Beyond *p* values:
Better Inference with the New Statistics

Bob Calin-Jageman

Neuroscience Program, Dominican University

Geoff Cumming

Psychological Science, La Trobe University

Overview

- What is the New Statistics?
- Better inference with the New Statistics
- There are no panaceas

The first principle is that you must not fool yourself – and you are the easiest person to fool.

- Richard Feynman, 1974

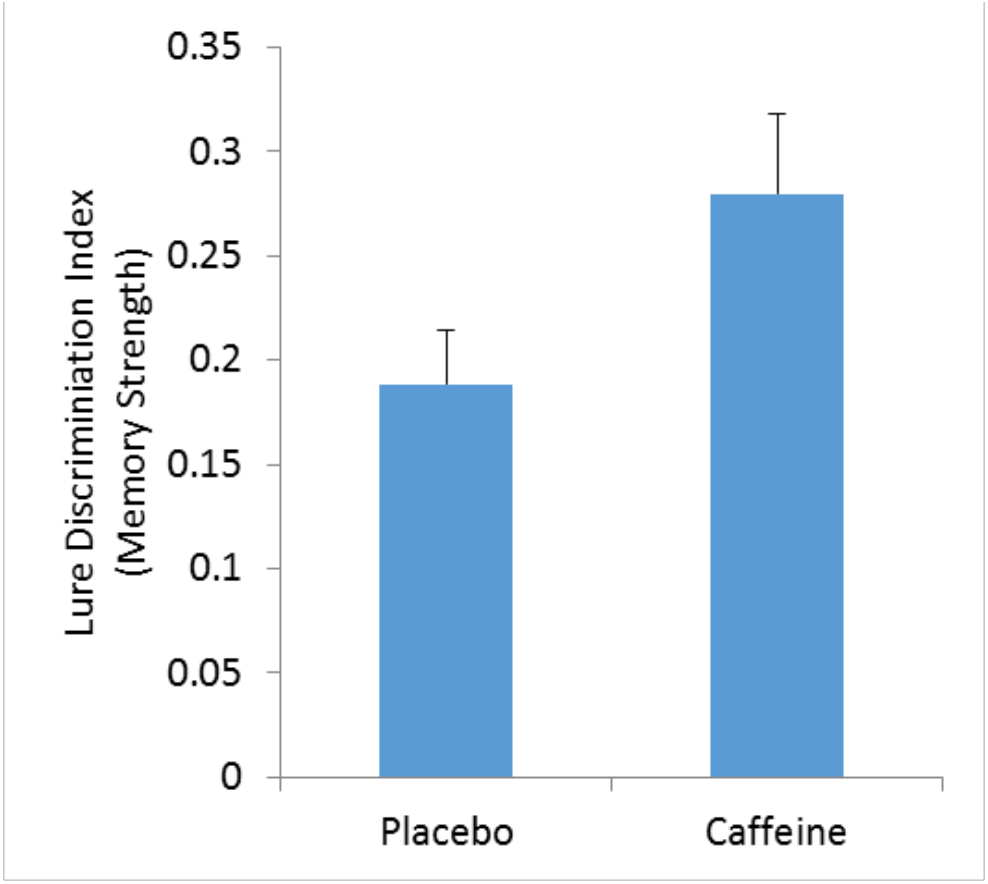
What is the New Statistics?

- Ask quantitative questions and give quantitative answers (effect sizes).
- Countenance uncertainty by visualizing and interpreting confidence intervals or credible intervals.
- Seek replication and use meta-analysis as a matter of course.
- Use Open Science practices to enhance the interpretability of research results.

There is no uncertainty that we can't quantify.

- Emery Brown, MIT

Does Caffeine Affect Memory?



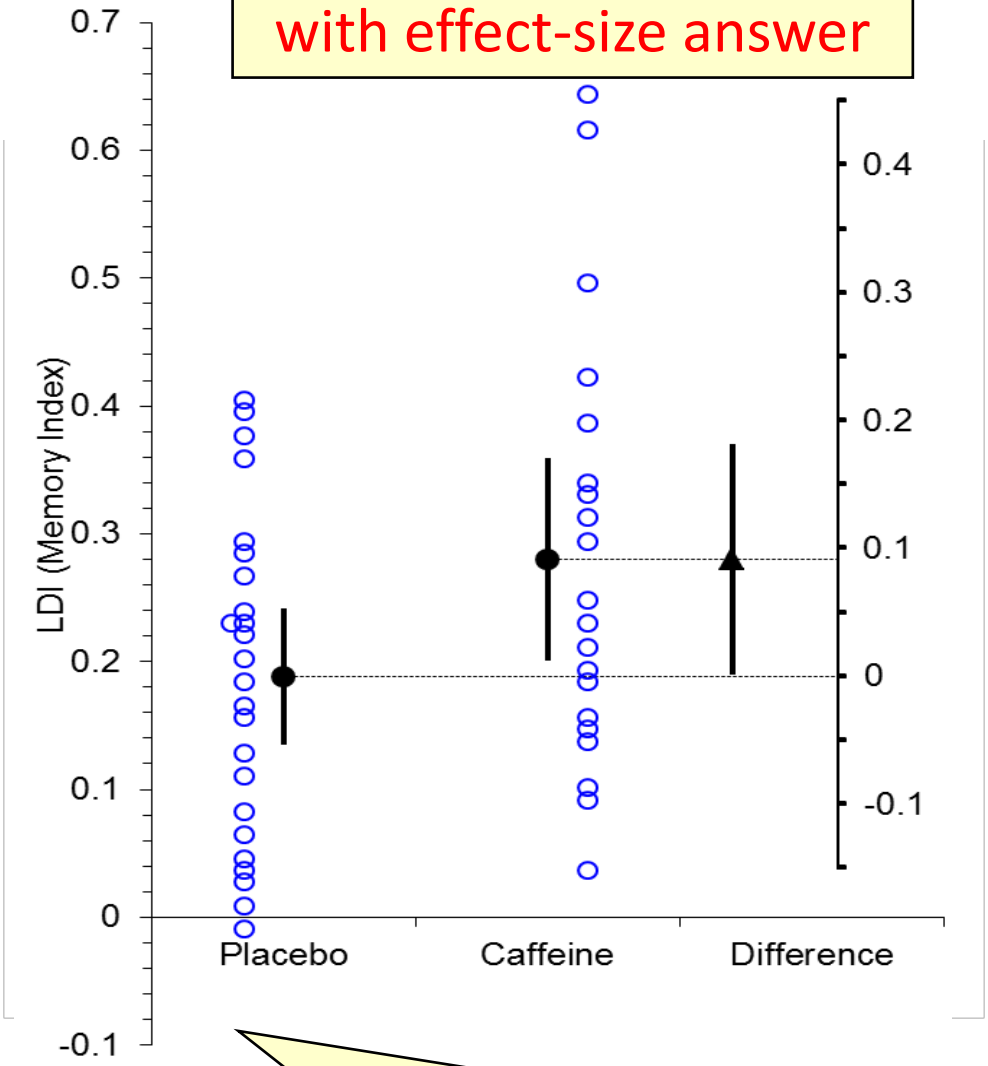
Group difference was statistically significant ($t(42) = 2.04$, $p = 0.05$).

Caffeine improves memory.

Data modeled after Borota et al. (2014)

How Much Does Caffeine Affect Memory?

Quantitative question with effect-size answer



Plot shows data, effect size, and CI

CI represents uncertainty in analysis & conclusion

Caffeine benefit:
d = 0.61, 95% CI[0.01, 1.22]

The data are consistent with anywhere from a vanishingly small up to a very large effect.

Need more data; caffeine probably doesn't hurt memory.

Data modeled after Borota et al. (2014)

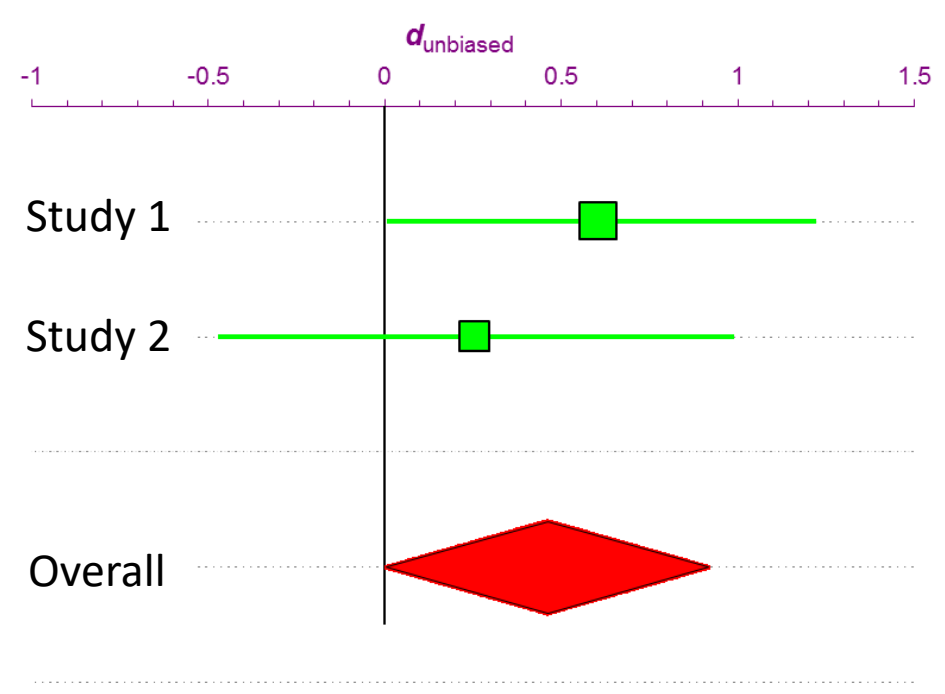
How Much Does Caffeine Affect Memory?

- We repeated the experiment….
- We combined data across the two experiments for the placebo and 200mg caffeine conditions to increase power.
- We found that performance for the 200mg caffeine condition was higher than that for placebo ($t(71) = 2.0$, $p = 0.049$)

Data modeled after Borota et al. (2014)

How Much Does Caffeine Affect Memory?

Seek replication; Use meta-analysis



Across two studies,
d = 0.46, CI[0.004, 0.92].

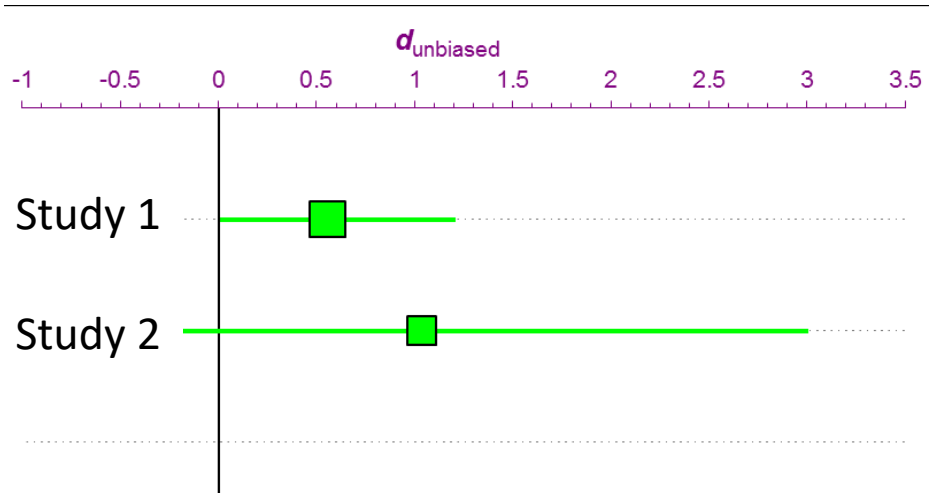
The balance of the evidence suggests that caffeine could improve memory from a very, very small up to a very large amount.

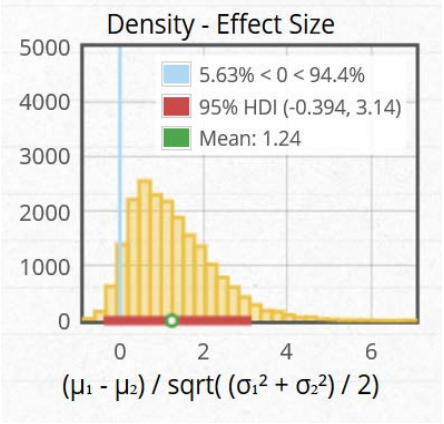
Overall, the data is not especially informative.

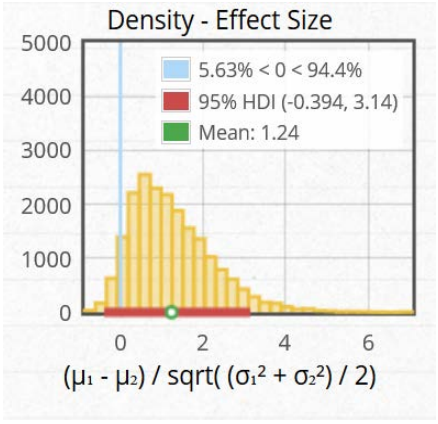
Data modeled after Borota et al. (2014)

How Much Does Caffeine Affect Memory?

Seek replication; Use meta-analysis







Bayesian Credible Intervals

Across two studies,
d = 0.46, CI[0.004, 0.92].

The balance of the evidence suggests that caffeine could improve memory from a very, very small up to a very large amount.

Overall, the data is not especially informative.

Data modeled after Borota et al. (2014)

Don't test; estimate!

Testing Framework

- Qualitative questions
- *p* value or Bayes Factor
  - Focuses on plausibility of a specific null hypothesis
- Not in same format as meta-analysis

New Statistics (aka Estimation)

- Quantitative questions
- Effect size with CI
  - Focuses on uncertainty and practical significance
- Directly amenable to meta-analysis

Estimation is for everyone:

Frequentists: Confidence Intervals

Bayesians: Credible Intervals

Switching is Easy:

Frequentists: Same foundation

Bayesians: Everything is easy for you

Don’t test; estimate!

Testing Framework

- Qualitative questions
- *p* value or Bayes Factor
  - Focuses on plausibility of a specific null hypothesis
- Not in same format as meta-analysis

$$Test\ Statistic = \frac{Effect\ Size}{Standard\ Error}$$

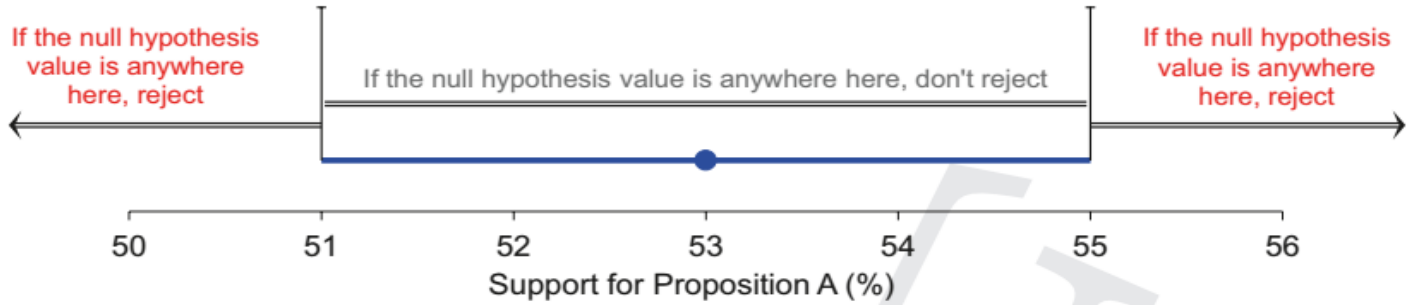
If *Test_Statistic* > *Critical Value*, reject H_0

New Statistics (aka Estimation)

- Quantitative questions
- Effect size with CI
  - Focuses on uncertainty and practical significance
- Directly amenable to meta-analysis

$$MoE = Standard\ Error\ * Critical\ Value$$

$$CI = Effect\ Size\ \pm MoE$$



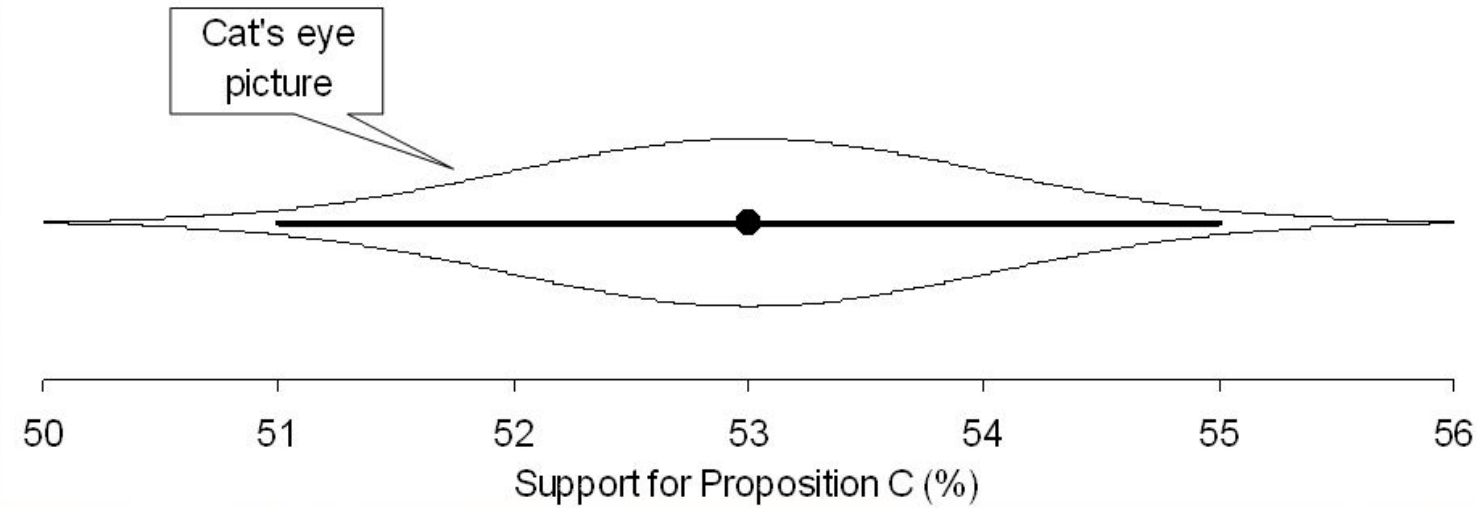
Don’t test; estimate!

Testing Framework

- Qualitative questions
- *p* value or Bayes Factor
  - Focuses on plausibility of a specific null hypothesis
- Not in same format as meta-analysis

New Statistics (aka Estimation)

- Quantitative questions
- Effect size with CI
  - Focuses on uncertainty and practical significance
- Directly amenable to meta-analysis



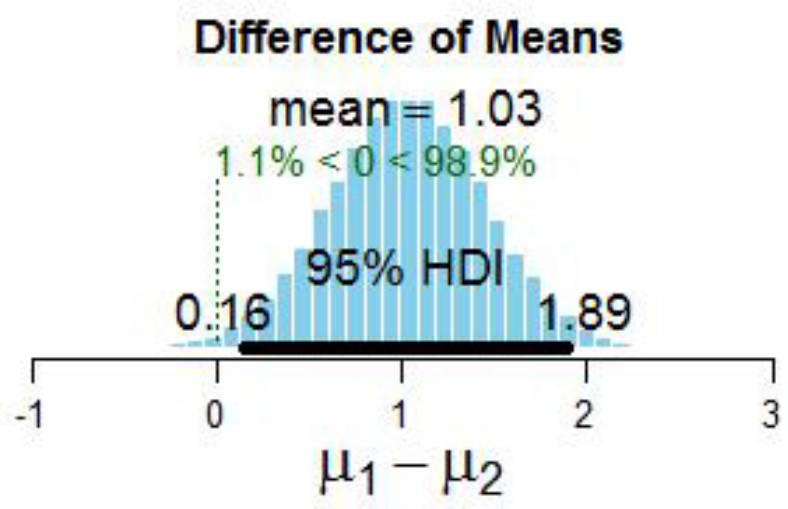
Don’t test; estimate!

Testing Framework

- Qualitative questions
- *p* value or Bayes Factor
  - Focuses on plausibility of a specific null hypothesis
- Not in same format as meta-analysis

New Statistics (aka Estimation)

- Quantitative questions
- Effect size with CI
  - Focuses on uncertainty and practical significance
- Directly amenable to meta-analysis



Bayesian Credible Interval (BEST by Kruschke, 2013)

Overview

- What is the New Statistics?
- **Better inference with the New Statistics**
- There are no panaceas

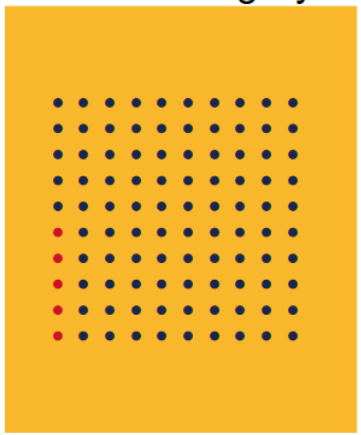
The first principle is that you must not fool yourself – and you are the easiest person to fool.

- Richard Feynman, 1974

Evaluating approaches to inference

- Mathematically sound
- Empirically validated
- Supports good statistical cognition

Of 100 women who have surgery



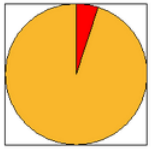
5 out of **100** women will require additional treatment

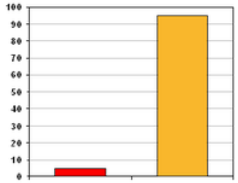
NOT

20% less women will required additional treatment

5% of women will required additional treatment

OR





Gigerenzer et al 1995, Feldman-Stewart et al 2000, Fagerlin et al review 2007

p -> Sound but dangerous

Evaluating approaches to inference

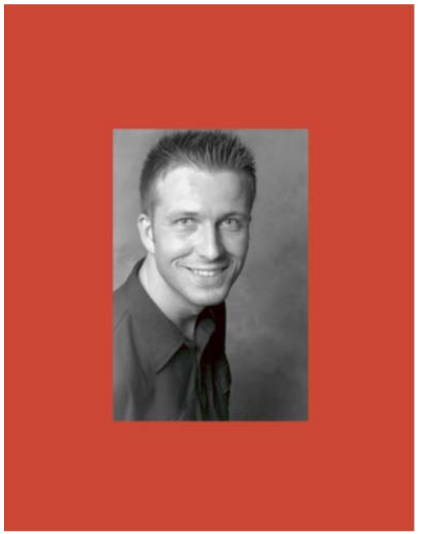
- Mathematically sound
- Empirically validated
- Supports good statistical cognition

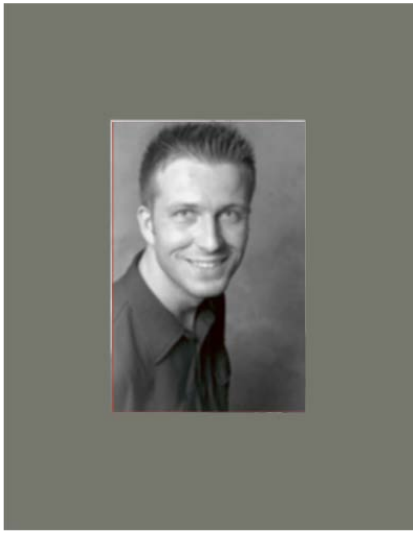
Our claim: The New Statistics helps support good inference.

This is an empirical claim that requires more research.

- Better judgements of consistency
- Reduced over-confidence in small samples
- Avoids dichotomous thinking which might help limit publication bias
- And more…

CIs make consistency more clear



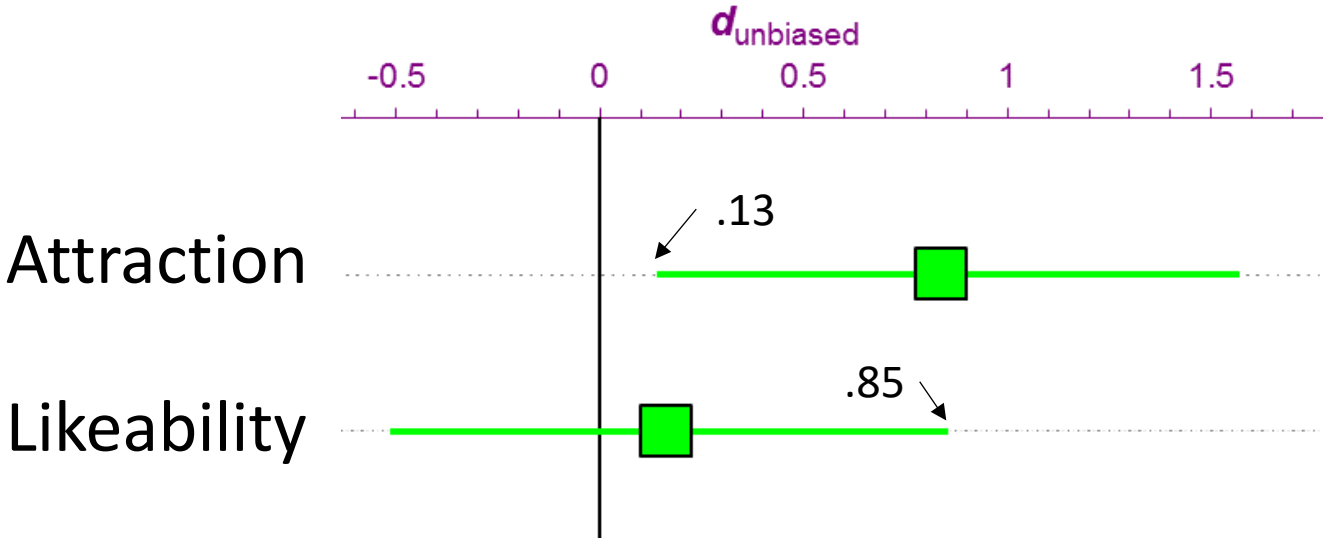


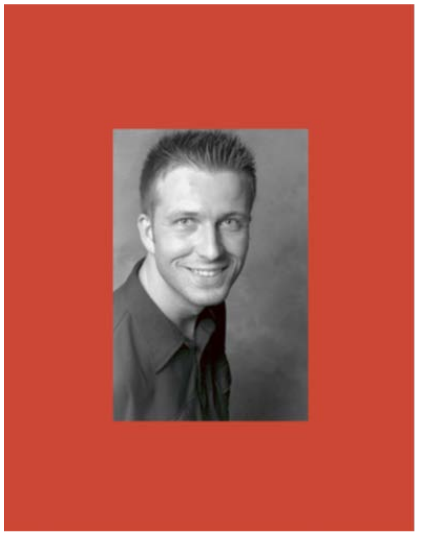
- Red increased physical attraction:
  - $t(32)=2.44$, p=.05, d=0.86
- Red did not influence likeability:
  - $t(32) = 0.48$, $p = 0.63$, $d = 0.17$

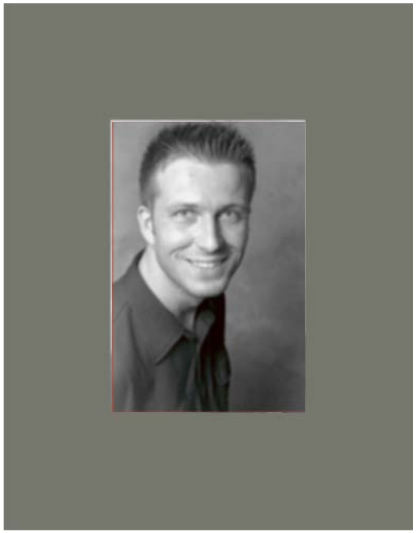
Red selectively increases physical attraction.

Elliot et al., 2010, Experiment 3

CIs make consistency more clear







- Red increased physical attraction:
  - $t(32)=2.44$, p=.05, d=0.86
- Red did not influence likeability:
  - $t(32) = 0.48$, $p = 0.63$, $d = 0.17$

~~**Red selectively increases physical attraction.**~~
Need more data.

Elliot et al., 2010, Experiment 3

Overcoming the "~~Law~~ Fallacy of Small Numbers"

Researchers have too much confidence in small samples:

- In psychology and cognitive neuroscience studies, typical power to detect small, medium, and large effects is 0.12, 0.44, 0.73 (Szucs & Ioannidis, 2016)
- Typical power is 0.08 in neuroimaging and .18-.31 in animal model neuroscience research (Button et al., 2013)

Focusing on uncertainty can help

Participants primed to feel lonely reported lower room temperatures ($t(63) = 2.02$, $p < .05$)

-- Zhong & Leonardelli, 2008, *Psychological Science*

Never directly replicated

So far, cited 634 times

I suspect that the main reason they [CIs] are not reported is that they are so embarrassingly large!

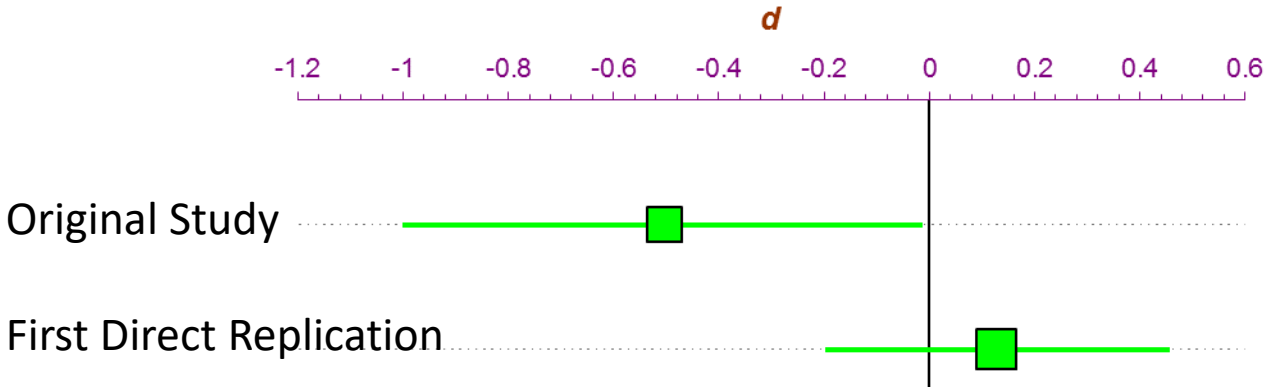
- Cohen, 1994

Overcoming the "~~Law~~ Fallacy of Small Numbers"

Researchers have too much confidence in small samples:

- In psychology and cognitive neuroscience studies, typical power to detect small, medium, and large effects is 0.12, 0.44, 0.73 (Szucs & Ioannidis, 2016)
- Typical power is 0.08 in neuroimaging and .18-.31 in animal model neuroscience research (Button et al., 2013)

Focusing on uncertainty can help.

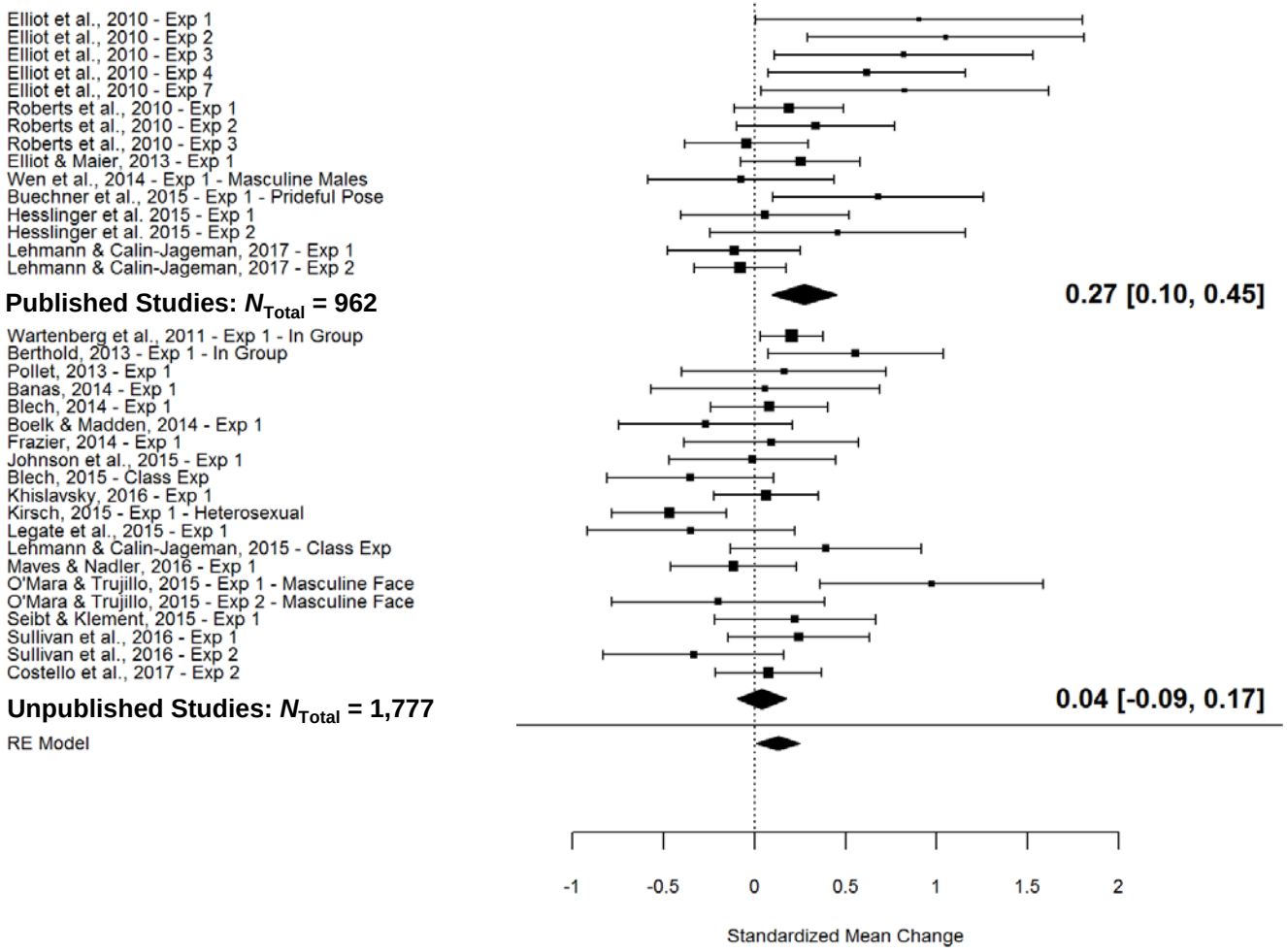


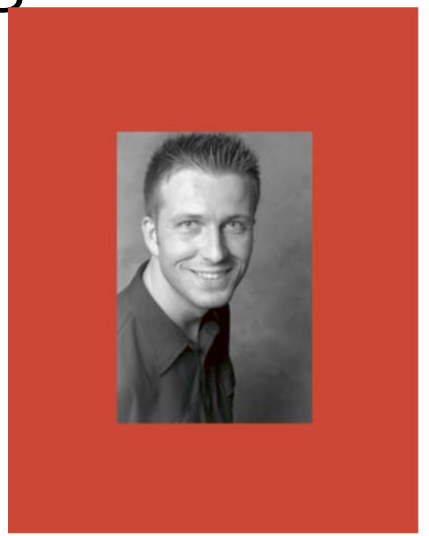
Original results uninformative.

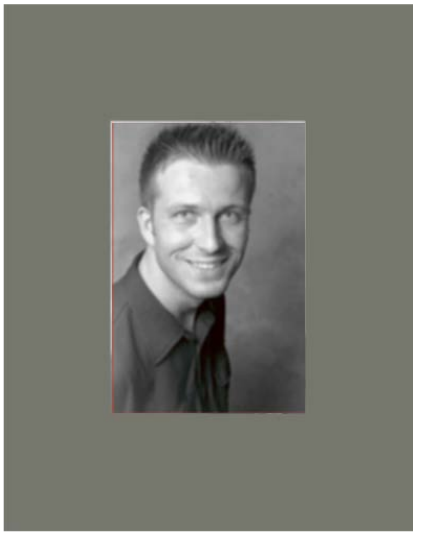
Wide range of possible replication results expected.

I suspect that the main reason they [CIs] are not reported is that they are so embarrassingly large!
- Cohen, 1994

Dichotomous Thinking: Right or Wrong?





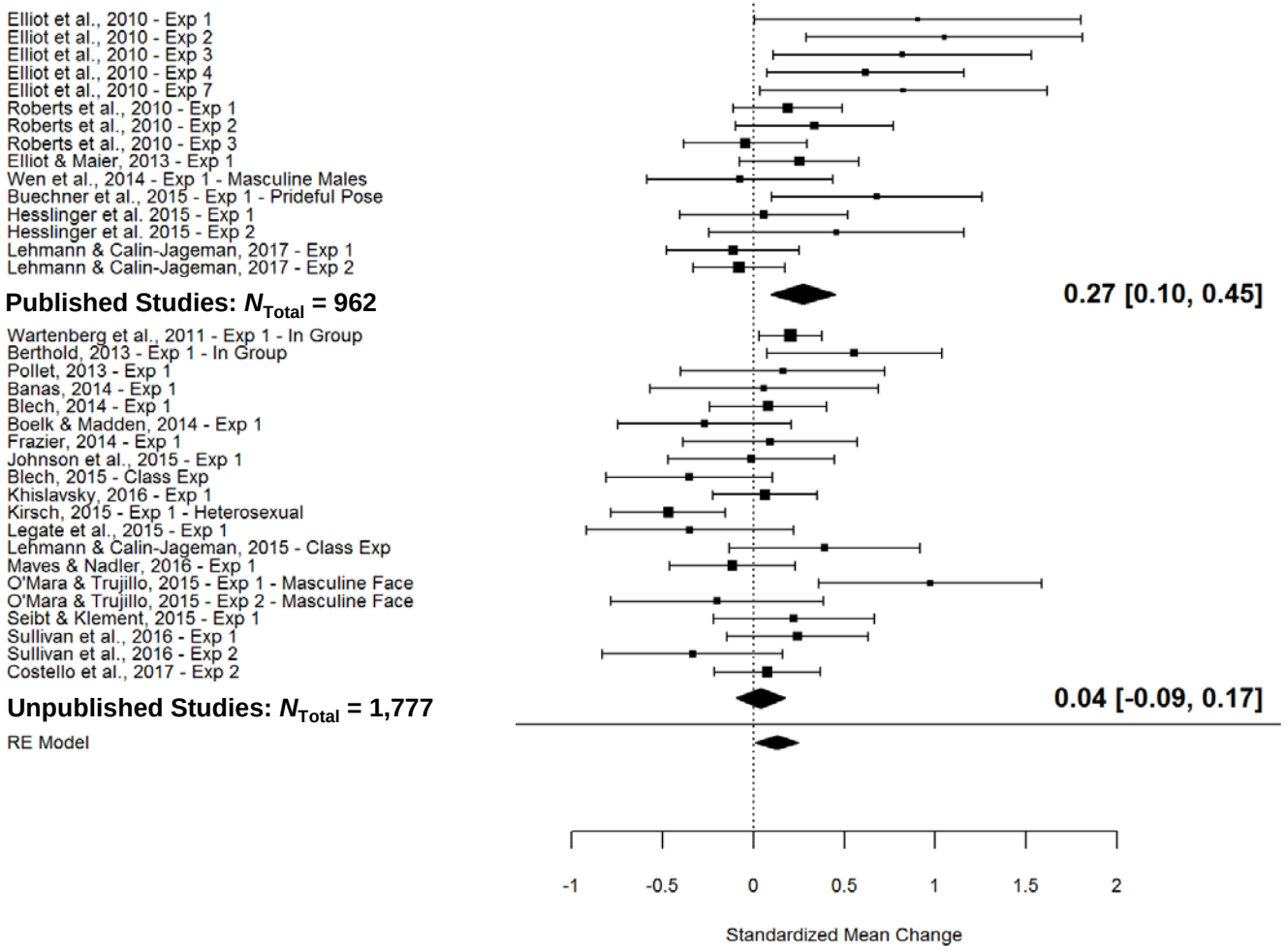


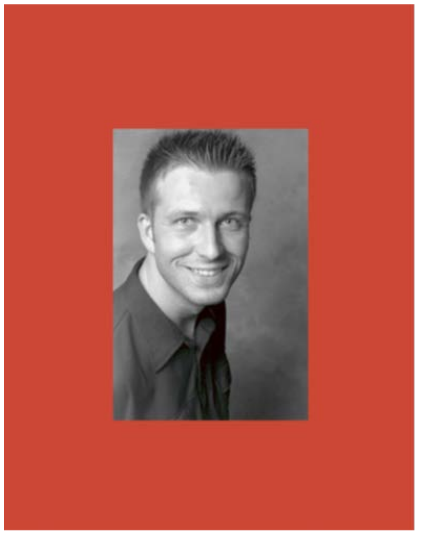
Positive/Negative dichotomy is a license for publication bias.

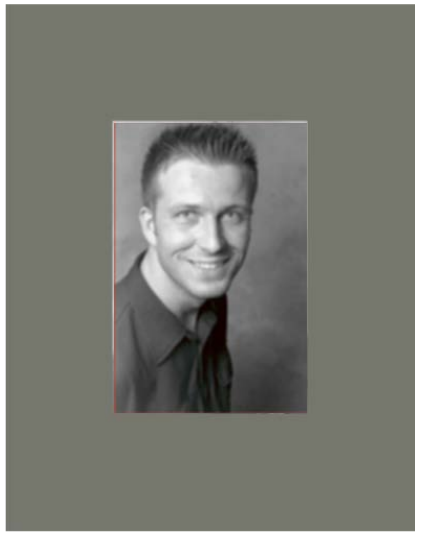
How much does red increase attraction?

Lehmann, Elliot, & Calin-Jageman, in prep

Dichotomous Thinking Ruins Science







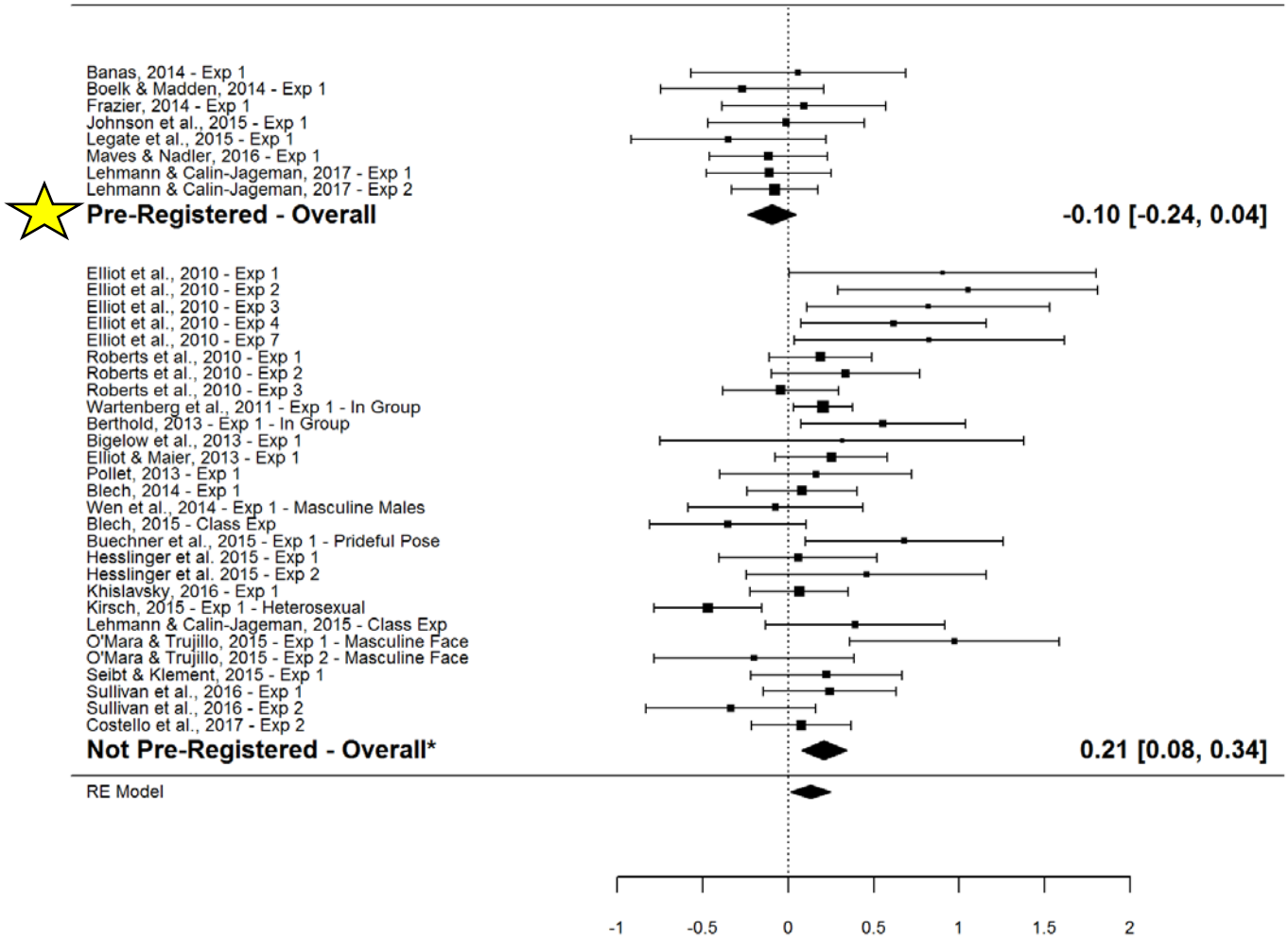
Estimation approach shuns dichotomization. Estimates are estimates.

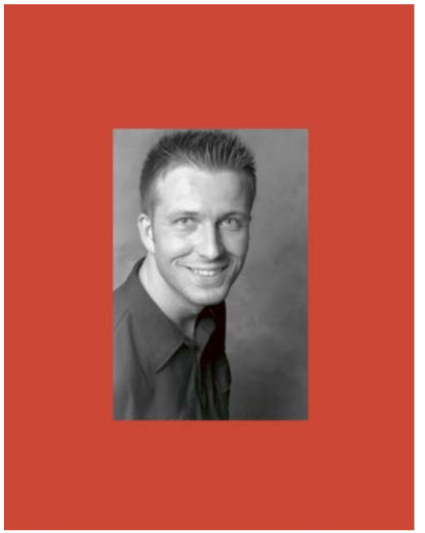
Focus should be on the quality of the estimates (measurement, design, etc.)

How much does red increase attraction?

Lehmann, Elliot, & Calin-Jageman, in prep

Dichotomous Thinking Ruins Science





Estimation approach shuns dichotomization. Estimates are estimates.

Focus should be on the quality of the estimates (measurement, design, etc.)

How much does red increase attraction?

Lehmann, Elliot, & Calin-Jageman, in prep

Overview

- What is the New Statistics?
- Better inference with the New Statistics
- **There are no panaceas**

Only peoples. Peoples is peoples. No is buildings. Is tomatoes, huh?
Is peoples, is dancing, is music, is potatoes. So, peoples is peoples. Okay?
- Pete, Muppets Take Manhattan

There are no panaceas…

- Issues related to reproducibility are not merely statistical
  - Structural issues (incentives, growth, etc.)
  - Measurement and design issues
  - Theory/model development and cumulative science

- Also, science is hard. Good judgement is always required, but often lacking.
  - It is easy to report CIs but maintain dichotomous thinking

- Still
  - We think estimation is a step in the right direction
  - We think research into statistical cognition can help develop statistical practices that will better support good judgement.

Where we stand on some key issues…

- *p* values:
  - Not needed when CI is given; too dangerous

- Neglected factors (measurement, design, effect size consideration, and theory)
  - $Yes_1 … Yes_{1000}$!
  - We think the New Statistics helps put the focus here; critical thinking required regardless of the estimate obtained.

- Additional training in advanced math
  - Cognitive biases aren't dispelled this way

- ASA statement
  - Yes, but now a statement on best practices is needed

ASA Statement on Best Practices in Inference

Pluralism and domain-specificity:

- There's no one right way, but different approaches may be more/less suitable
- Good statistical practice may be different in different domains

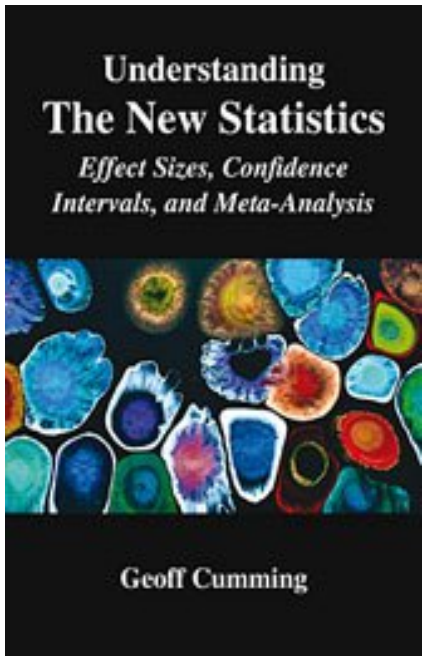
Still, there are some general principles we endorse:

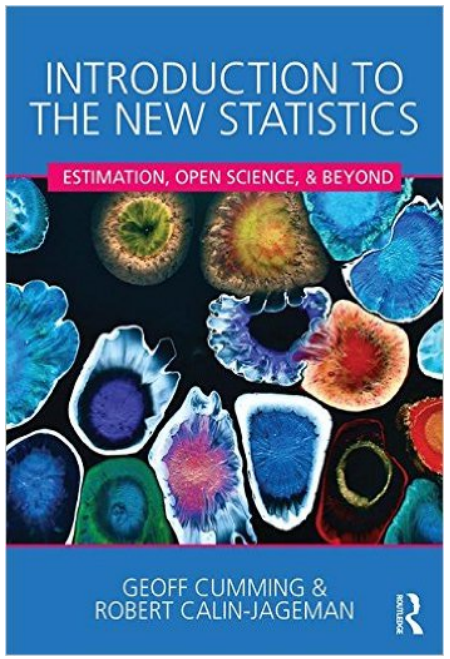
- Look at the data
- Quantify, visualize, report, and interpret uncertainty (there are multiple ways to do this)
- Be complete in your reporting
- Distinguish as clearly as possible planned analyses and predictions from exploratory analyses and post-hoc explanations

Get Started with the New Statistics

https://thenewstatistics.com/

- Getting Started page
  - List of resources, articles, software packages, etc.
- ESCI
- Video tutorials
  - For fellow researchers
  - For students





Thank you

- What is the New Statistics?
- Better inference with the New Statistics
- There are no panaceas

Co-authors or time-lapse photography?

The first principle is that you must not fool yourself – and you are the easiest person to fool.

- Richard Feynman, 1974

It’s not new!

- Neither was “New Math”
- The New is about it being new to standard inferential practice in many scientific domains

Confidence intervals don’t tell you what you want to know

- Use Bayesian approaches if your hung up on this
- I disagree—A CI tells me I’m using a reasonable method but that results may vary and so I need epistemic angst about each individual result that drives me towards replication

We need to make decisions

- You can make decisions based on CIs.
- It is better preserve as much information about uncertainty as possible up to and beyond the point of decision making.
  - We recommend this drug: it improves symptoms by an average of 10% with a margin of error of 1%.
  - We recommend this drug: it improves symptoms by an average of 10% with a margin of error of 9.9%.