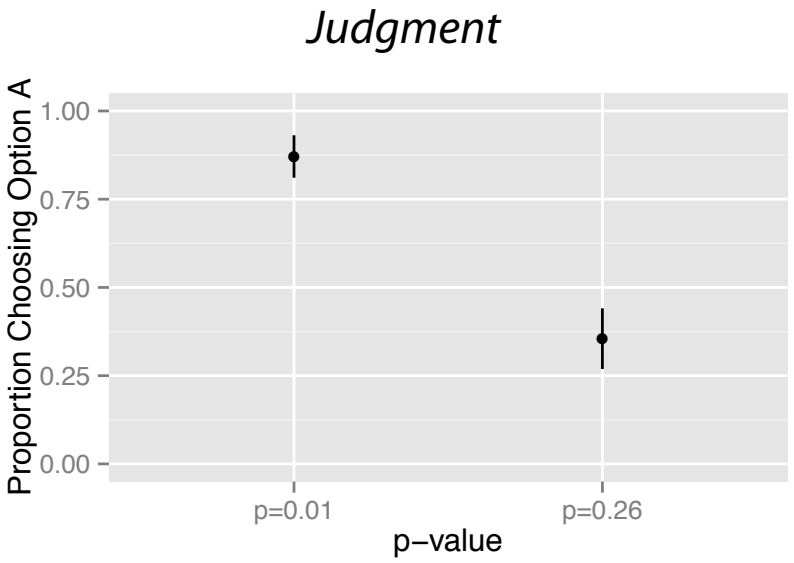
    C. I would advise that there is no difference between Drug A and Drug B.

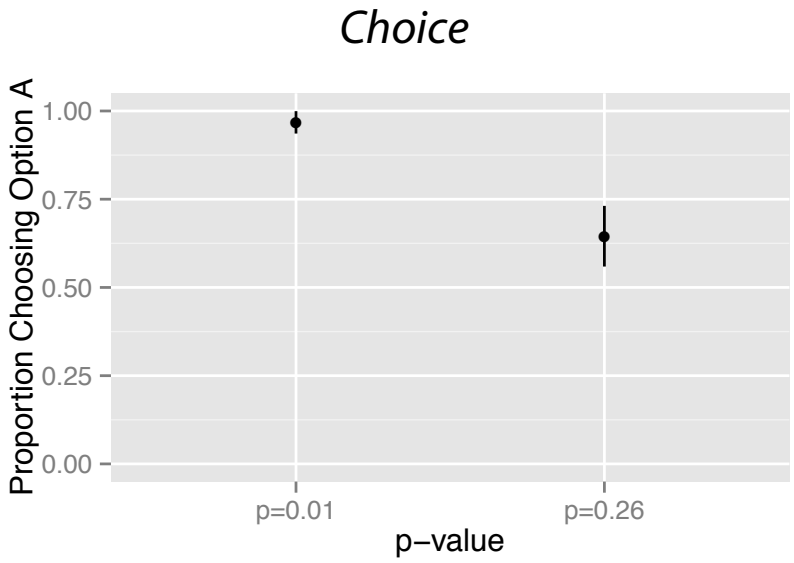
Study 2: Judgment and Choice
American Journal of Epidemiology Authors



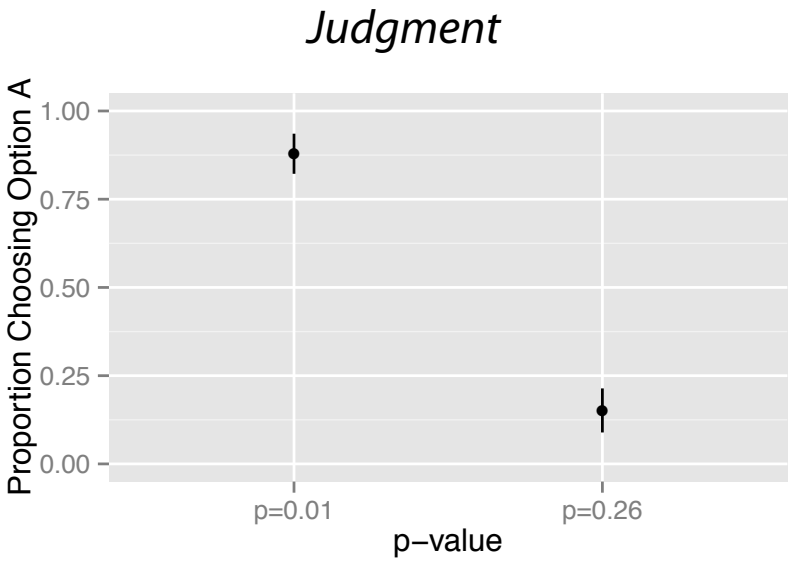
- Large Treatment Difference: 57% versus 39%.
- Small Treatment Difference: 52% versus 44%.

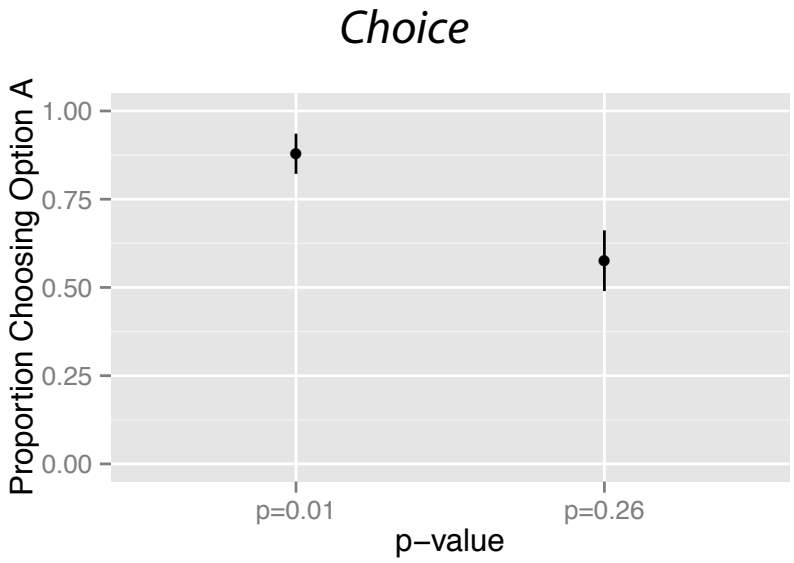
Study 2: Judgment and Choice
Cognition Editorial Board



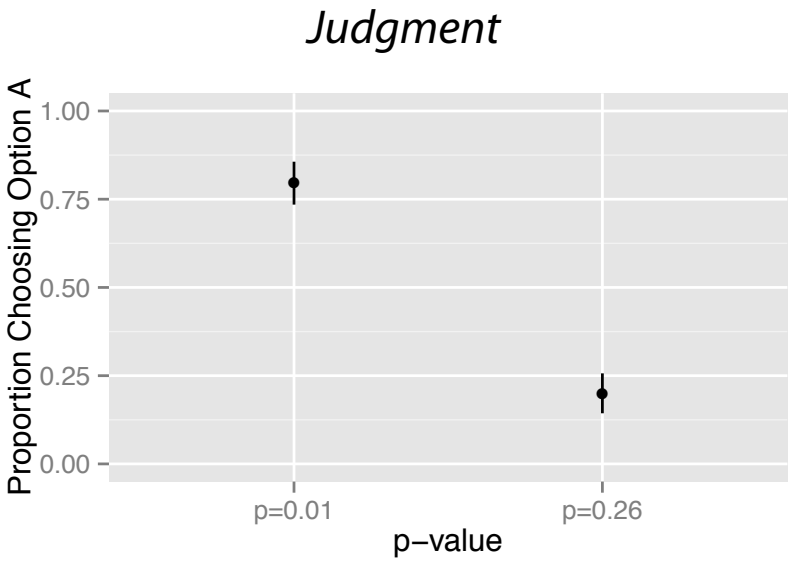


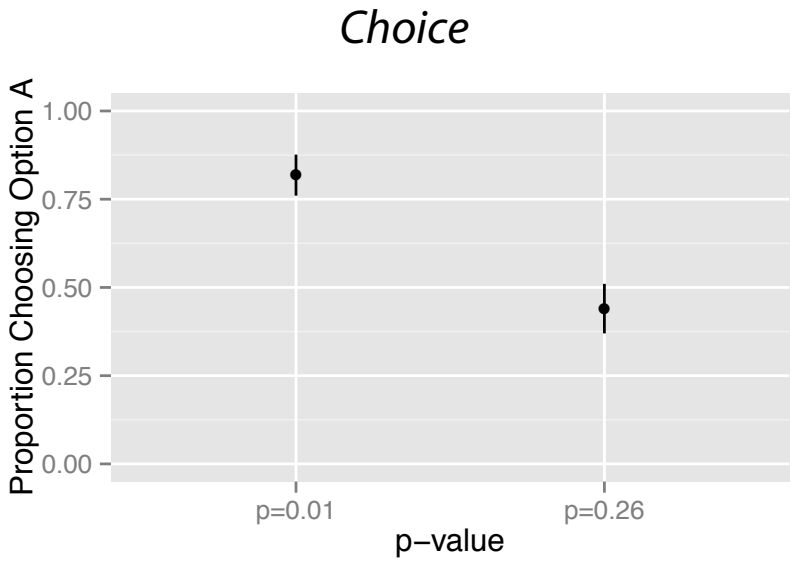
Study 2: Judgment and Choice
Social Psychological and Personality Science Ed. Board



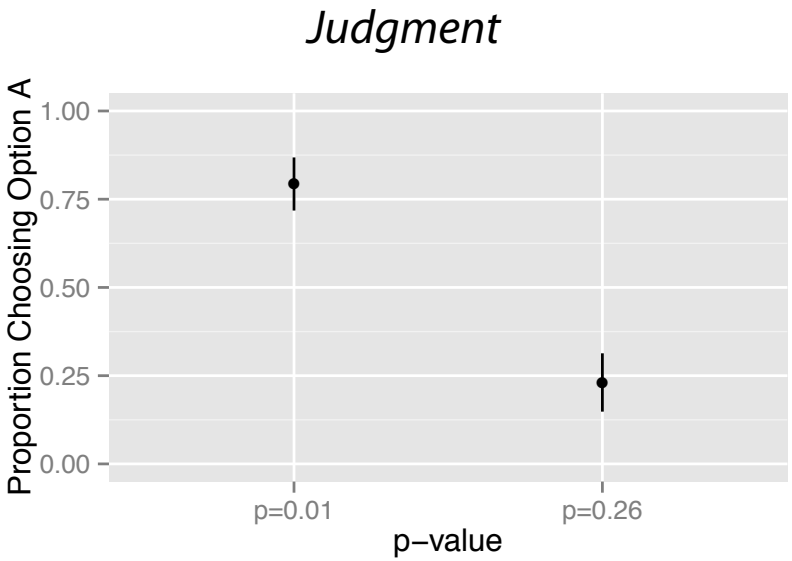


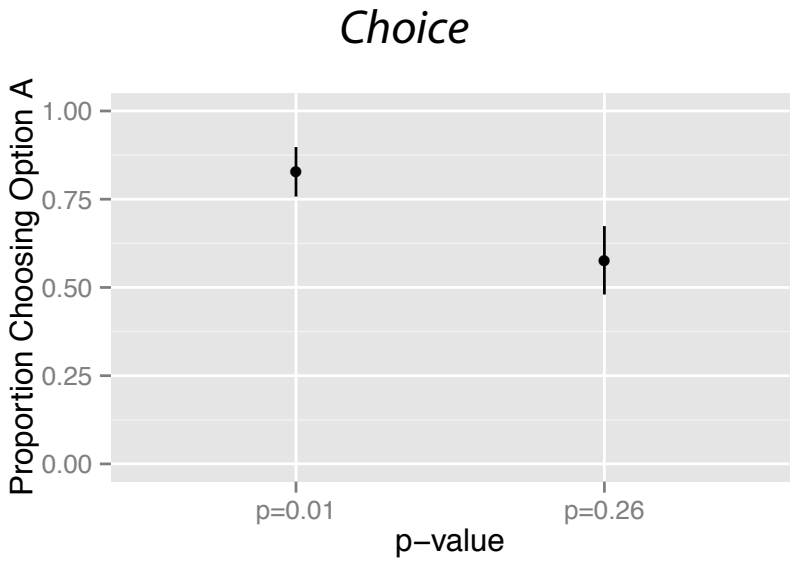
Study 2: Judgment and Choice
American Economic Review Authors



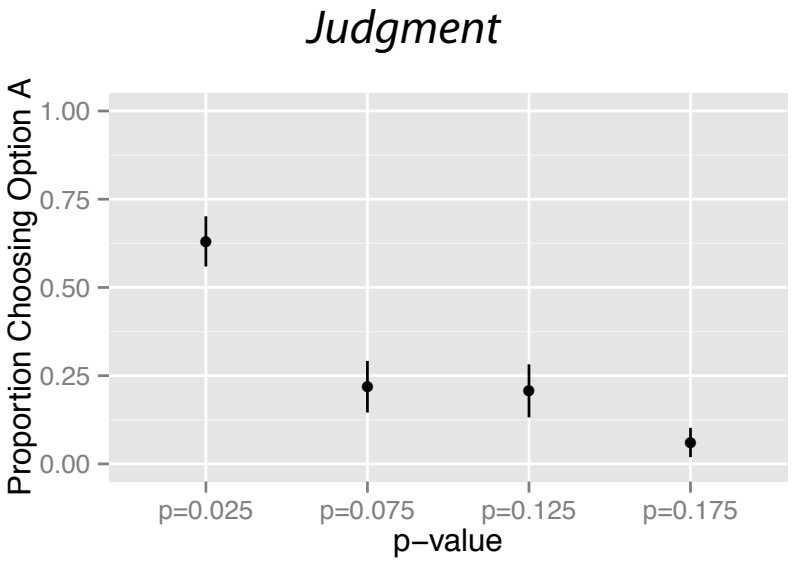


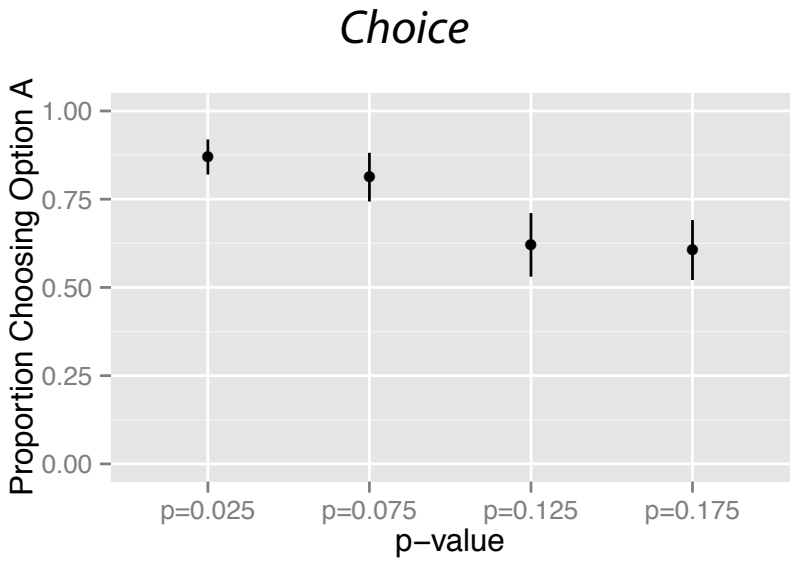
Study 2: Judgment and Choice
Quarterly Journal of Economics Authors



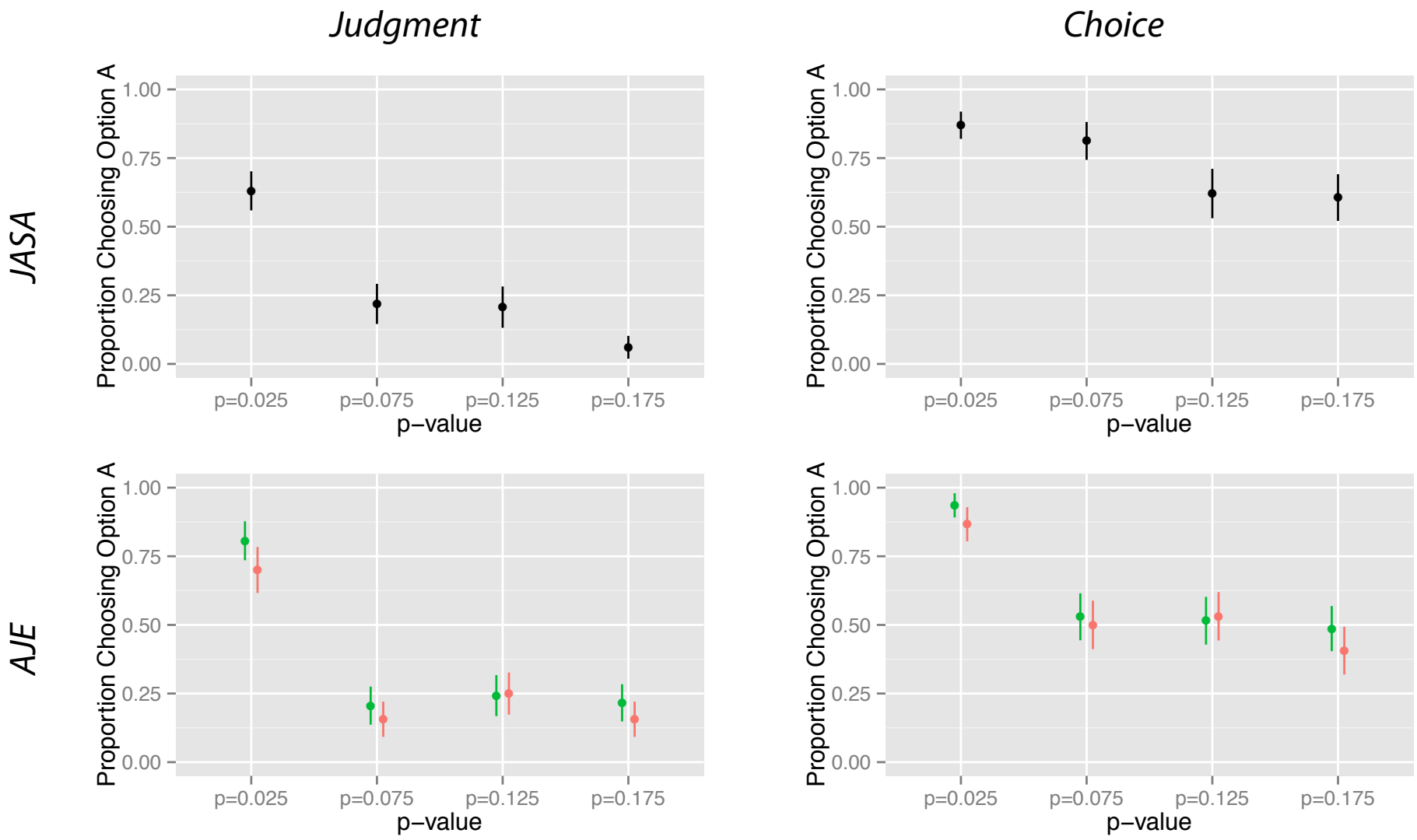


Study 2: Judgment and Choice
Journal of the American Statistical Association Authors

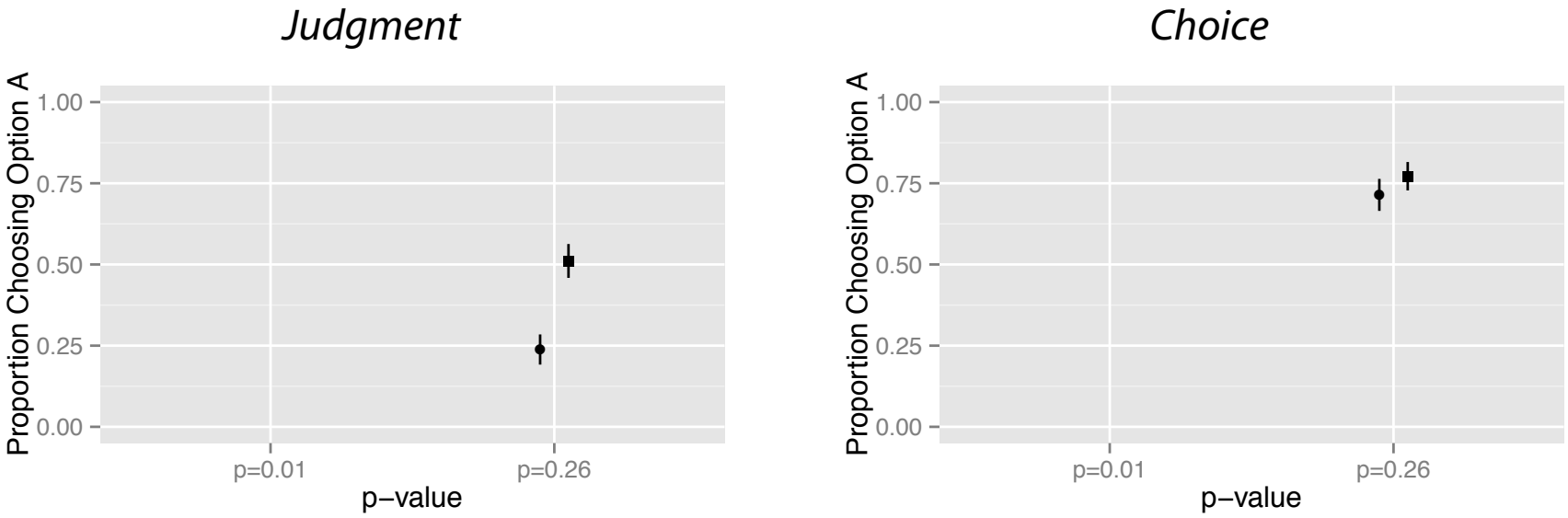




Study 2: Judgment and Choice
JASA & *AJE* Authors



Study 2: Judgment and Choice Economists (*AER* and *JPE* Authors)



- ●: No posterior probability.
- ■: Posterior probability.
- *Note*: Given results with $p = 0.26$, there is a 1 - 0.26/2 = 0.87 posterior probability that Drug A is more effective than Drug B under a non-informative prior.

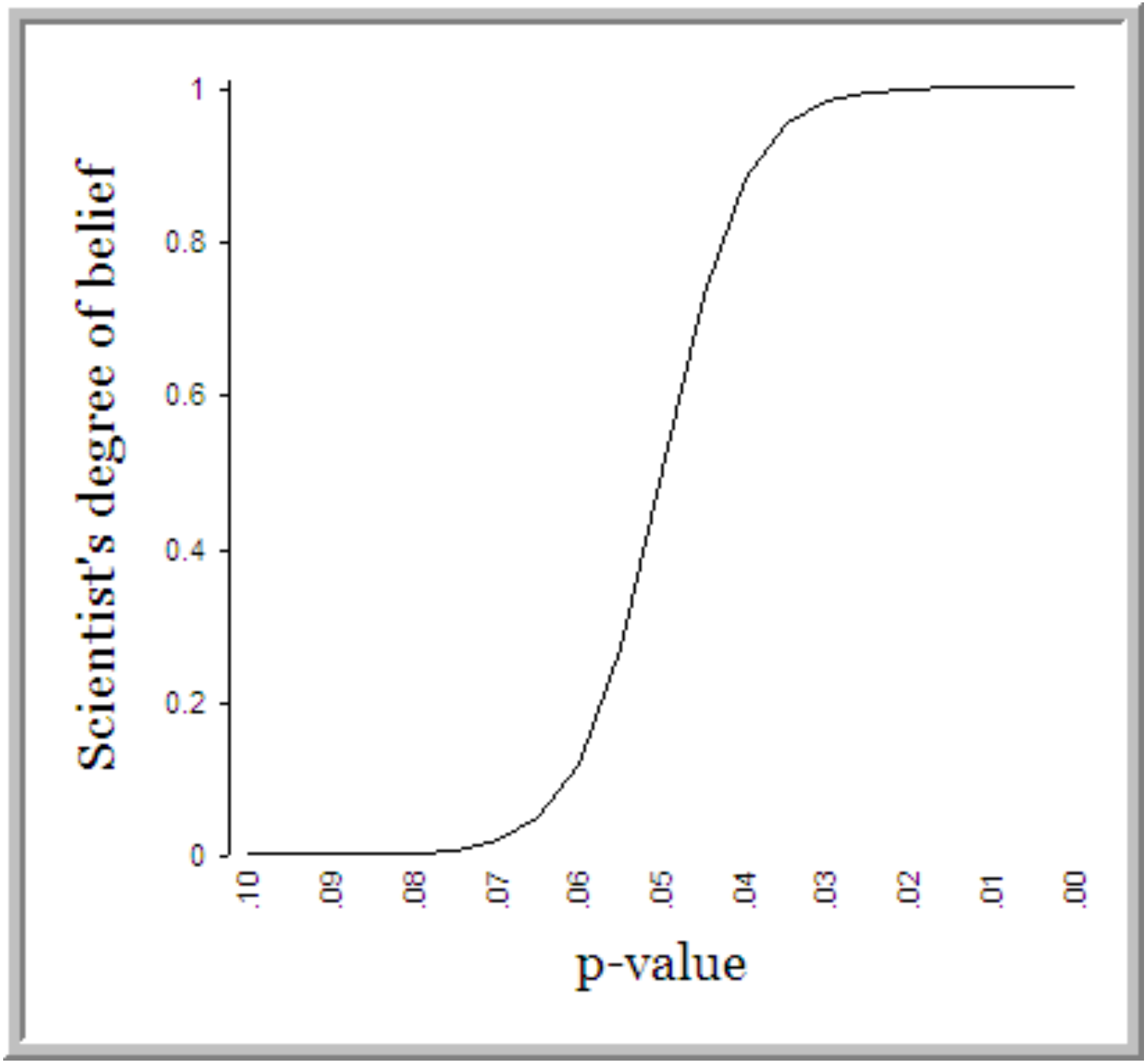
Discussion

- Researchers across a variety of fields–including expert statisticians–are likely to make erroneous statements and judgments when presented with evidence that fails to attain statistical significance.

- Quantitative results in tandem with researchers' own explanations of their reasoning suggest that the preponderance of researchers focus primarily or even exclusively simply on whether or not the *p*-value is below or above the "magic number" of 0.05.

- Most discouraging findings regarding statisticians:
  - About half the subjects in Study 1 failed to identify differences that were not statistically significant as different.
  - The vast majority of the subjects in Study 2 failed to select option A for both the likelihood judgment and choice question (i.e., because the posterior probability that Drug A was more effective than Drug B was larger than 90% in each of the four *p*-value settings).

- It was most encouraging–if not entirely surprising–that statisticians performed better than applied researchers
  - This suggests a deep as opposed to cursory training in statistics that includes exposure to forms of statistical reasoning outside NHST can help attenuate dichotomous thinking even if it cannot entirely eliminate it.



Source: Dan Goldstein, Decision Science News

What Can Be Done?

- Our results suggest the dominant NHST paradigm and the rote and recipe-like manner in which it is typically taught and practiced can impair reasoning.
    - The problem seems to lie with the dichotomization of evidence intrinsic to NHST: the assignment of evidence to the different categories “statistically significant” and “not statistically significant” appears to be simply too strong an inducement to the conclusion that the items thusly assigned are categorically different.

- A greater focus on effect sizes, their variability, and the uncertainty in estimates of them will naturally lead researchers to think of evidence as lying on a continuum.

- Instead of thinking of effects as being “there” or “not there,” careful consideration of study-level and individual-level variation as well as moderators of this variation can lead researchers to develop deeper and richer theories.

- Researchers should also move away from focusing solely on statistical considerations: real world costs and benefits, size and scientific importance of results, data quality, propriety of the statistical analysis, etc.

What Can Be Done? Abandon Statistical Thresholds!

- Perhaps most importantly we should move away from any forms of dichotomous or categorical reasoning whether in the form of NHST or otherwise:
  - Confidence intervals evaluated only on the basis of whether or not they contain zero or some other number.
  - Posterior probabilities evaluated only on the basis of whether or not they are above some particular threshold.
  - Bayes Factors evaluated only in terms of discrete categories.
  - ...

- Each is a form of statistical alchemy that falsely promises to transmute randomness into certainty, an "uncertainty laundering" [Gelman, 2016] that begins with data and concludes with dichotomous declarations of truth or falsity—binary statements about there being "an effect" or "no effect"—based on some *p*-value or other statistical threshold being surpassed.