Robust tests of theory with randomly sampled experiments

Joachim Vandekerckhove, Beth Baribault, Chris Donkin, Daniel R. Little, Jennifer Trueblood, Zita Oravecz, Don van Ravenzwaaij, Corey White

University of California, Irvine

and various other institutions of great renown

Reproducibility

Reproducibility...

Reproducibility...

– ... is widely considered to be in crisis (Baker, 2016)

Reproducibility...

- ... is widely considered to be in crisis (Baker, 2016)
- ... is the topic of a National Academy of Sciences review (Cicerone, 2015)

Reproducibility...

- ... is widely considered to be in crisis (Baker, 2016)
- ... is the topic of a National Academy of Sciences review (Cicerone, 2015)
- ... is now one of the fundamental review criteria for all grants submitted to the National Institutes of Health

Reproducibility...

- ... is widely considered to be in crisis (Baker, 2016)
- ... is the topic of a National Academy of Sciences review (Cicerone, 2015)
- ... is now one of the fundamental review criteria for all grants submitted to the National Institutes of Health
- ... has been discussed in Congressional hearings

Reproducibility...

- ... is widely considered to be in crisis (Baker, 2016)
- ... is the topic of a National Academy of Sciences review (Cicerone, 2015)
- ... is now one of the fundamental review criteria for all grants submitted to the National Institutes of Health
- ... has been discussed in Congressional hearings
- ... "threatens the entire biomedical research enterprise" (Rosenblatt, 2016).

Reproducibility...

Poor reproducibility of empirical findings, in a number of areas, is at this point sufficiently established. Much of the literature is suspect.

Reproducibility...

Poor reproducibility of empirical findings, in a number of areas, is at this point sufficiently established. Much of the literature is suspect.

There is broad support for reproducibility-boosting corrective measures (Baker, 2016):

Reproducibility...

Poor reproducibility of empirical findings, in a number of areas, is at this point sufficiently established. Much of the literature is suspect.

There is broad support for reproducibility-boosting corrective measures (Baker, 2016):

- Better statistics

Reproducibility...

Poor reproducibility of empirical findings, in a number of areas, is at this point sufficiently established. Much of the literature is suspect.

There is broad support for reproducibility-boosting corrective measures (Baker, 2016):

- Better statistics
- Internal and external validation of research results

Reproducibility...

Poor reproducibility of empirical findings, in a number of areas, is at this point sufficiently established. Much of the literature is suspect.

There is broad support for reproducibility-boosting corrective measures (Baker, 2016):

- Better statistics
- Internal and external validation of research results
- Robust experimental designs

Robustness

Robustness

When empirical effects fail to replicate, we often hear a mantra of hidden moderators:

Robustness

When empirical effects fail to replicate, we often hear a mantra of hidden moderators:

"Well, but . . ."

Robustness

When empirical effects fail to replicate, we often hear a mantra of hidden moderators:

"Well, but . . ."

- ". . . your study took place in Amsterdam; ours in Nijmegen."

Robustness

When empirical effects fail to replicate, we often hear a mantra of hidden moderators:

"Well, but . . ."

- ". . . your study took place in Amsterdam; ours in Nijmegen."
- ". . . your study took place in 2017; ours in 1999."

Robustness

When empirical effects fail to replicate, we often hear a mantra of hidden moderators:

"Well, but . . ."

- ". . . your study took place in Amsterdam; ours in Nijmegen."
- ". . . your study took place in 2017; ours in 1999."

I would question the societal value of scientific claims that do not generalize beyond 1999 Nijmegen.

Robustness

The concept of hidden moderators is as old as science.

Robustness

The concept of hidden moderators is as old as science.

nial cinnabar: for though in our furnaces it hath been very ſucceſsfully made, yet not only we have afterwards failed of making it, but we have ſeen much more expert chymiſts, and who, becauſe of the high value they do (not undeſervedly) place upon that medicine, imploy themſelves oftener than we in making it, divers times unſucceſsfully attempt the preparing it. And it may be perhaps alſo from ſome diverſity either in antimonies or irons, that eminent chymiſts have (as we have obſerved) often failed in their endeavours to make the ſtarry regulus of *Mars* and antimony. Inſomuch that divers artiſts fondly believe and teach (what our experience will not permit us to allow) that there is a certain reſpect to times and conſtellations requiſite to the producing of this (I confeſs admirable) body. Upon which ſubject I muſt

Boyle (1772), on the suspected importance of astrological time to chemical purification of antimony.

Terminology

Terminology

- Replicability: How can we ensure that other people can repeat the study procedures we used?

Terminology

- Replicability: How can we ensure that other people can repeat the study procedures we used?
- Reproducibility: How can we ensure that other people can repeat the analysis we performed?

Terminology

- Replicability: How can we ensure that other people can repeat the study procedures we used?
- Reproducibility: How can we ensure that other people can repeat the analysis we performed?
- Robustness: How can we ensure that other people obtain the same findings?

Terminology

- Replicability: How can we ensure that other people can repeat the study procedures we used?
- Reproducibility: How can we ensure that other people can repeat the analysis we performed?
- Robustness: How can we ensure that other people obtain the same findings?
- Generalizability: How can we ensure that our results translate to very different contexts?

Terminology

- Replicability: How can we ensure that other people can repeat the study procedures we used? → Open materials
- Reproducibility: How can we ensure that other people can repeat the analysis we performed? → Code hygiene; open data
- Robustness: How can we ensure that other people obtain the same findings? → ? Robust design?
- Generalizability: How can we ensure that our results translate to very different contexts? → ? Higher abstraction?

Robustness

Robustness is borrowed from a number of other fields:

Robustness

Robustness is borrowed from a number of other fields:

- Robust statistics [have] good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal

Robustness

Robustness is borrowed from a number of other fields:

- Robust statistics [have] good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal
- Robust programming is a style of programming that focuses on handling . . . unexpected actions. It requires code to handle these . . . actions gracefully

Robustness

Robustness is borrowed from a number of other fields:

- Robust statistics [have] good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal
- Robust programming is a style of programming that focuses on handling . . . unexpected actions. It requires code to handle these . . . actions gracefully
- [In engineering,] Robustness is the state where the technology, product, or process performance is minimally sensitive to factors causing variability (either in the manufacturing or users environment)

Robustness

What constitutes “robust design” in that incremental list of requirements?

Robustness

What constitutes “robust design” in that incremental list of requirements?

- A low bar for robustness would seem to be “extend to a future repetition of an identical procedure”

Robustness

What constitutes "robust design" in that incremental list of requirements?

- A low bar for robustness would seem to be "extend to a future repetition of an identical procedure"
- A slightly higher (but still low) bar would be "extend to a procedure that only differs in irrelevant ways" (e.g., lab)

Robustness

What constitutes "robust design" in that incremental list of requirements?

- A low bar for robustness would seem to be "extend to a future repetition of an identical procedure"
- A slightly higher (but still low) bar would be "extend to a procedure that only differs in irrelevant ways" (e.g., lab)
- One way of thinking about this is to imagine a population of largely interchangeable experiments

Robustness

What constitutes "robust design" in that incremental list of requirements?

- A low bar for robustness would seem to be "extend to a future repetition of an identical procedure"
- A slightly higher (but still low) bar would be "extend to a procedure that only differs in irrelevant ways" (e.g., lab)
- One way of thinking about this is to imagine a population of largely interchangeable experiments
- We have methods to license statements about populations: sampling

Randomized design

Randomized design and the metastudy

When we perform manipulations to elicit experimental effects, we expect that effect to extend to other situations – but which ones?

Randomized design and the metastudy

When we perform manipulations to elicit experimental effects, we expect that effect to extend to other situations – but which ones?

Three independent groups invoke similar sounding metaphors:

- Universe of generalization (Cronbach et al., 1963)
- Constraints of generality (Simons et al., 2017)
- Boundary of meaning (Kenett & Rubinstein, 2017)

Randomized design and the metastudy

When we perform manipulations to elicit experimental effects, we expect that effect to extend to other situations – but which ones?

Three independent groups invoke similar sounding metaphors:

- Universe of generalization (Cronbach et al., 1963)
- Constraints of generality (Simons et al., 2017)
- Boundary of meaning (Kenett & Rubinstein, 2017)

These metaphors imply the existence of some spatially-arranged population of possible experiments where the effect holds.

Randomized design and the metastudy

When we perform manipulations to elicit experimental effects, we expect that effect to extend to other situations – but which ones?

Three independent groups invoke similar sounding metaphors:

- Universe of generalization (Cronbach et al., 1963)
- Constraints of generality (Simons et al., 2017)
- Boundary of meaning (Kenett & Rubinstein, 2017)

These metaphors imply the existence of some spatially-arranged population of possible experiments where the effect holds.

To generalize across this population of experiments, we could sample from it.

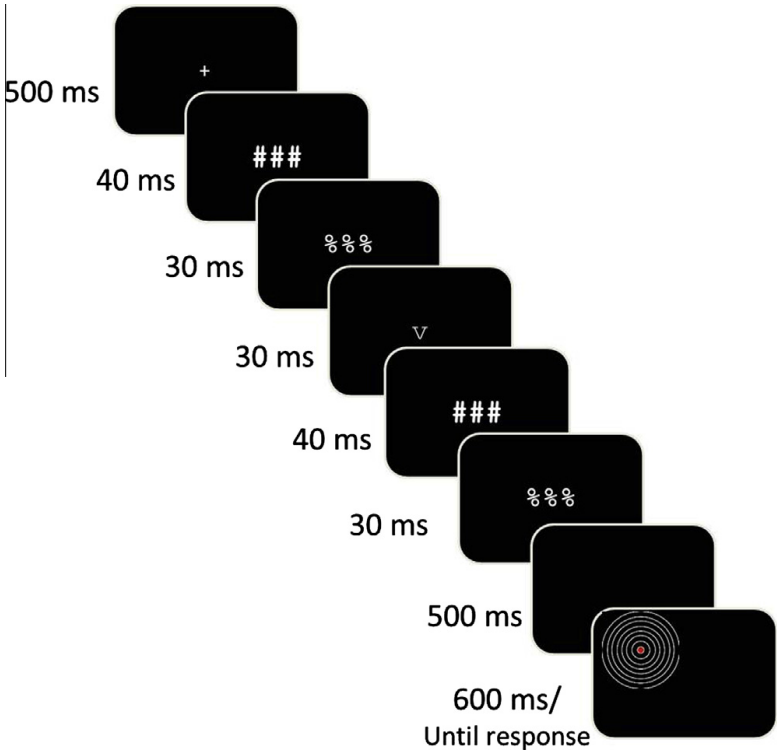
Randomized design and the metastudy

Start with a basic experiment:

Randomized design and the metastudy

Start with a basic experiment:

- Reuss et al. (2015) report on a subliminal cueing effect

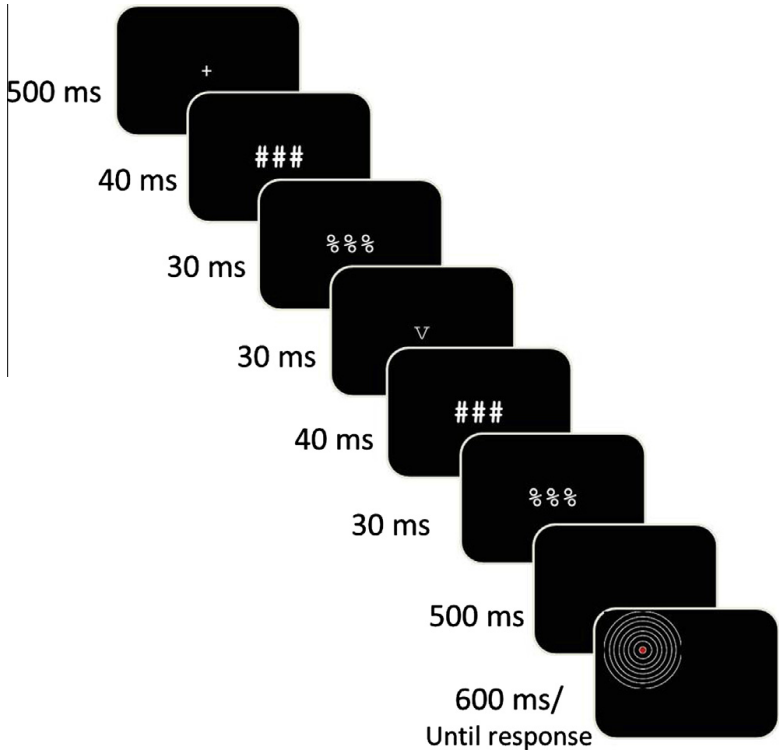


Time course of one trial.

Randomized design and the metastudy

Start with a basic experiment:

- Reuss et al. (2015) report on a subliminal cueing effect
- The cue ƀ or ⩛ tells the participant to favor speed or accuracy in the task

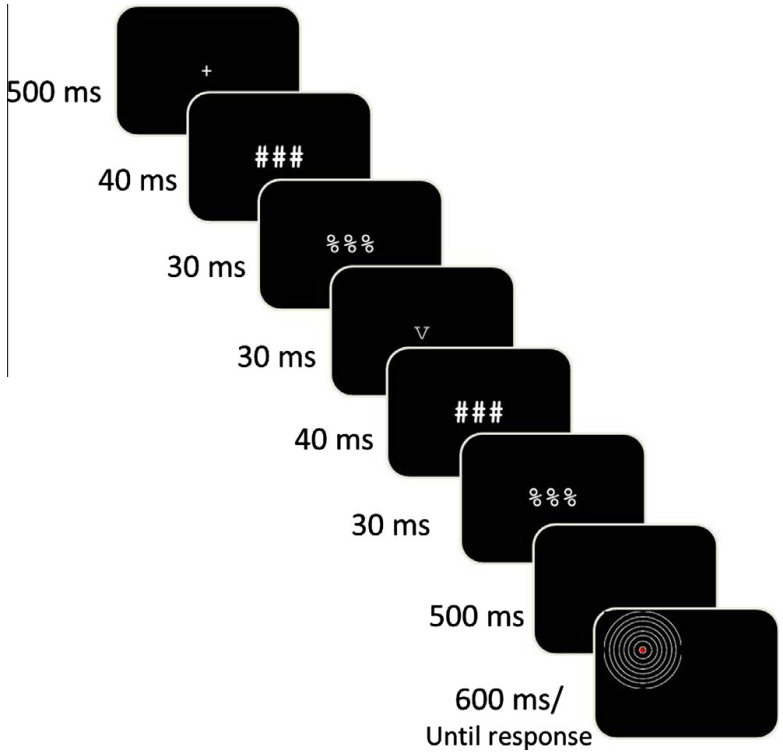


Time course of one trial.

Randomized design and the metastudy

Start with a basic experiment:

- Reuss et al. (2015) report on a subliminal cueing effect
- The cue ɓ or ⱱ tells the participant to favor speed or accuracy in the task
- When the cue is masked, it becomes subliminal

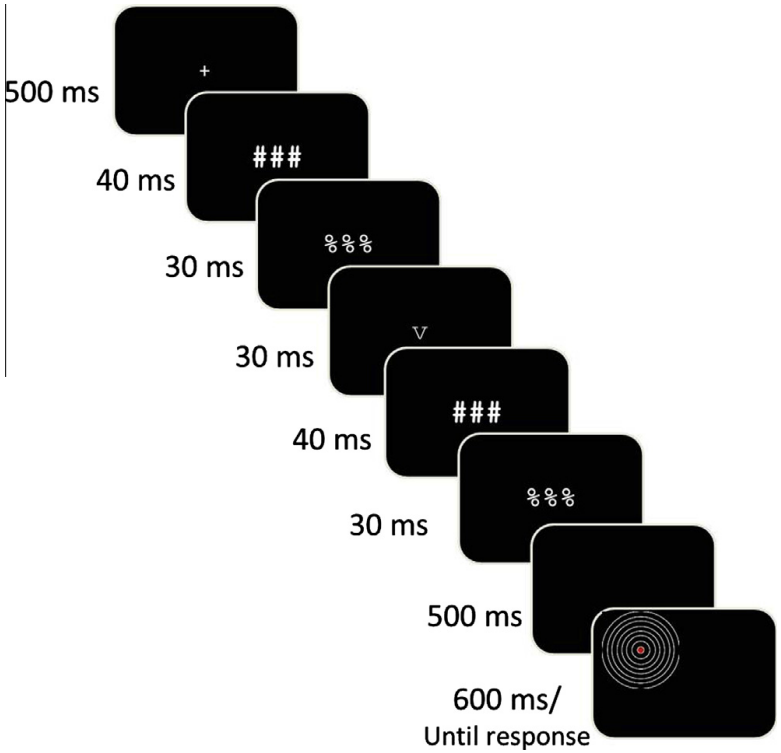


Time course of one trial.

Randomized design and the metastudy

Start with a basic experiment:

- Reuss et al. (2015) report on a subliminal cueing effect
- The cue ƀ or ⩛ tells the participant to favor speed or accuracy in the task
- When the cue is masked, it becomes subliminal
- Even so, participants changed their speed-accuracy trade-off



Time course of one trial.

Radical randomization and the metastudy

Identify features of the experiment that a reasonable colleague might or would have implemented differently:

- Characters used

Radical randomization and the metastudy

Identify features of the experiment that a reasonable colleague might or would have implemented differently:

- Characters used
- Color of the stimuli

Radical randomization and the metastudy

Identify features of the experiment that a reasonable colleague might or would have implemented differently:

- Characters used
- Color of the stimuli
- Brightness of the stimuli

Radical randomization and the metastudy

Identify features of the experiment that a reasonable colleague might or would have implemented differently:

- Characters used
- Color of the stimuli
- Brightness of the stimuli
- Identity of the participants

Radical randomization and the metastudy

Identify features of the experiment that a reasonable colleague might or would have implemented differently:

- Characters used
- Color of the stimuli
- Brightness of the stimuli
- Identity of the participants
- ...

Radical randomization and the metastudy

We can identify many such features that are theoretically inert – and some that are not.

Radical randomization and the metastudy

We can identify many such features that are theoretically inert – and some that are not.

These "theoretically inert features" of an experiment are largely ones that could go lost in the translation from a verbal claim to an experimental design.

Radical randomization and the metastudy

We can identify many such features that are theoretically inert – and some that are not.

These "theoretically inert features" of an experiment are largely ones that could go lost in the translation from a verbal claim to an experimental design.

These independent-variable dimensions span a space that we'll call the method space.

Radical randomization and the metastudy

We can identify many such features that are theoretically inert – and some that are not.

These "theoretically inert features" of an experiment are largely ones that could go lost in the translation from a verbal claim to an experimental design.

These independent-variable dimensions span a space that we'll call the method space.

Theoretical statements often correspond to functions over the method space (e.g., an effect size is nonzero over some region).

Radical randomization and the metastudy

We can identify many such features that are theoretically inert – and some that are not.

These "theoretically inert features" of an experiment are largely ones that could go lost in the translation from a verbal claim to an experimental design.

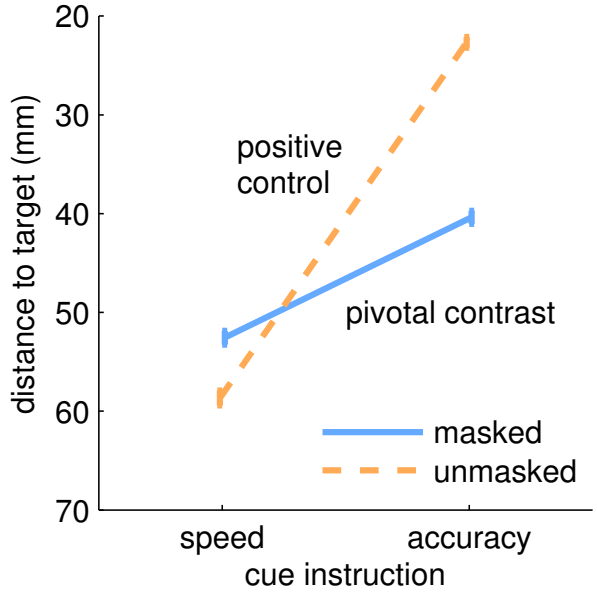
These independent-variable dimensions span a space that we'll call the method space.

Theoretical statements often correspond to functions over the method space (e.g., an effect size is nonzero over some region).

Experiments are points or small areas in that space.

Radical randomization and the metastudy

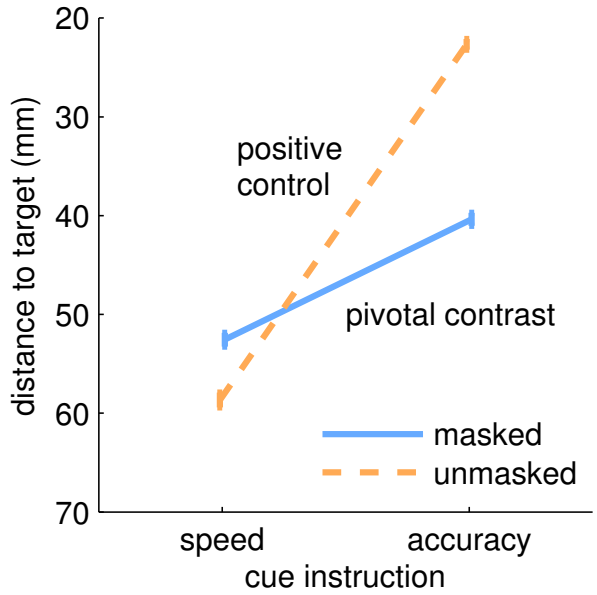
– Now we sample from these points and conduct one micro-experiment ($n = 32$) for each location



Each micro-experiment has 32 trials in a 2×2 design followed by 32 trials of a cue identification task.

Radical randomization and the metastudy

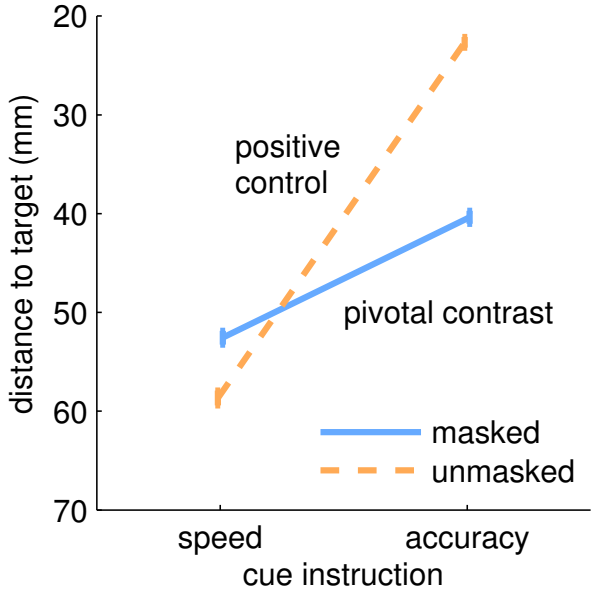
- Now we sample from these points and conduct one micro-experiment ($n = 32$) for each location
- Often these are exchangeable for our theoretical purposes



Each micro-experiment has 32 trials in a 2×2 design followed by 32 trials of a cue identification task.

Radical randomization and the metastudy

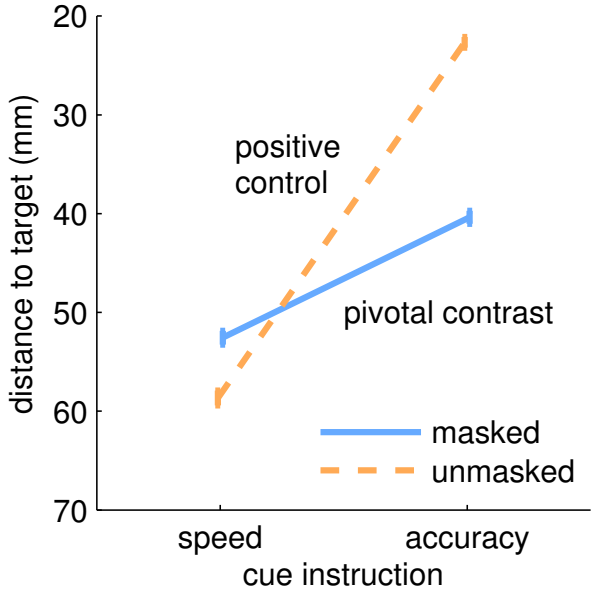
- Now we sample from these points and conduct one micro-experiment ($n = 32$) for each location
- Often these are exchangeable for our theoretical purposes
- Then jointly analyze in a "planned" meta-analysis



Each micro-experiment has 32 trials in a 2×2 design followed by 32 trials of a cue identification task.

Radical randomization and the metastudy

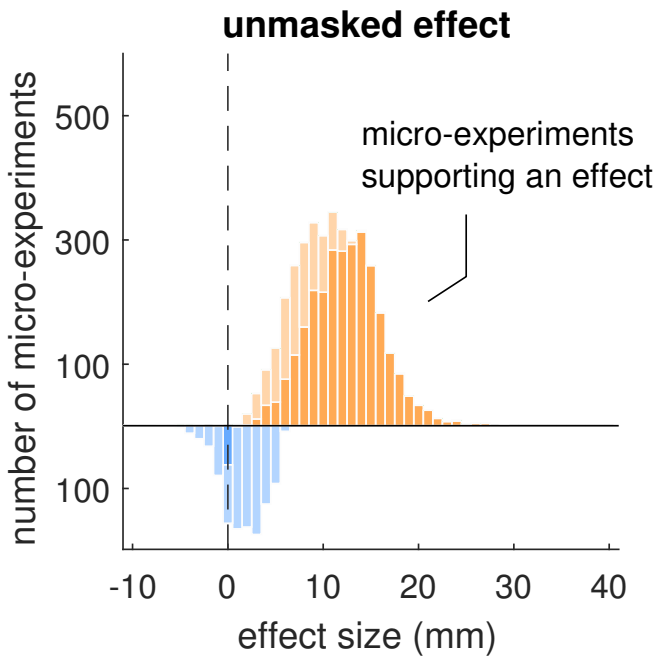
- Now we sample from these points and conduct one micro-experiment ($n = 32$) for each location
- Often these are exchangeable for our theoretical purposes
- Then jointly analyze in a "planned" meta-analysis
- We should expect to see some level-2 variability



Each micro-experiment has 32 trials in a 2×2 design followed by 32 trials of a cue identification task.

Planned meta-analysis

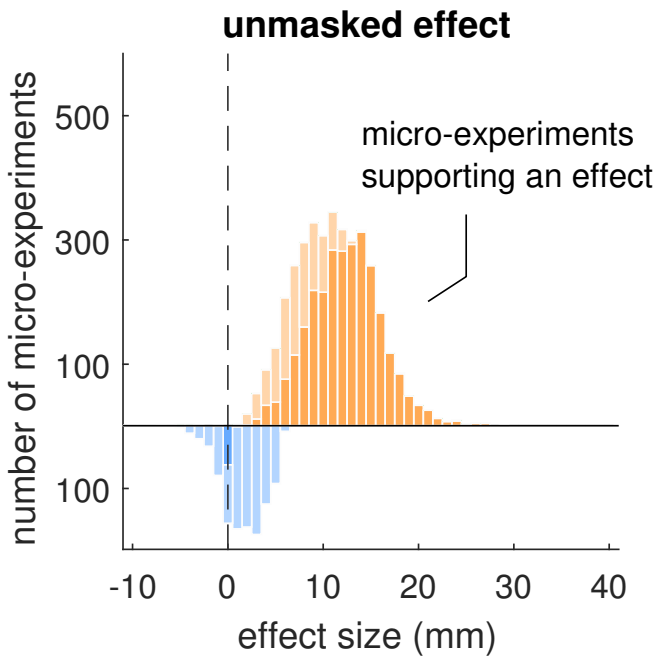
– Focusing only on unmasked trials, the effect seems reasonably robust



Histogram of effect sizes Δmm (unmasked condition).

Planned meta-analysis

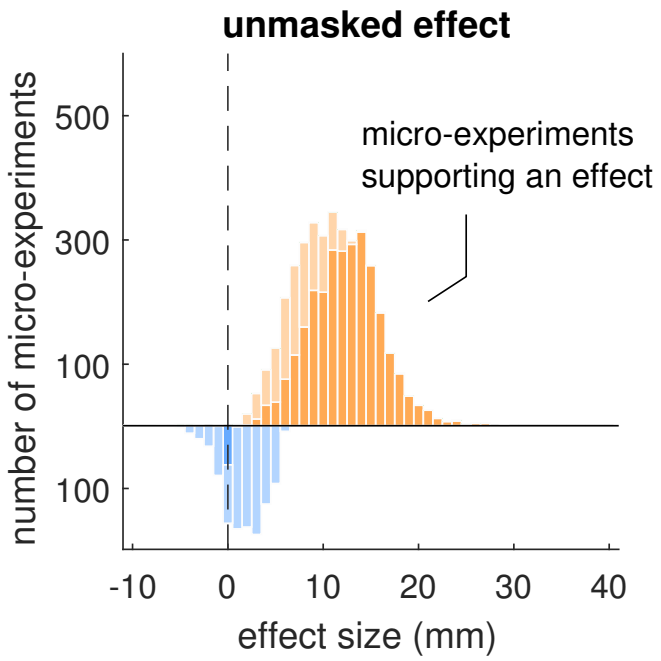
- Focusing only on unmasked trials, the effect seems reasonably robust
- Participants are 10-15mm more accurate after an accuracy prompt



Histogram of effect sizes Δmm (unmasked condition).

Planned meta-analysis

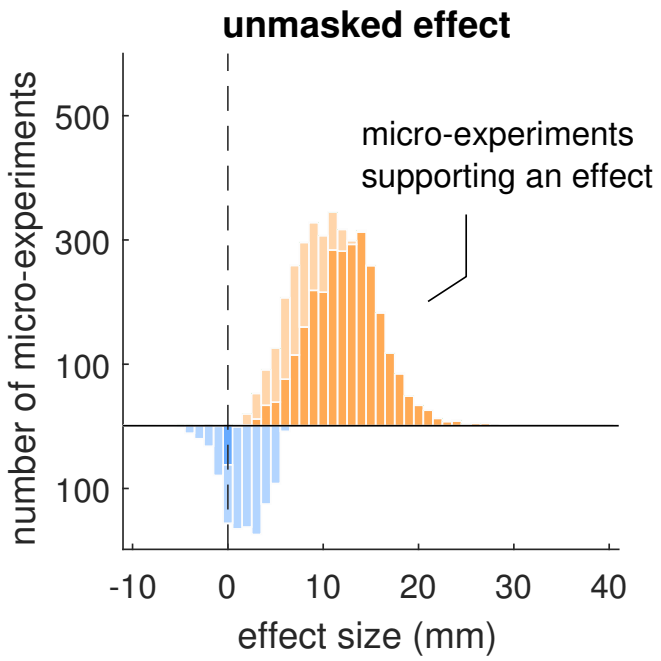
- Focusing only on unmasked trials, the effect seems reasonably robust
- Participants are 10-15mm more accurate after an accuracy prompt
- 78% of micro-experiments show evidence for the effect



Histogram of effect sizes Δmm (unmasked condition).

Planned meta-analysis

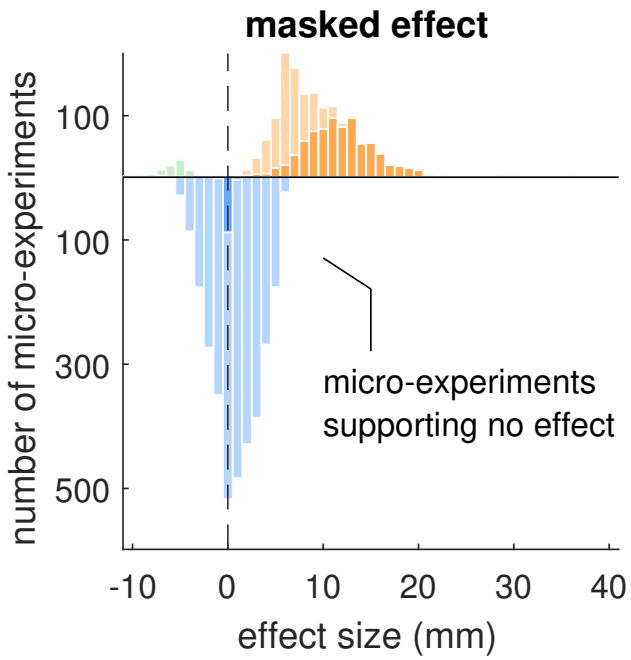
- Focusing only on unmasked trials, the effect seems reasonably robust
- Participants are 10-15mm more accurate after an accuracy prompt
- 78% of micro-experiments show evidence for the effect
- This is not surprising since this is just the positive control



Histogram of effect sizes Δmm (unmasked condition).

Planned meta-analysis

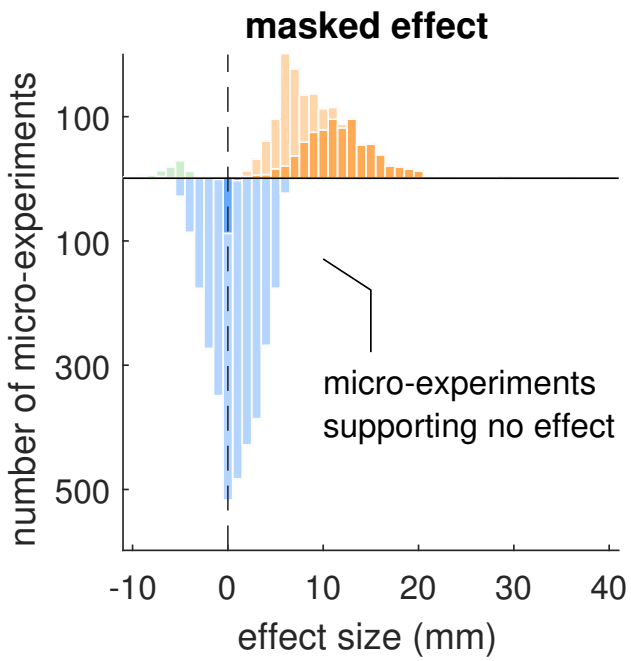
– Focusing only on masked trials, the effect seems fickle



Histogram of effect sizes Δmm (masked condition).

Planned meta-analysis

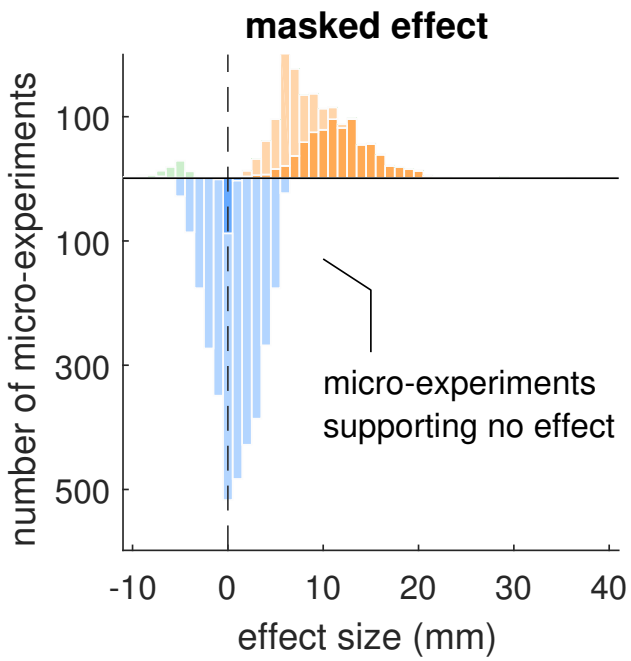
- Focusing only on masked trials, the effect seems fickle
- 75% of micro-experiments are more consistent with an effect of 0mm than 10mm



Histogram of effect sizes Δmm (masked condition).

Planned meta-analysis

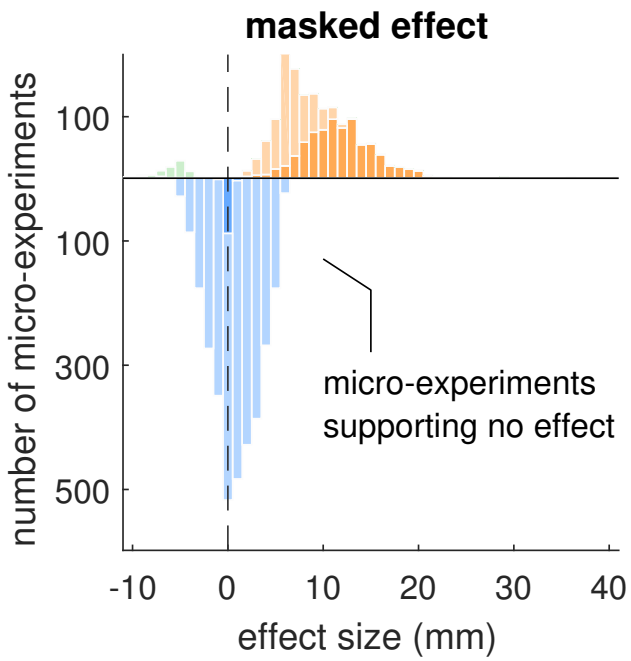
- Focusing only on masked trials, the effect seems fickle
- 75% of micro-experiments are more consistent with an effect of 0mm than 10mm
- Lots of level-2 variability



Histogram of effect sizes Δmm (masked condition).

Planned meta-analysis

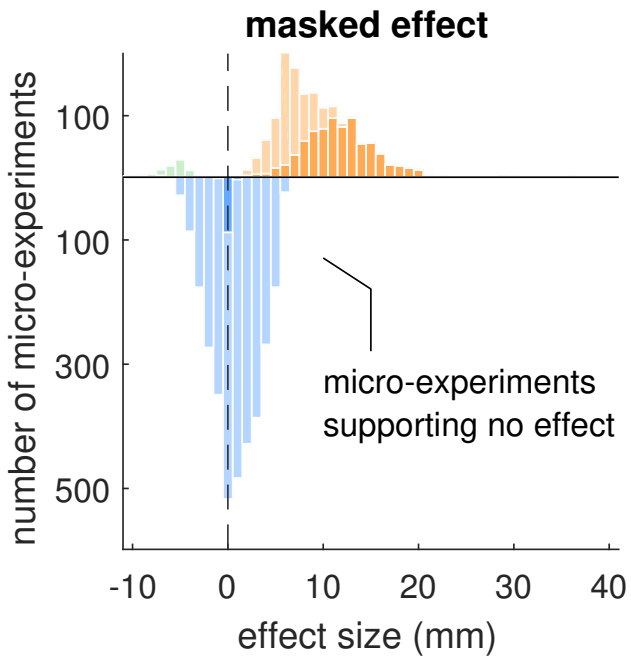
- Focusing only on masked trials, the effect seems fickle
- 75% of micro-experiments are more consistent with an effect of 0mm than 10mm
- Lots of level-2 variability
- Does not support a robust subliminal cueing effect



Histogram of effect sizes Δmm (masked condition).

Planned meta-analysis

- Focusing only on masked trials, the effect seems fickle
- 75% of micro-experiments are more consistent with an effect of 0mm than 10mm
- Lots of level-2 variability
- Does not support a robust subliminal cueing effect
- But there might be an effect under limited conditions



Histogram of effect sizes Δmm (masked condition).

Planned meta-regression

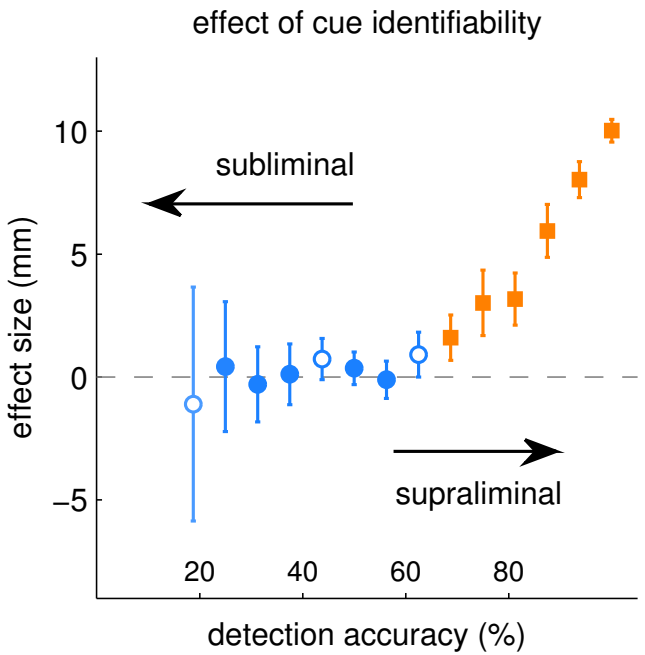
– As a manipulation check, we measured cue identifiability under the unique conditions of each micro-experiment

Planned meta-regression

- As a manipulation check, we measured cue identifiability under the unique conditions of each micro-experiment
- We can now introduce cue detection performance as a covariate in a meta-regression

Planned meta-regression

- As a manipulation check, we measured cue identifiability under the unique conditions of each micro-experiment
- We can now introduce cue detection performance as a covariate in a meta-regression
- The effect size function is flat in the subliminal range



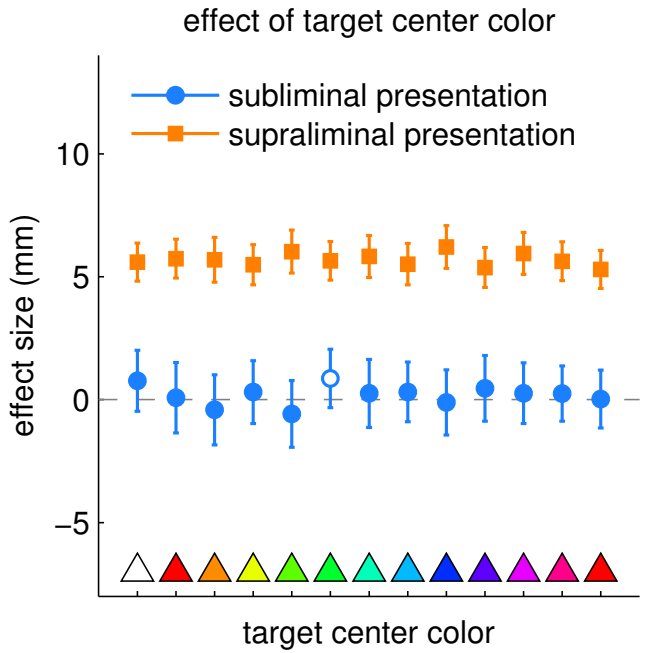
Effect sizes (Δmm) as a function of detection performance.

Planned meta-regression

– As robustness check, include target color as a covariate

Planned meta-regression

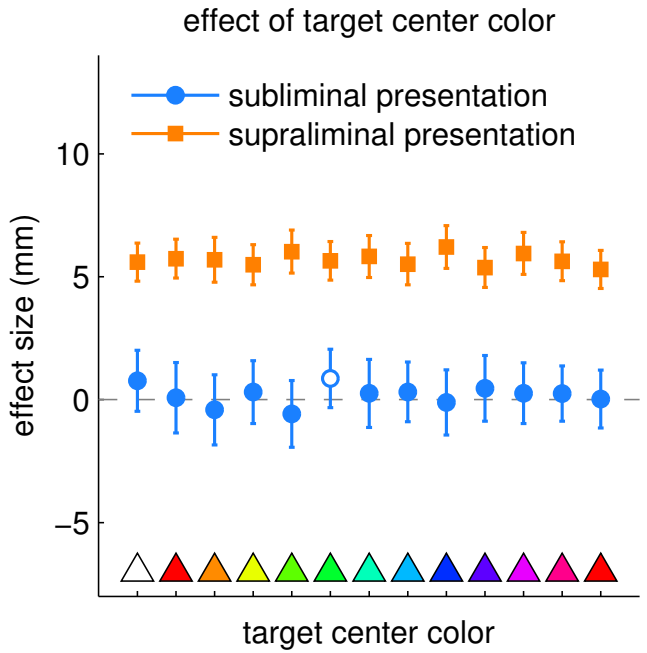
- As robustness check, include target color as a covariate
- Flat function indicates robustness to target color



Effect sizes (Δmm) as a function of target color, split by detection performance.

Planned meta-regression

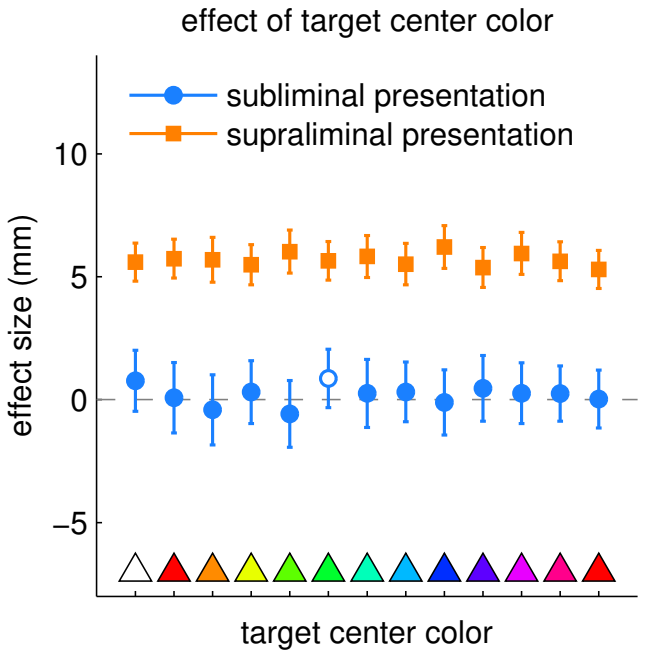
- As robustness check, include target color as a covariate
- Flat function indicates robustness to target color
- Conditional on supraliminality, the subliminal effect seems robustly present



Effect sizes (Δmm) as a function of target color, split by detection performance.

Planned meta-regression

- As robustness check, include target color as a covariate
- Flat function indicates robustness to target color
- Conditional on supraliminality, the subliminal effect seems robustly present
- Conditional on subliminality, the subliminal effect seems robustly absent



Effect sizes (Δmm) as a function of target color, split by detection performance.

Metastudies

This metastudy (Baribault et al., 2018). . .

Metastudies

This metastudy (Baribault et al., 2018)...

- permits statements at a higher level of abstraction: the existence of an effect in a well-defined and formalized universe of intended generalization (e.g., the varying design choices that other researchers might have made in this study)

Metastudies

This metastudy (Baribault et al., 2018)...

- permits statements at a higher level of abstraction: the existence of an effect in a well-defined and formalized universe of intended generalization (e.g., the varying design choices that other researchers might have made in this study)
- allows for defensive design: a design strategy that targets replicability, reproducibility, robustness, and generalizability

Metastudies

This metastudy (Baribault et al., 2018)...

- permits statements at a higher level of abstraction: the existence of an effect in a well-defined and formalized universe of intended generalization (e.g., the varying design choices that other researchers might have made in this study)
- allows for defensive design: a design strategy that targets replicability, reproducibility, robustness, and generalizability
- turns out to be feasible and relatively cheap

Metastudies

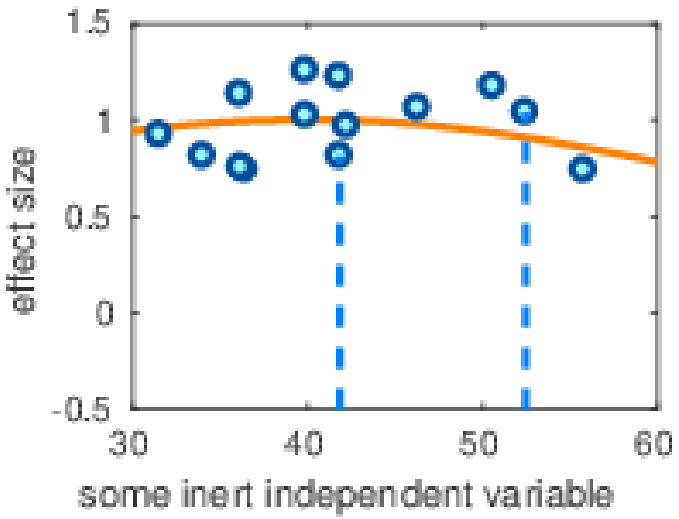
This metastudy (Baribault et al., 2018)...

- permits statements at a higher level of abstraction: the existence of an effect in a well-defined and formalized universe of intended generalization (e.g., the varying design choices that other researchers might have made in this study)
- allows for defensive design: a design strategy that targets replicability, reproducibility, robustness, and generalizability
- turns out to be feasible and relatively cheap
- doesn't require very complicated statistical methods (yet)

Discussion

Discussion

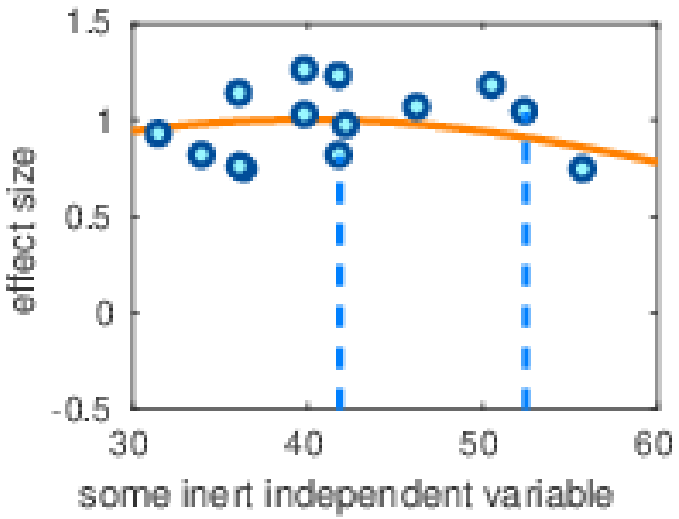
– I think most real effects are probably robust



What I think real effects behave like.

Discussion

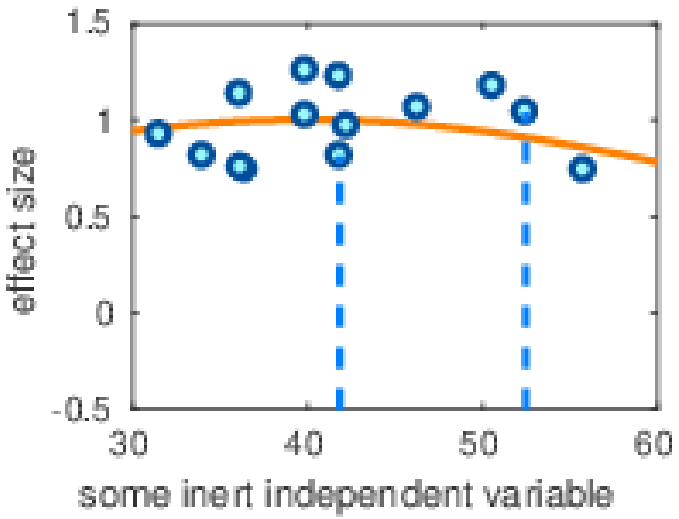
- I think most real effects are probably robust
- The Many Labs projects (e.g., Klein et al., 2014) lend little support to the idea that effects routinely wink in and out of detectability



What I think real effects behave like.

Discussion

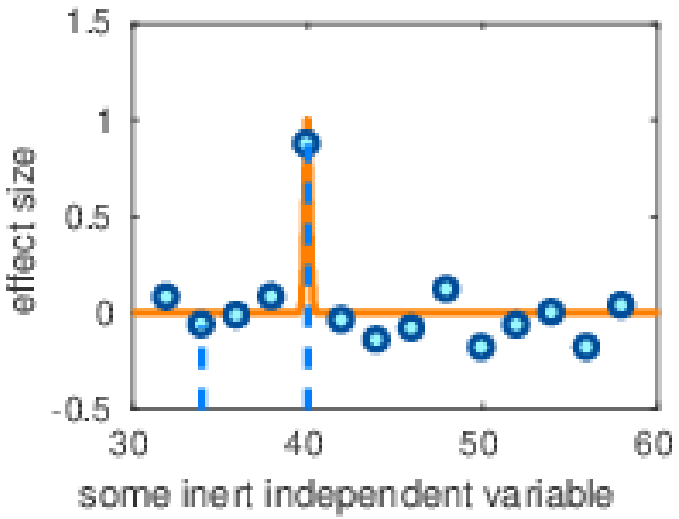
- I think most real effects are probably robust
- The Many Labs projects (e.g., Klein et al., 2014) lend little support to the idea that effects routinely wink in and out of detectability
- Only robust claims have predictive power, which presumably aids practical applicability



What I think real effects behave like.

Discussion

– I suspect many social scientists believe effects are often fickle

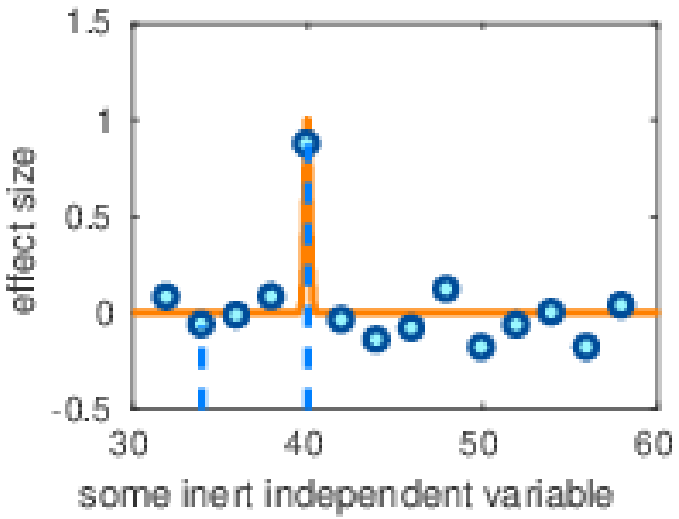


What I think many social scientists believe.

Discussion

– I suspect many social scientists believe effects are often fickle

  . . . making it important—and a bit of an art—to conduct very finely tuned experiments

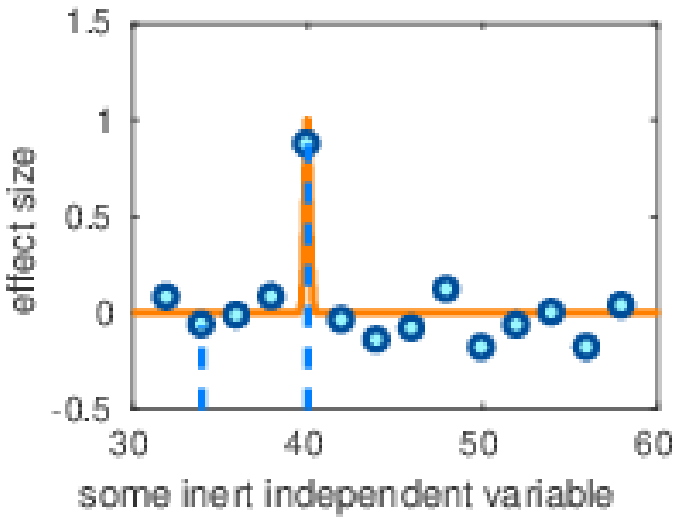


What I think many social scientists believe.

Discussion

- I suspect many social scientists believe effects are often fickle

  . . . making it important—and a bit of an art—to conduct very finely tuned experiments
- Valid or not, this mindset invites irreproducibility

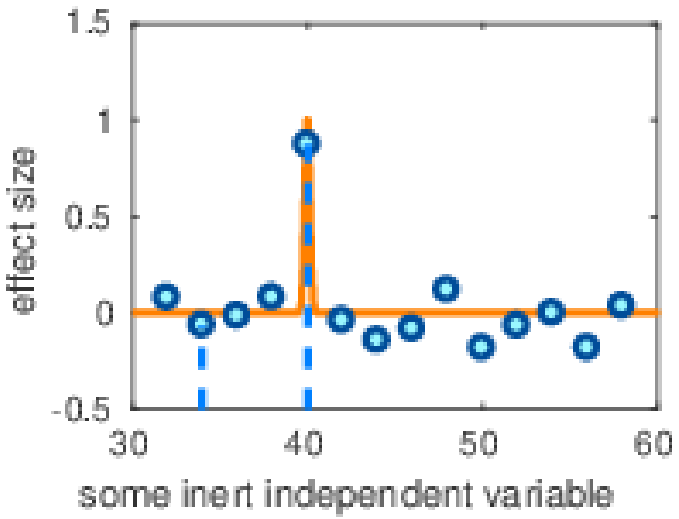


What I think many social scientists believe.

Discussion

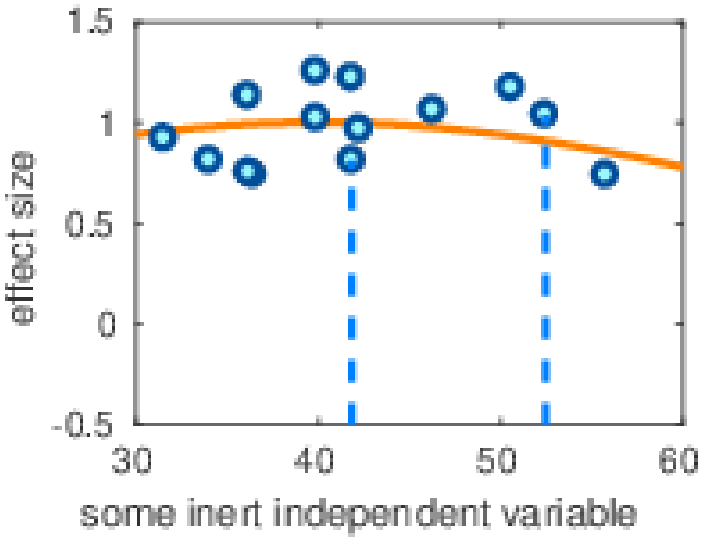
- I suspect many social scientists believe effects are often fickle

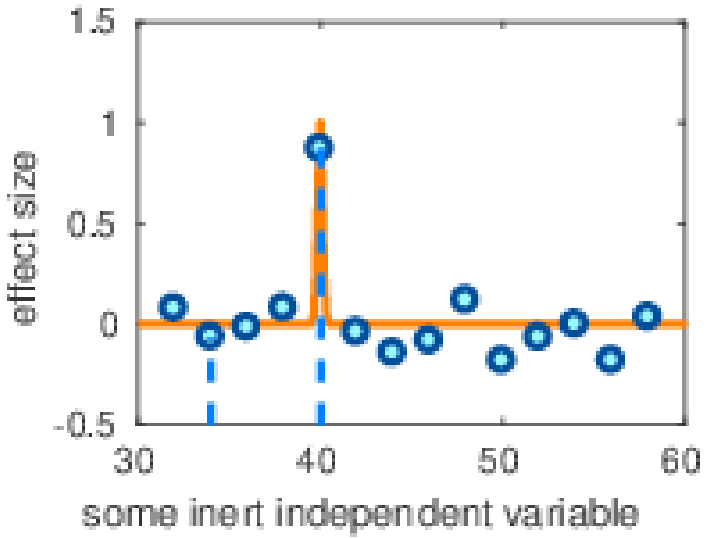
  . . . making it important—and a bit of an art—to conduct very finely tuned experiments
- Valid or not, this mindset invites irreproducibility
- Fickle effects are systematically hard to distinguish from flukes



What I think many social scientists believe.

Discussion





These two conceptions of real effects probably call for different kinds of empirical programs – strategies with different balances of exploration vs. exploitation.

Discussion

More generally I think there exist methods that are optimized for sensitivity, robustness, and generalizability, that could be used more in psychology and social sciences.

Discussion

More generally I think there exist methods that are optimized for sensitivity, robustness, and generalizability, that could be used more in psychology and social sciences.

Techniques like adding variability to designs to improve robustness and generalizability come directly out of the statistical / psychometrical literature on random effects.

Discussion

More generally I think there exist methods that are optimized for sensitivity, robustness, and generalizability, that could be used more in psychology and social sciences.

Techniques like adding variability to designs to improve robustness and generalizability come directly out of the statistical / psychometrical literature on random effects.

These techniques are not difficult to implement.

Discussion

More generally I think there exist methods that are optimized for sensitivity, robustness, and generalizability, that could be used more in psychology and social sciences.

Techniques like adding variability to designs to improve robustness and generalizability come directly out of the statistical / psychometrical literature on random effects.

These techniques are not difficult to implement.

$\sim fin \sim$

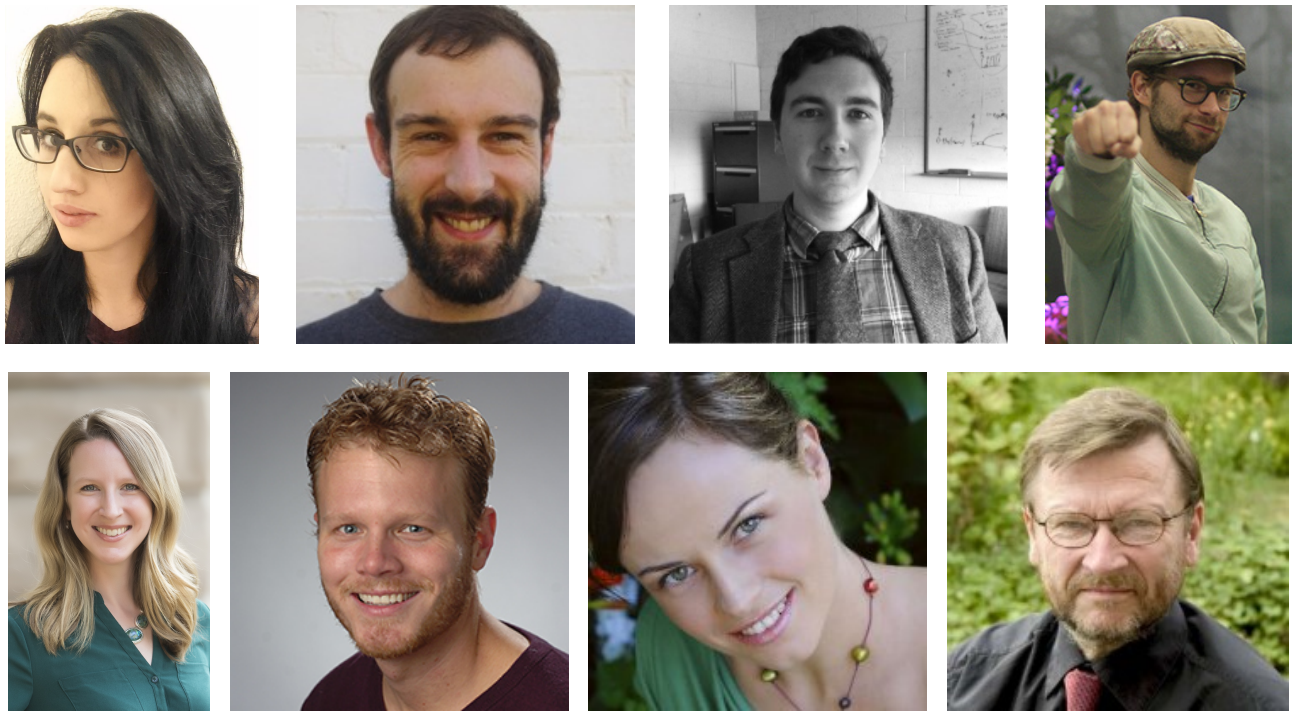
Some (current and recent) benefactors

UCIrvine
UNIVERSITY OF CALIFORNIA, IRVINE

JOHN TEMPLETON
FOUNDATION

A veritable mosaic of most excellent people

This is joint work with Beth Baribault, Chris Donkin, Daniel Little, Don van Ravenzwaaij, Jennifer Trueblood, Corey White, Zita Oravecz, and Paul de Boeck.



References i

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 453–455.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., et al. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 201708285.

Boyle, R. I. (1772). *The works of the Honourable Robert Boyle, In six volumes, To which is prefixed the life of the author* (Vol. 2). J. and F. Rivington.

Cicerone, R. (2015). *Research reproducibility, replicability, reliability.* National Academy of Sciences (NAS) Annual Meeting.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.

References ii

Kenett, R. S., & Rubinstein, A. (2017). A generalization approach to reproducibility claims. (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035070)

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social psychology*, *45*(3), 142.

Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62.

Rosenblatt, M. (2016). An incentive-based approach for improving data reproducibility. *Science Translational Medicine*, *8*(336), 336ed5–336ed5.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.