Why and How to Replace Statistical Significance Tests with Better Methods

Andreas Schwab
Ivy College of Business
Iowa State University

CARMA Webinar

October 26, 2018

© 2018 Andreas Schwab All Rights Reserved
Research Method Division
Professional Development Workshop or Symposium

Why and How to Replace Statistical Significance Tests with Better Methods

William Starbuck
Andreas Schwab
Eric Abrahamson
Sam Holloway
Chet Miller

Bruce Thompson
Donald Hatfield
Jose Cortina
Raymond Hubbard
Lisa Lambert

Impending methodological changes ... really?
Impending methodological changes … really?

Since 2013
Center of Open Science
(Brian Nosek and Jeffrey Spies)

- Replicated 100 psychology studies
- Replicated ES only half in magnitude
- Statistically significant effects:
  - 97% in original studies
  - 36% in replicated studies

Impending methodological changes... really?

- 2016
  SMJ paper suggests 24 – 40% of published management findings will not replicate

Goldfarb, B., & King, A. A. 2016.
Impending methodological changes ... really?

- 2016 American Statistical Association’s “Statement on p-Values”

Impending methodological changes ... really?

- 2016
  SMJ revised author guidelines
  - ... will no longer publish papers that report cutoff levels of statistical significance.
  - ... requests that authors explicitly discuss and interpret effect sizes.

Bettis at al. (2016).
What is at the core of these method change discussions?

Limitations of Statistical Significance Tests
The limitations of statistical significance tests are “old news” ...

- Fisher proposed statistical significance in 1925
- Solid criticism by leading methodologist including:
  - Gerd Gigerenzer
    Max-Plank Institute
  - Karl Pearson (1857 - 1936)
    University College of London
  - Jacob Cohen (1923 - 1998)
    New York University
  - Herbert Simon (1916 – 2001)
    Carnegie Mellon University
  - Frank Schmidt
    University of Iowa
The limitations of statistical significance tests are “old news”...

However...

- Statistics textbooks teach statistical significance without much reference to these complaints.
- Many scholars remain unaware of the strong arguments against statistical significance tests.
Prevalence of Statistical Significance Tests in Quantitative Empirical Studies

In media res ...

Statistical Significance Tests: What are the problems?
Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests

Andreas Schwab, Iowa State University
Eric Abrahamson, Columbia University
Bill Starbuck, University of Oregon
Fiona Fidler, La Trobe University

2011
Organization Science
22(4), 1105-1120.

Beyond Statistical Significance Webpage
https://sites.google.com/site/nhstresearch/
In media res ...

Statistical Significance Tests: What are the problems?
Problem 1: Statistical significance tests make assumptions that many studies do not satisfy

- Statistical significance estimates the probability for random sampling to have affected observed sample parameters (under assumption that $H_0$ is true)
  - Non-random samples
    => no justification for random-sampling distribution
  - Population data
    => no need for probability of sampling effects

... these applications make no sense!
Problem 2: Nil Hypotheses

- Statistical significance calculates the probability of observing the sample parameter under the condition that the null-hypothesis of “no effect” is true.

- In most cases, we have very limited confidence in $H_0$ being true
  - Absolutely no effect is a highly unlikely hypothesis.
  - Any reasonable intervention will have “some” effect.
  - Hence, we tend to answer a question to which we already know the answer. [no new deep insights]
  - “Thus, only asking ‘Are the effects different’ [from zero] is foolish.” (Tukey, 1991)

... because it is not enough!! (e.g., effect sizes)
Problem 3: Sample-size sensitivity of statistical significance tests

- As sample size increases, confidence intervals shrink and eventually no longer include zero.
  - By increasing sample size, researchers can reject any null-hypothesis.
  - ... even if absolutely no effects because random effects and measurement errors always guarantee some difference from zero.
  - ... for directional hypothesis still 50% success rate guaranteed.
- Hence, test outcomes are known in advance and under the researchers control. (... advanced IT facilitates)
Problem 4: Single fixed significance-level threshold

- Statistical significance levels ($p < .05$)
  - What if outcomes are extremely negative or positive?
  - "Cliff effects" amplify very small differences in the data into very large differences in implications. (Simonsohn et al. 2013; Rosnow and Rosenthal, 1989)
  - Instability of p-values facilitates "cliff effects" as small changes in sample can substantially affect p-values. (Gelman & Stern, 2006)
"Cliff Effects"


- We know this distribution cannot possibly be true.
- The only “Positive” of Cliff Effects: Reveals degree of research manipulation across studies.

O’Boyles et al. (2016)
A Meta-Analysis of Interactions in Management and Applied Psychology Research.
How easy is it to find statistical significant correlations between variables in management research?

- Webster & Starbuck (1988)
  - In management research, the average correlation between unrelated variables is not zero but 0.09.
  - When choosing two variables at random, statistical significance offers a 67% chance of significant findings on the first try, and a 96% chance with three tries for average reported sample sizes.

- Simmons, Nelson & Simonsohn (2011)
  - Current flexibilities in data collection, analysis and reporting increase false-positive rates.
  - Statistical significance (p<.05) encourages a high number of false positive
Second-order consequences: “Significant” findings often do not replicate

- “Statistically significant findings” may not replicate! (Open Science Project, 2015; Goldfarb & King, 2015)
- Management research does not conduct and publish replication studies
- Management research has accumulated lots of empirical evidence, but it is unclear what findings are real.
Problem 5: Statistical significance tests encourage dichotomous thinking

- Statistical significance portrays results as dichotomous and definite
  - Either reject or fail to reject the null hypothesis
- No explicit discussing and reporting of detailed uncertainty information in research reports
Problem 6: No probability estimate for the hypothesis of interest ($H_1$)

- We reject the Null-Hypothesis and conclude that the proposed hypothesis is the only alternative explanation.

- We have no direct probability statement if $H_1$ is true based on the observed data. (*inverse probability fallacy*)
Problem 6: No probability estimate for the hypothesis of interest (H₁)

- Statistical significance provides
  - Probability of observing data assuming null-hypothesis is true
    \[ \text{Pr(data|H₀)} \] (inverse probability fallacy)

- Question of interest:
  - Probability of proposed hypothesis being true given the observed data
    \[ \text{Pr(H₁|data)} \]

- The p-value provides no probability information about H₁ or H₀ being true (Bayesian approaches can!)
Overall effects on management research

- Management researchers spent substantial time and effort to collect empirical data to conduct statistical significance tests.
- Limitations of statistical significance practices described have contributed to the accumulation of published empirical findings heavily contaminated with false-positive findings.
- These practices prevent scientific progress and practical impact of management research.
... but here is the good news!!

"We have met the enemy and he is us"

... and we can do better!

Pogo (Walt Kelly) 1971
Statistical Significance Tests impede scientific progress

How can we do better?
Beyond Statistical Significance

Effect Sizes
1: Focus on effects size measures and tailor them to research contexts

What metrics make sense for dependent variables?

- Describe effects in the meaningful units used to measure dependent variables – number of tons, people, bales, barrels or dollars earned.

- Example: Nutrition training of nurses
  - Number of correct answers on knowledge tests
  - Different measures for different outcomes and stakeholders (nurses, patients, hospitals, society)
  - Complementary measures (e.g., benefits and costs)
1: Focus on effects size measures and tailor them to research contexts

- For standardized scales:
  Change in DV associated with 1 SD change in IV

- Other general effect size measures (Cumming, 2011)
  - Pearson’s r, Δ odds, Δ R² (more familiar)
  - Cohen’s q, f² (variance-explained)
  - Cohen’s d, Glass’ Δ, Hedges’ g, ω² (omega-sq), η² (eta-sq) (mean differences)
  - Cohen’s w and h, relative risk, Δ risk (categorical DV)
Recommended literature

Understanding the New Statistics: Effect sizes, Confidence Intervals, and Meta-Analysis.

Geoff Cumming
La Trobe University
2011

Free Video Lecture Series
Recommended literature

1: Focus on effects size measures and tailor them to research contexts

Stop talking about statistical significance

Report and discuss effect sizes
Beyond Statistical Significance

Effect Uncertainty
2: Report the uncertainty associated with measures of effects

- Effect size and effect uncertainty crucially important for theory development.
- Emerging emphasis for effect size evaluations (e.g., SMJ author guidelines)
- ... current statistical significance tests truncate explicit evaluation and modeling of effect uncertainty.
2: Report the uncertainty associated with measures of effects

- Report uncertainty of effects using measures of variability
  - CI, SD, Ranges, quartiles ...

- Show graphs of complete distributions – say, the probability distribution of effect sizes.
2: Report the uncertainty associated with measures of effects

Effects of training on nurses’ knowledge about nutrition

- Null-Hypothesis test showed “statistically significant” effect of training, but it did not show how much knowledge changed.

This simple graph tells more:
- Average improvement
- Bi-model post-training performance distribution
2: Report the uncertainty associated with measures of effects

Effects of training on nurses’ knowledge about nutrition

This graph tells even more:
- Some nurses knew less after training. Why?

Performance of some nurses decreased!
2: Report the uncertainty associated with measures of effects

Current Advancement in Graph Design

- Simple and powerful software tools
- Animated and interactive graphs
- Move toward completely digital journals
2: Report the uncertainty associated with measures of effects

- The investigation and evaluation of both effect size and effect uncertainty crucially important for theory development and management advice.  

“We have to discuss, model and breath uncertainty because it is inherent in what we study.”

John W. Tukey
Beyond Statistical Significance

Baseline Models
3: Compare new data with meaningful baselines rather than no-effect hypotheses

- Alternative baseline models:
  - Alternative treatments, competing explanations
    Example: Is the new training program better than the old
  - Simple random processes
    Example: Organizational survival as a random walk.
  - Crude stability or momentum processes.
    Example: Tomorrow will be similar to today
    Yesterday’s trends will continue today
3: Compare new data with baseline models rather than null hypotheses.

Use of baseline models in academic papers.
Beyond Statistical Significance

Bayesian Statistics
4: Can Bayesian statistics help?

- Revisit: Statistical significance provides conditional probability statement of limited value.
  - Probability of observing data assuming null-hypothesis is true
    \[ \Pr(\text{data}|\ H_0) \] (inverse probability fallacy)

- Question of interest:
  - Probability of proposed hypothesis being true given the observed data
    \[ \Pr(\ H_1|\ \text{data}) \]

- Bayesian approaches try to answer the relevant questions!
4: Can Bayesian statistics help?

- Bayesian statistics offer an established methodology for drawing inferences and assessing plausibility of theories. (needs understanding for meaningful applications)

- Bayesian analysis advantages
  - Bayesian analyses estimate posterior distributions of hypothesized effects based on prior distribution and observed data
  - Posterior distributions enable evaluation of both the strength and uncertainty of hypothesized effects.
2: Report the uncertainty associated with measures of effects

Bayesian Posterior Distribution of M&A Effects on Firm Performance

19% probability of decreasing performance.

Posterior distributions enable powerful interpretations of the empirical data.

81% probability of improving firm performance.

... but this graph also conveys a wealth of other information about effect size and probability.
- Mean
- Range
- Functional form
- Thickness of tails

Available software and faster computers have made numerical methods practical

<table>
<thead>
<tr>
<th>Software</th>
<th>Google Search</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>WinBugs</td>
<td>“winbugs”</td>
<td>Low-level programming language: Free software from University of Cambridge. Supports bootstrapping.</td>
</tr>
<tr>
<td>BugsXLA</td>
<td>“BugsXLA”</td>
<td>Free add-in for Excel. Provides Excel interface for WinBugs analyses of linear models: Available from University of Cambridge or philwoodward.co.uk.</td>
</tr>
<tr>
<td>Stata</td>
<td>“Stata”</td>
<td>Many universities have site licenses. Supports bootstrapping.</td>
</tr>
<tr>
<td>SAS/ STAT</td>
<td>“SAS”</td>
<td>Many universities have site licenses.</td>
</tr>
<tr>
<td>R</td>
<td>“R-project”</td>
<td>Open-source software from the R-project. Low-level programming language.</td>
</tr>
<tr>
<td>SPSS</td>
<td>“SPSS”</td>
<td>Many universities have site licenses.</td>
</tr>
</tbody>
</table>

(adopted from Starbuck, 2017)
Promote and support methodological change

How can you help?
How you can help!

- Improve your own work
  - Report and discuss effect sizes (ES)
  - Report and discuss uncertainty of effects (e.g., CI and graphs of effect distributions)
  - Evaluate hypotheses based on "substantive" or "practically" importance using reasonable baselines
  - Avoid rituals, instead carefully account for research question, design and empirical context in your analyses and interpretations of data.
How you can help!

- Please speak up in seminars or as a reviewer
  - When researchers miss apply or misinterpret statistical significance tests.
  - When researchers do not report and discuss effect sizes and effect uncertainty.
  - Insist that studies consider multiple perspectives, baselines, research context and design in their interpretation of empirical data. (“dig deeper!”)
... and support your colleagues when they raise such issues!
Method changes are coming!

- We are experiencing early stages of some fundamental methodological changes in our field.
- We should embrace these changes.
- They are opportunities to have impact and for our work to make more of a difference.
Why and How to Replace Statistical Significance Tests with Better Methods

Andreas Schwab
Iowa State University

CARMA Webinar
October 26, 2018

© 2018 Andreas Schwab All Rights Reserved