# Mathematically Elegant Answers to Questions No One is Asking

**Uri Simonsohn**

esade

# The overarching concern motivating this talk

- **Reality check**
  - Stat folks: sorry, we have mere *supporting* roles
  - Our research has no intrinsic value
  - Extrinsic value: help researchers answer *their* questions
- **As JDMer I worry**
  - "Do we study things *we* find interesting, but aren't useful?"
- **As Methodologist I worry**
  - "Do we study things *we* find interesting, but aren't useful?"
  - But it's worse
    - Most MBA students can decide whether 'embodied cognition'* is silly
    - Most researchers can't decide whether 'Random Effects' are silly
- **It's on us to be more transparent about what a method actually does**
  - Stop taking the math literally
  - Start taking researchers seriously

# I think of it as a transparency issue

- Important that other methodologists can check our work
- Also important: researchers can evaluate if our work is useful
  - Need to transparently (non-technically) explain actual trade-offs
  - Not philosophical platitudes (likely to be misinterpreted)

# How do researchers study things?
# How they choose study designs?
(meta-analytical mean; random effects; Bayes factors)

| **Taking math literally** | **Taking researchers seriously** |
|---|---|
| Drawn at random | Carefully curated, actively non-random |
| From defined populations | From undefined/inexistent populations (generally) |
| With known distributions | No population → no distribution<br>If they exist, each researcher their own |
| Goal: estimate population mean effect | Goal: local test of *this* effect<br>Qualitative generalization based on thinking |

# Outline

**My Claim: Researchers don't want the answers provided by these tools**

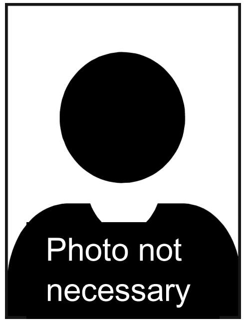
1. Mixed models (Platonic generalizability)
2. Meta-analysis (Overall means or subgroup means)
3. Bayes Factors (testing some average hypothesis)

Free Trial

# Non-Random Effects:

## Designing & Analyzing Experiments with Multiple Stimuli (in The Real World)


**Uri Simonsohn**
*ESADE, Barcelona*

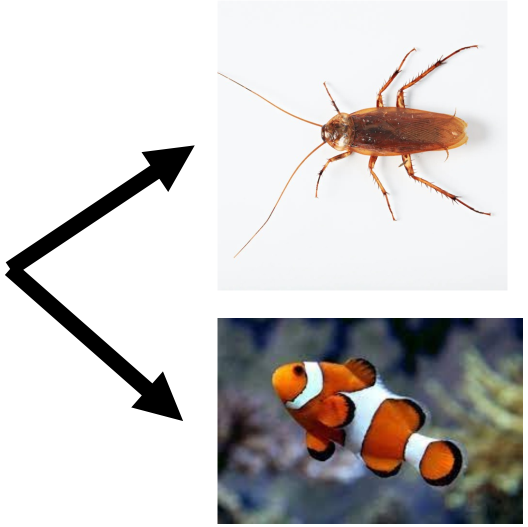



**Andres Montealegre**
Cornell (PhD Student)

**Ioannis Evangelidis**
*ESADE, Barcelona*

# Psychology's unique experimental challenge

- **Hard & applied sciences**
  - What's the impact of this vaccine? Randomize vaccine → Got Covid?
  - What's the impact of defaults? Randomize default → % organ donors?
- **Psychology**
  - What's the impact of disgust on moral judgments?



→

**Moral judgment**
Is this ok? (1-7)

# Psychology experiments produce mere correlations

(this seems simultaneously obvious and earth shattering)

- We randomly assign stimuli to participants
- We do not randomly assign attributes to stimuli
- *Stimuli* are confounded

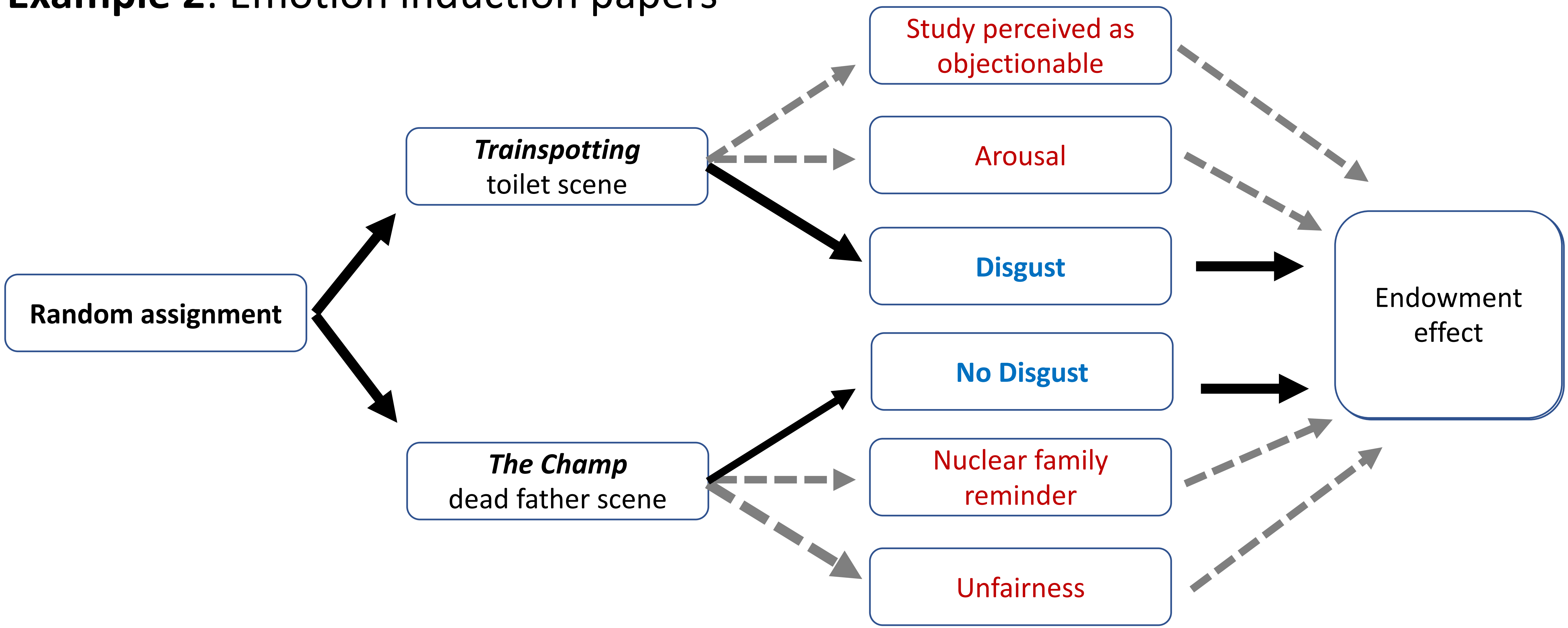
➔ psychology experiments are confounded

- **Example 1**: Rubenstein et al (1971)
  - Homophonic words: slower recognition
  - Participants randomly shown words, e.g. `Pray` & `Pest`
  - `Pray` NOT randomly assigned to have homophone
  - Reaction time to `Pray` vs `Pest` is confounded

# Psychology experiments produce mere correlations

(this seems simultaneously obvious and earth shattering)

**Example 2**: Emotion induction papers



# Mixed-model consensus

- **Concern is external validity**
  - Generalize beyond chosen stimuli

[Clark 1971] >2,900 citations
[Baayen, Davidson, & Bates, 2008] >8,400 citations
[Barr, Levy, Scheepers, & Tily, 2013] >8,100 citations
[Judd, Westfall, & Kenny, 2012] >1,100 citations

- **Recommendations:**
  - Many stimuli
  - Use mixed models

- **Says nothing on :**
  - How to select stimuli (beyond, choose many, at random)
  - How to learn from stimuli variation

**Skipping:**

Our paper proposes "Match-and-Mix 1.0"

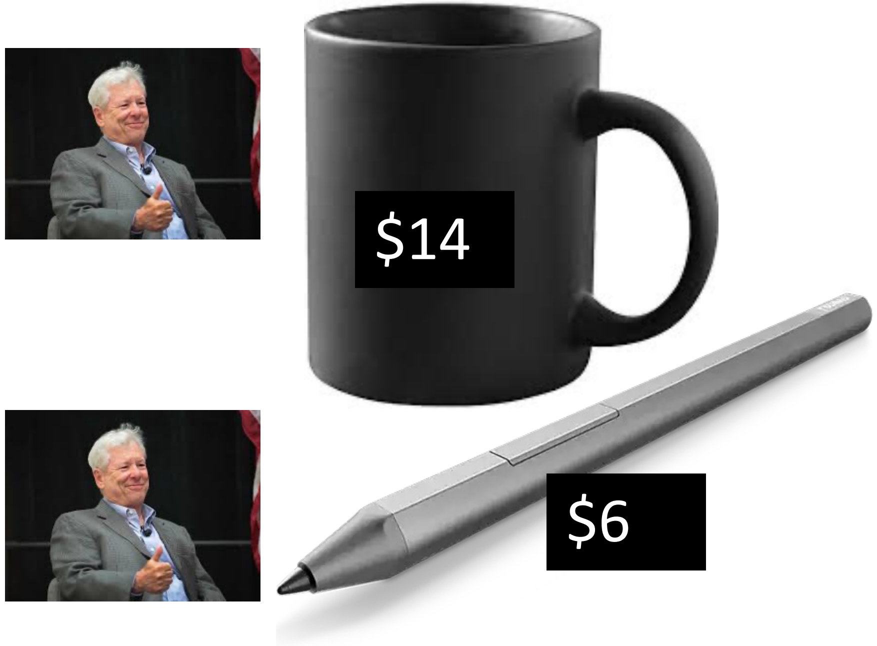
6 steps to choosing (a few) stimuli

**For this talk:**

Let's focus on the statistical analysis of multi-stimuli experiments

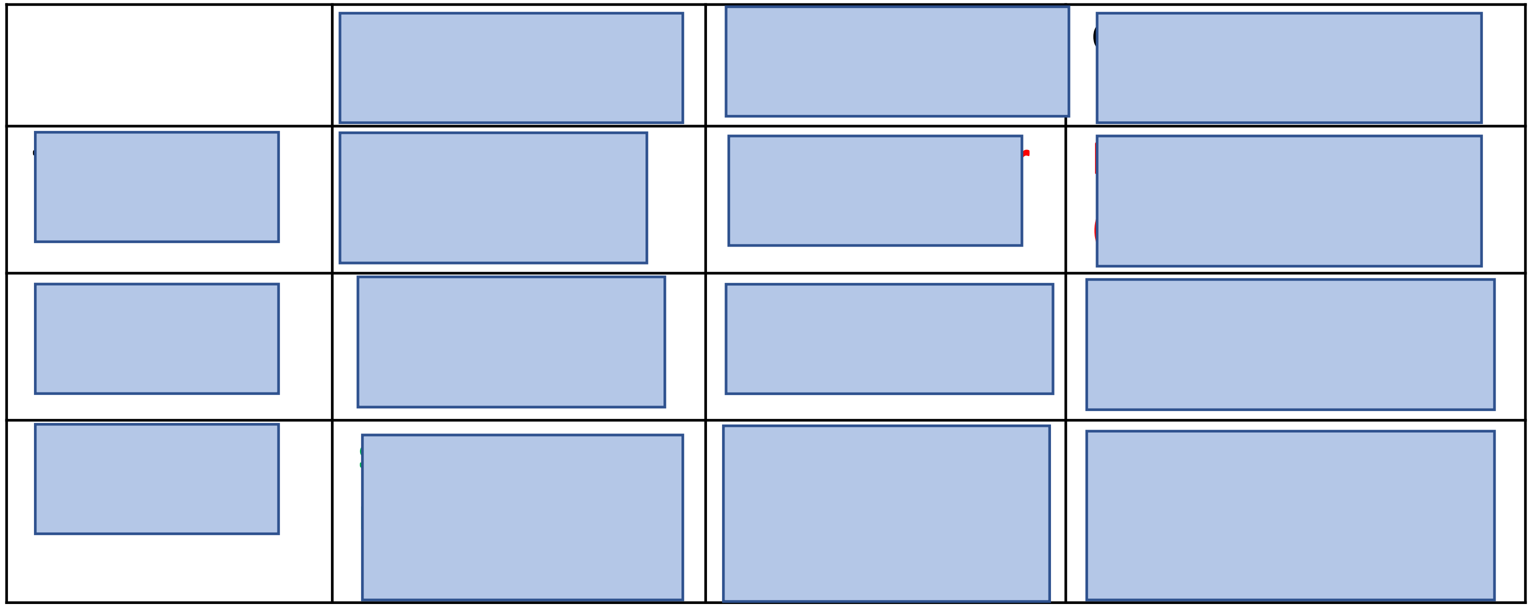
# Analyzing Studies with Many Stimuli

*Example: Endowment-effect*


CHALLENGES

TOOLS



# Platonic Generalizability

1. **Assume a population of all possible stimuli exist**
   - All goods that exist
   - All goods one could imagine

   → Now average them
     - People. 50:50 Women:Men
     - Endowment effect:
       - x% Mugs
       - y% Obama dinners
       - z% refurbished iPhone 11
   - "The" effect we estimate: weighed mean
2. **Assume stimuli were chosen at random from it**
3. **Assume researcher wants to generalize / estimate (1)**

(1) exists in theory only → We call it platonic generalizability.

If it were free to get platonic generalizability we ***may*** buy it.

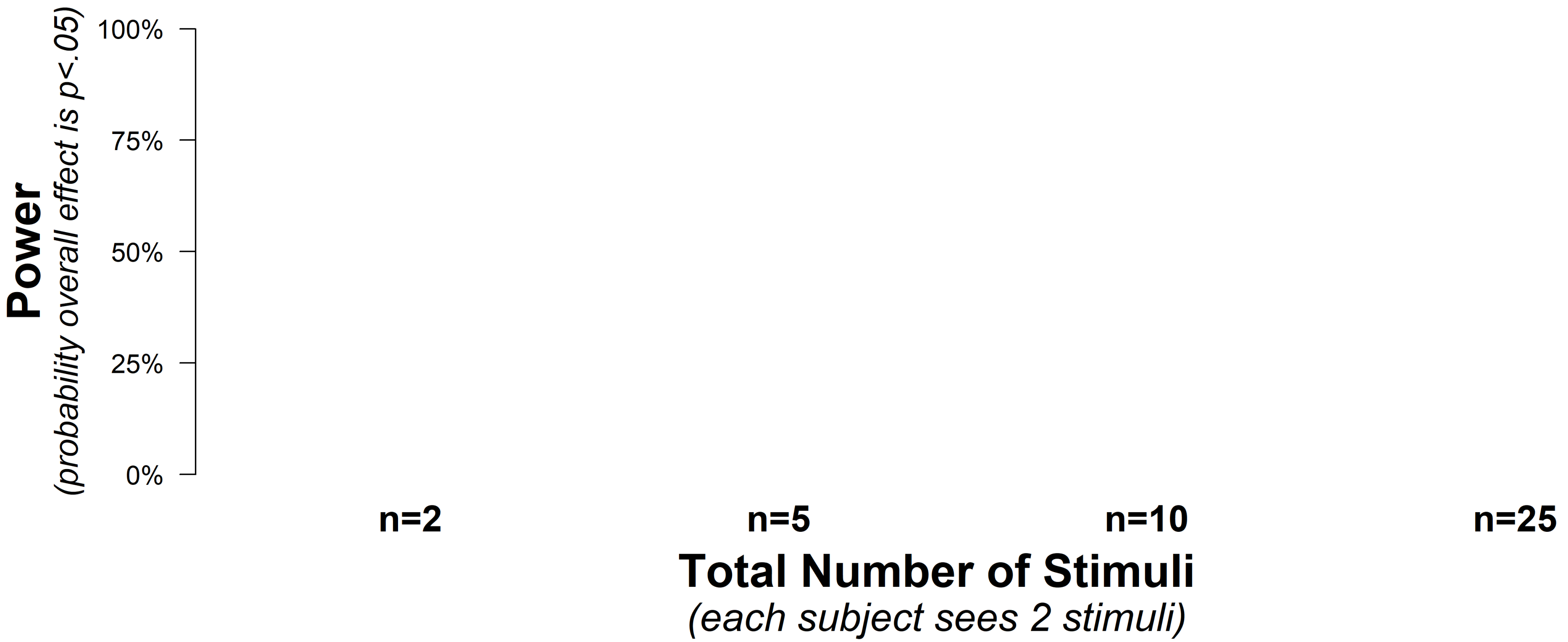
But it is very expensive.

**Next**. Simulations for statistical power

1) Participants see *n* out of *n* stimuli
2) Participants see 2 out of *n* stimuli

# Case 1. Subjects see *n* out of *n*

# Case 2. Subjects see 2 out of *n*




**Takeaway:**
**Controlling for stimuli increases power when k of n**

**Mixed model still expensive**

# Mixed model advocates know about power

- **But they don't care**
  - They worry t-tests have too many false-positives
- **We sure care about false-positives**
  - But not about *those*
  - We think they are *true*-positives.
- This can get philosophical…

…let's make it super concrete.

General Article

**False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant**

Psychological Science
22(11) 1359–1366
© The Author(s) 2011
Reprints and permission: sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

**Next.** Let's contrast those two perspectives in a figure.

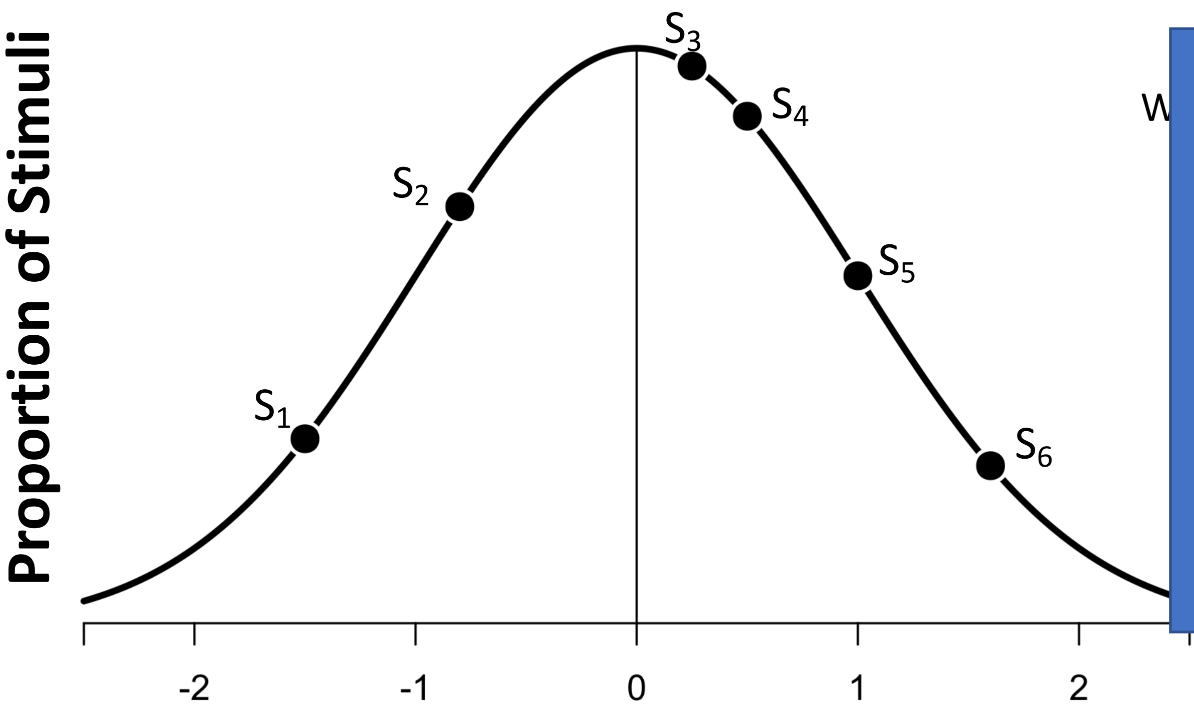
**Again, "the" population mean does not exist:**

Men:Women 50:50

Nail clipping men vs Trainspotting videos x%:y%?

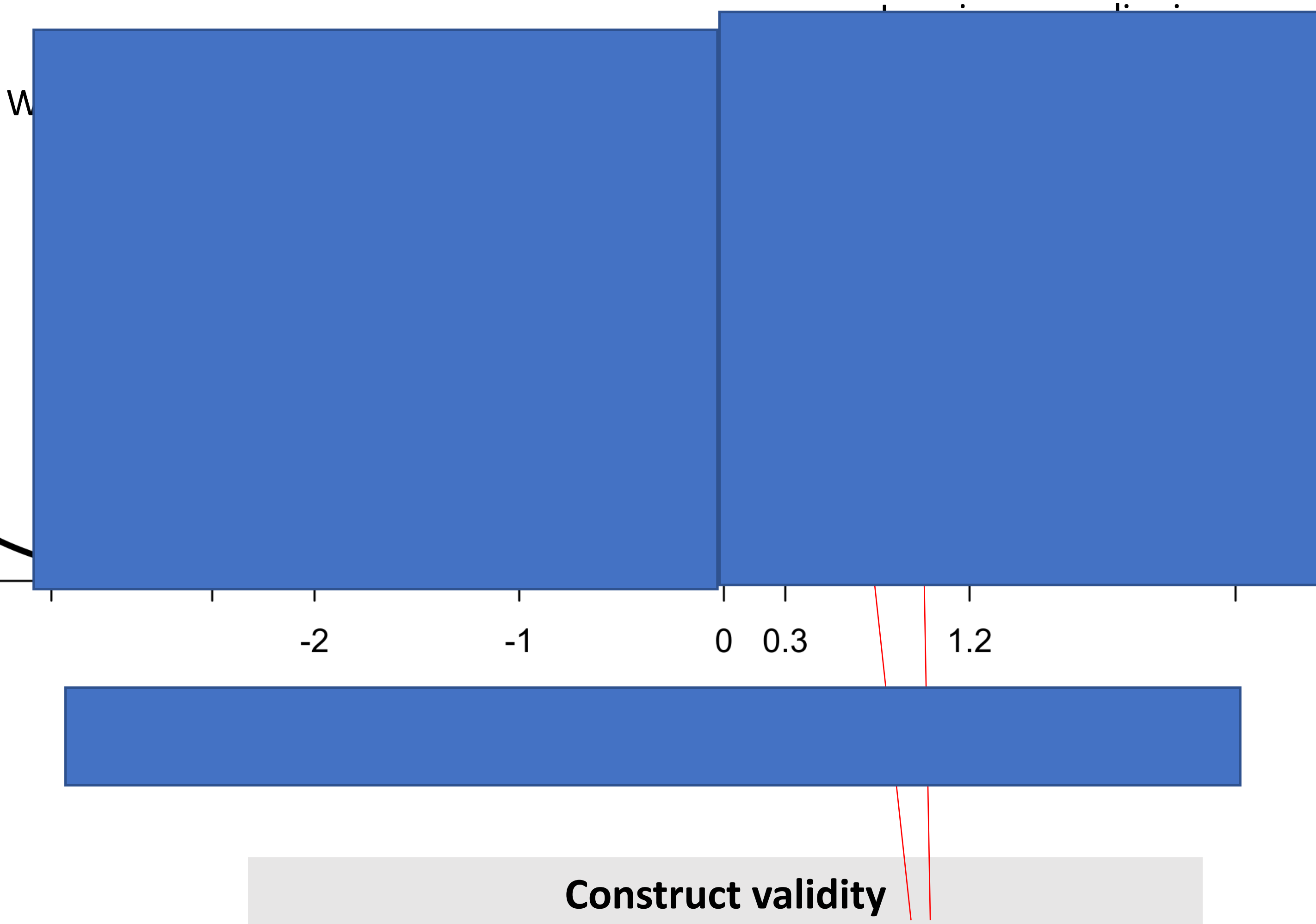
$S_4$ and $S_5$ are **true**-positive effects

## A. Textbook distribution of effect size




**Effect of stimulus on dependent *variable***

## B. Slightly more concrete & realistic




**Platonic generalizability**:
Mean of all possible stimuli is 0?

**Construct validity**
Do we *generally* get the effect when expected?

- Our interest as researchers should guide the tools we use

- Not vice versa

- We thus propose a tool to assess if you *generally* get an effect when you expect it.

**‘Stimuli Plots’**

# Stimuli Plots

- Compute effect for each matched-pair of stimuli in control condition
- Assess if effect is obtained in general
- Assess if variation identifies
  - Possible confounds
  - Interesting moderators
  - Ideas for the next study

**Next**: stimuli plots for three published papers

# Paper 1. Kupfer et al (2020)

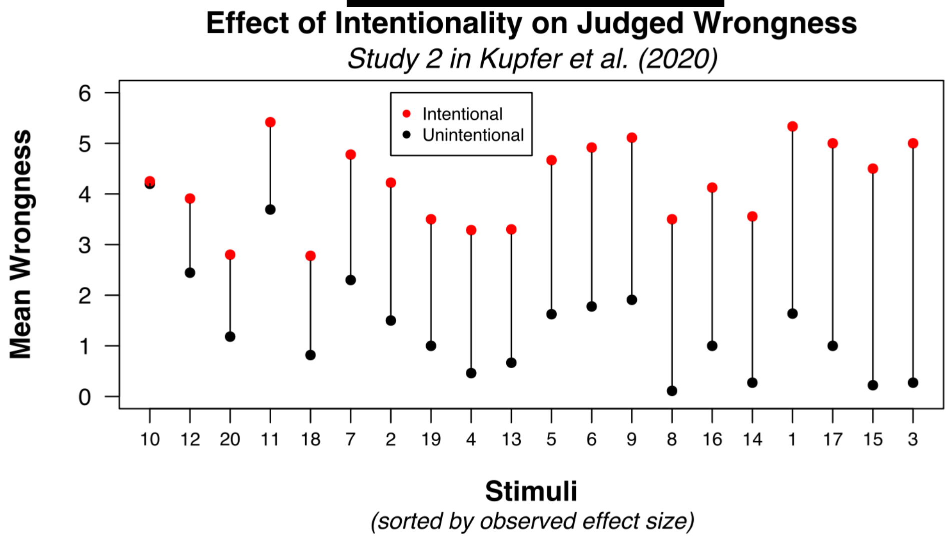
Registered Report Stage 2: Full Article

## Reexamining the role of intent in moral judgements of purity violations☆

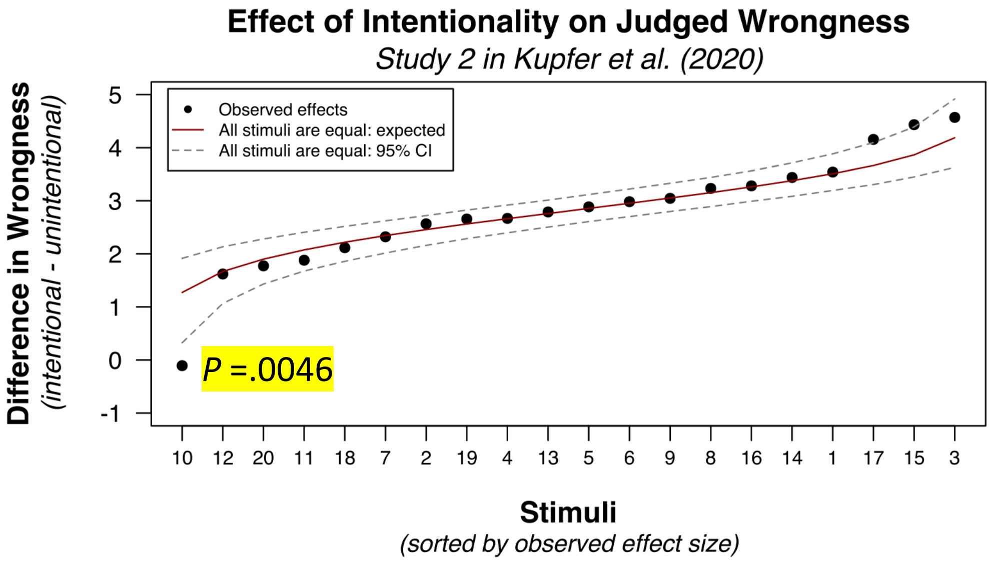
Tom R. Kupfer[a,*], Yoel Inbar[b], Joshua M. Tybur[a]

[a] Vrije Universiteit Amsterdam, the Netherlands
[b] University of Toronto, Canada

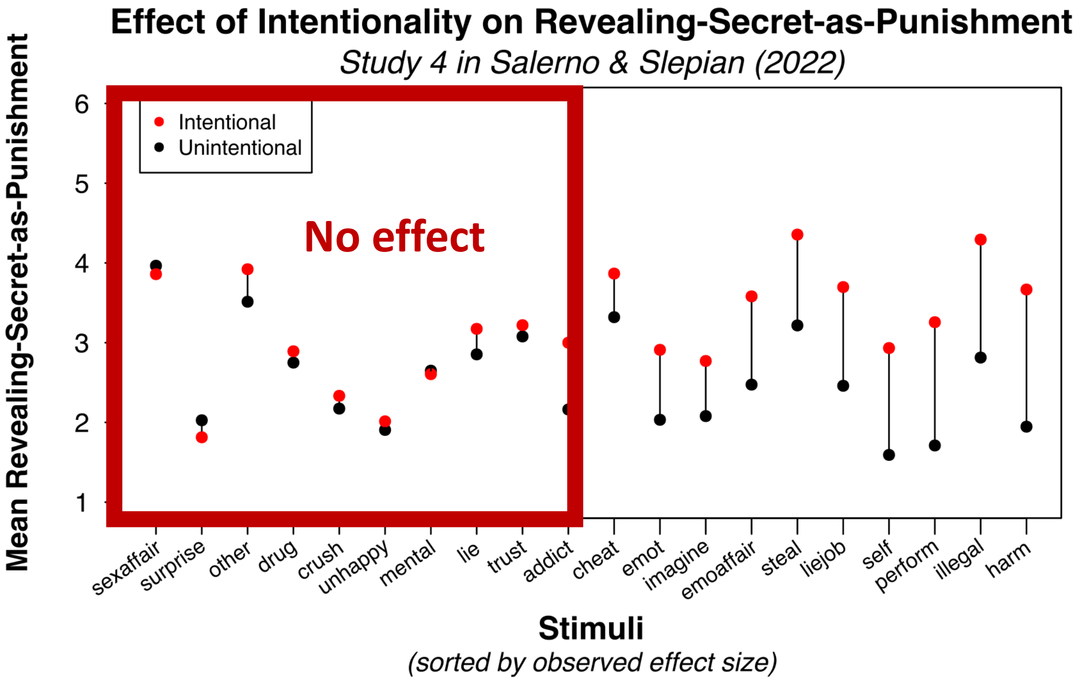
**Means by Stimuli**

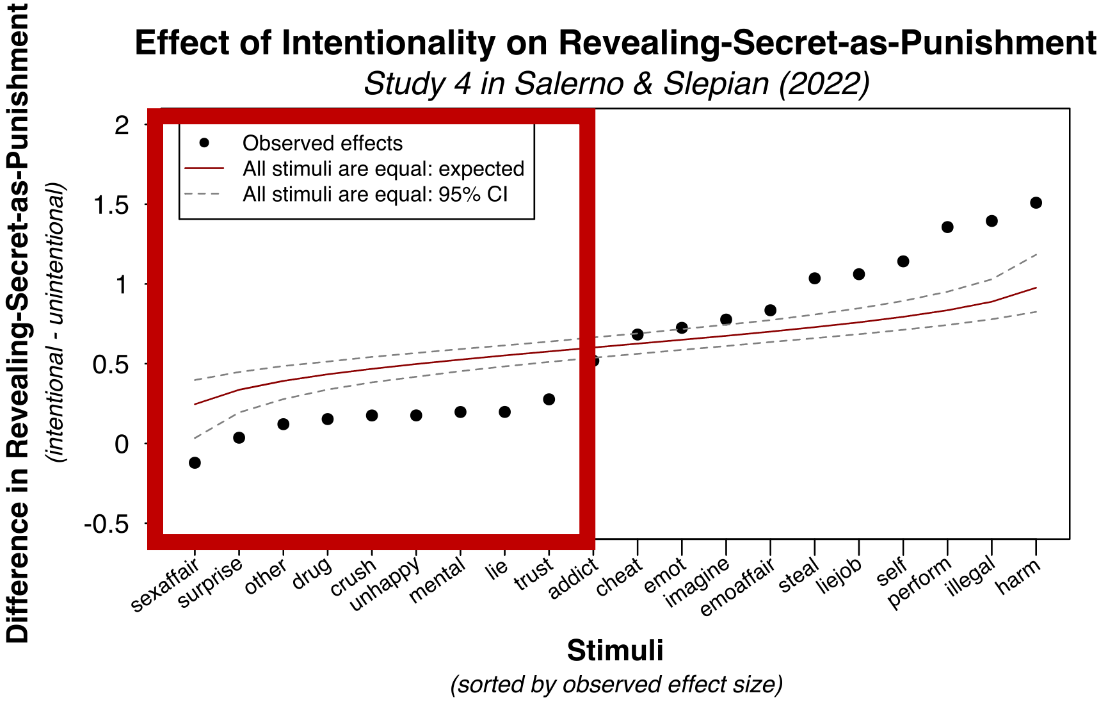



***Effects*** **by Stimuli**



# Paper 2. Salerno & Slepian (2022)

# Paper 3. Rottman & Young (2019)

## Two points from our perspective:

*Research Article*

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

1) Stimuli not matched purity/harm

Violations manipulated by magnitude

A person throws a [small/large] rock at a farm animal. A person eats a [small/large] amount of flesh from a dead person.

**Spe…**
**Distinguishes Impurity From Harm**

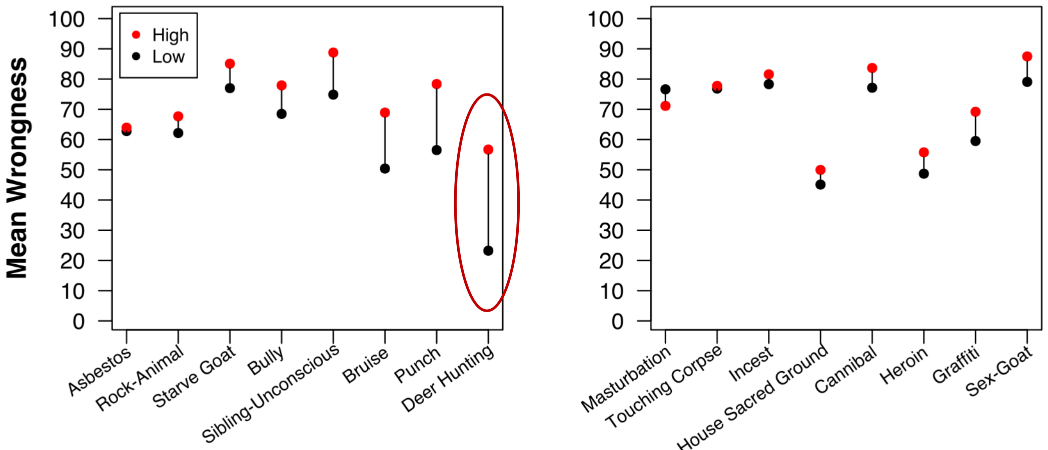
DOI: 10.1177/0956797619855382
www.psychologicalscience.org/PS
SAGE

2) Within harm: deer-hunting is an outlier

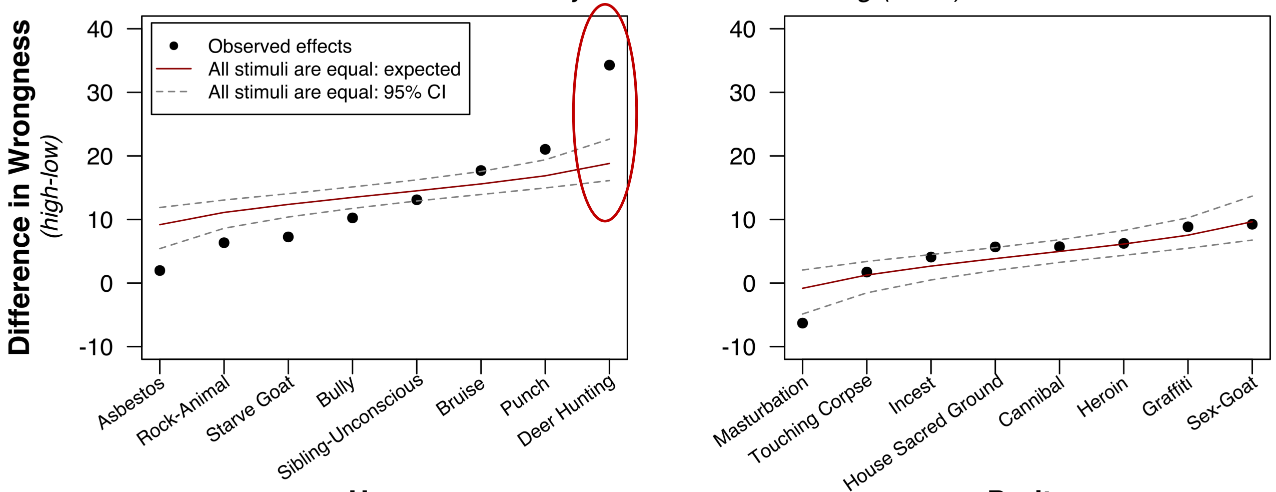
**Joshua Rottman**[1] **and Liane Young**[2]
[1]Departm…

A person kills [two/50] deer while hunting.

**Effect of Domain and Dosage on Wrongness**
*Study 1 in Rottman & Young (2019)*

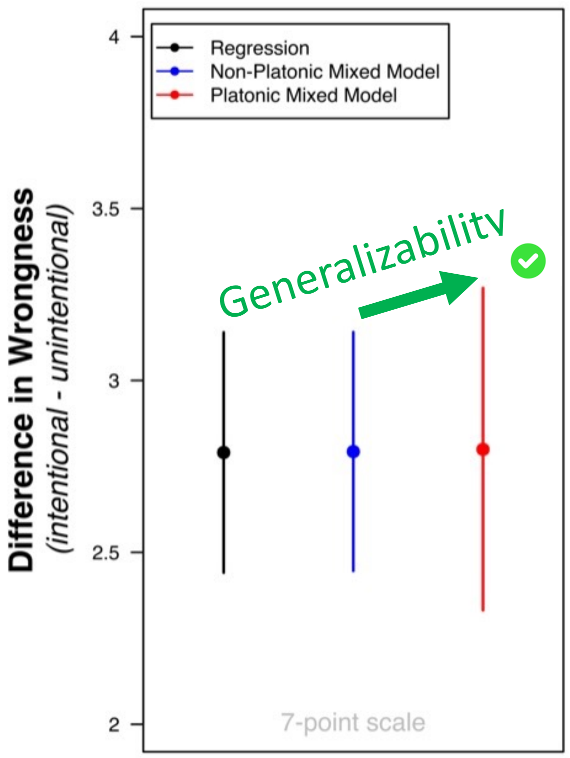



**Effect of Domain and Dosage on Wrongness**
*Study 1 in Rottman & Young (2019)*



- Contrast information provided by t-test & stimuli-level data
- With mixed-model results

## Confidence Interval for Mixed Model vs. Regression Using Published Papers




**Effect of Intentionality on Judged Wrongness**
Kupfer, Inbar, & Tybur (2020, JESP)

# Outline

**My Claim: Researchers don't want the answers provided by these tools**

1. Mixed models (Platonic generalizability)
2. Meta-analysis (Overall means or subgroup means)
3. Bayes Factors (average hypothesis)

# Also makes sense if taking math literally

1. Population of effects exists
2. Researchers sample at random
3. Estimand: overall mean

Thinking about evidence, and vice versa

nature reviews psychology

Explore content | About the journal | Publish with us | Subscribe

nature > nature reviews psychology > comment > article

Comment | Published: 02 September 2022

# Above averaging in literature reviews

Uri Simonsohn, Joseph Simmons & Leif D. Nelson

*Nature Reviews Psychology* **1**, 551–552 (2022) | Cite this article

## Meaningless Means Series

A Colada series, of indefinite length, on the meaninglessness of meta-analytical means

| | |
|---|---|
| Introduction | Colada [104] |
| #1 – The Average Effect of Nudging is d=.46 | Colada [105] |
| #2 – The average Effect of Nudging, by Academics, is 8.7% | Colada [106] |

# Why meaningless?

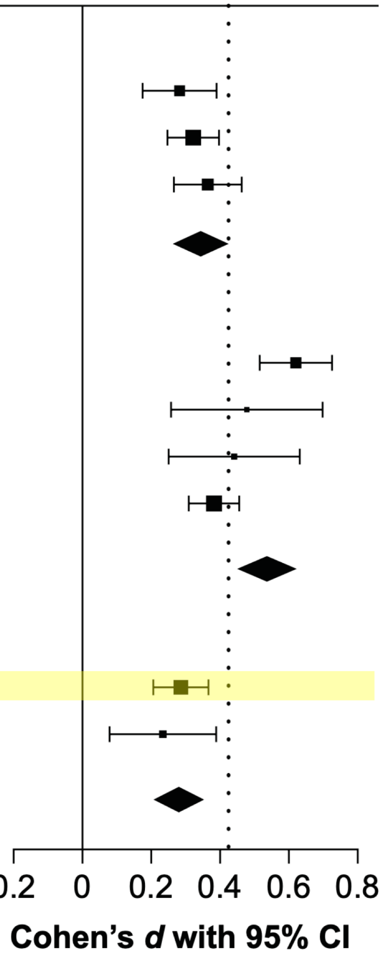
1) **No quality control (skip here)**
2) **Combining incommensurate results**

# Example #1 of Incommensurate Findings

## PNAS Nudge Meta-analysis

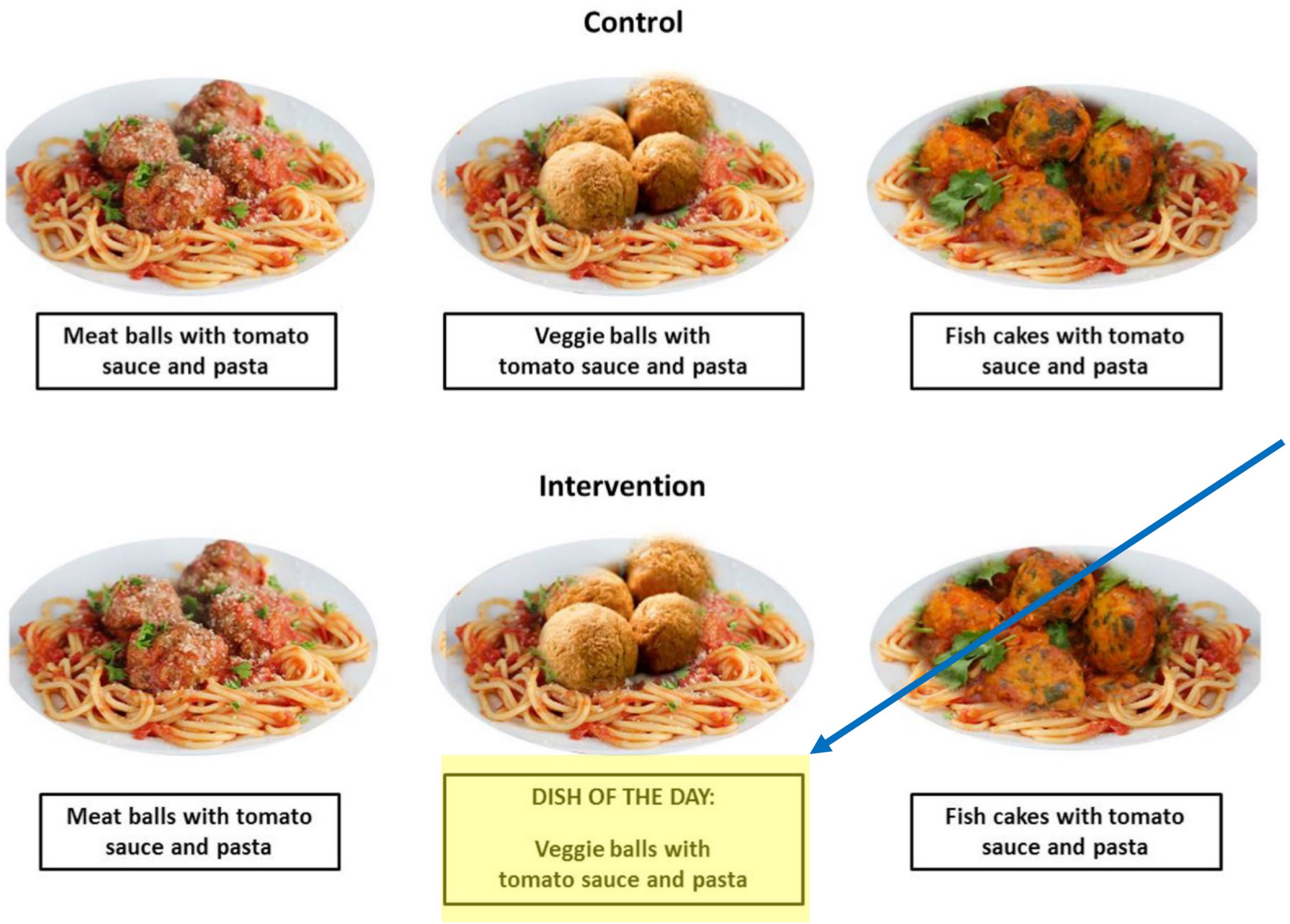
http://datacolada.org/105

| Intervention | *d* | 95% CI |
|---|---|---|
| **Decision information** | | |
| Translation[a] | 0.28 | [0.17, 0.39] |
| Visibility[b] | 0.32 | [0.25, 0.40] |
| Social reference[c] | 0.36 | [0.27, 0.46] |
| Average effect for category[g] | 0.34 | [0.27, 0.42] |
| **Decision structure** | | |
| Default[a,b,c,d,e,f] | 0.62 | [0.52, 0.73] |
| Effort | 0.48 | [0.26, 0.70] |
| Composition | 0.44 | [0.25, 0.63] |
| Consequence[d] | 0.38 | [0.31, 0.46] |
| Average effect for category[g,h] | 0.54 | [0.46, 0.62] |
| **Decision assistance** | | |
| Reminder[e] | 0.29 | [0.21, 0.37] |
| Commitment[f] | 0.23 | [0.08, 0.39] |
| Average effect for category[h] | 0.28 | [0.21, 0.35] |



# Estimate #1

Effect Size
d = - .12




The key manipulation was whether the veggie balls were labeled as the ”Dish of the Day”

**Fig. 1** How the dishes were presented in Control and Intervention groups

# Estimate #2

d= 1.18

# meta-analysis

- Our estimate of 'the' effect of reminders":

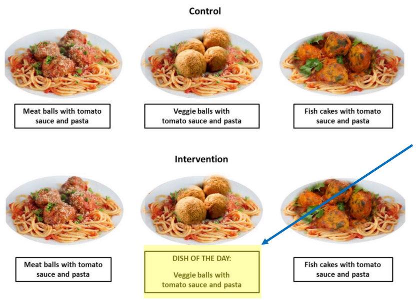
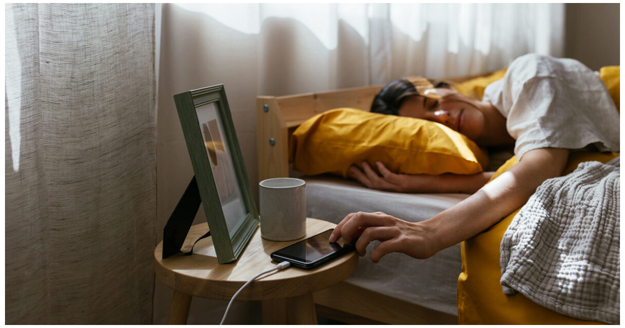



Fig. 1 How the dishes were presented in Control and Intervention groups

The key manipulation was whether the veggie balls were labeled as the "Dish of the Day"

-.12

+



+1.18

$$\frac{-.12 + 1.18}{2} = d=.53$$

Yeah. That's what we wanted to know

# Example #2 of Incommensurate Findings

## *Econometrica* Nudge Meta-analysis

[http://datacolada.org/106](http://datacolada.org/106)

D. Hummel and A. Maedche — Journal of Behavioral and Experimental Economics 80 (2019) 47–58

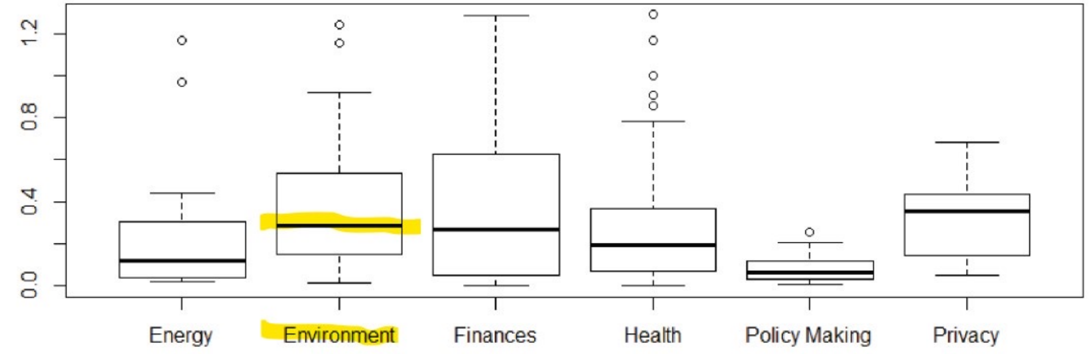



Fig. 4. Boxplot of relative effect sizes per context.

+ 51%


+ 7%


+ 4%

**The average environment nudge: ~21%**

**That average only makes sense if we take the math literally.**

- There is no population of effects
  (What % of nudges involve website defaults vs researchers stopping by?)
- Researchers do not run studies at random
- Readers do not want to know the average effect

# Outline

**My Claim: Researchers don't want the answers provided by these tools**

1. Mixed models (Platonic generalizability)
2. Meta-analysis (Overall means or subgroup means)
3. Bayes Factors (the average hypothesis)

# Data Colada [78]

## Drop That Bayes: A Colada Series on Bayes Factors

This series attempts to explain in simple terms what Bayes factors do, assume, mean and require people to be OK with if they want to use them. I (Uri) do not believe that many social scientists would embrace Bayes factors, if they understood them, and this is my attempt to convey that message.

**Post 1. DataColada[78a] – Milton and Minimum Wage**
The first post uses an example, where Milton predicts an effect between 1% and 10%, and upon seeing 1%, the Bayes factor deems this effect, which was predicted by Milton, as contradicting Milton's prediction. The example is used to convey the intuition of how Bayes factors assess if data are consistent with a theory, and contrasts it to how researchers do.

**Post 2. DataColada[78b] – Hyp-Chart: the missing link between *p*-values and Bayes factors**
This posts introduced Hyp-Chart, a plot that shows how consistent the data are with every possible hypothesis, compared to the null hypothesis. The Bayes factor is but a (bad) summary of Hyp-Chart.

**Post 3. DataColada[78c] – Looking at 10 papers in *Psych Science* that report Bayes factors**
In three Psych Science papers obtaining a non-significant effect ($p$>.05), the Bayes factor is shown to be non-diagnostic of whether the data do or do not support the null.

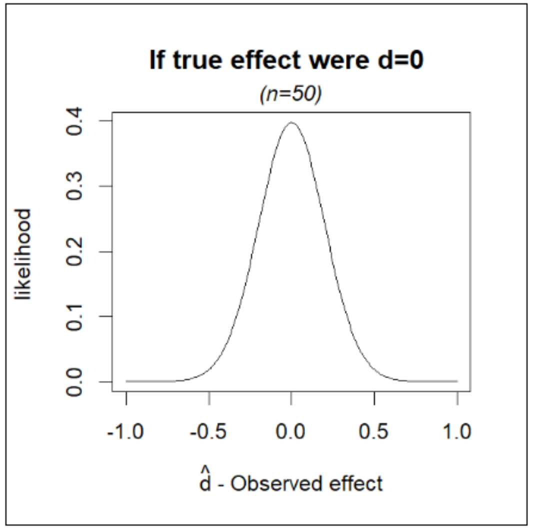
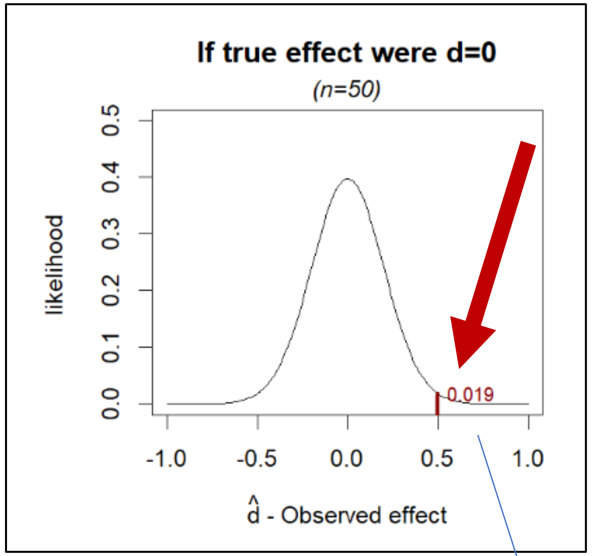
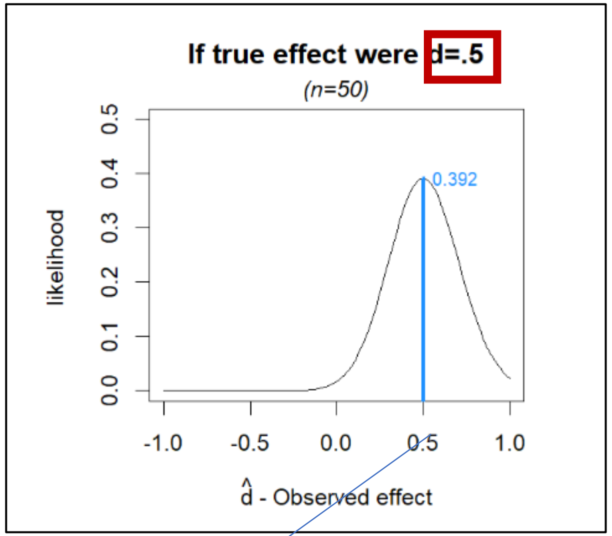



Fig 1. *If null is true, $d_{true}$=0, what's the likelihood of each estimate?*




**Likelihoods observing d=.5**

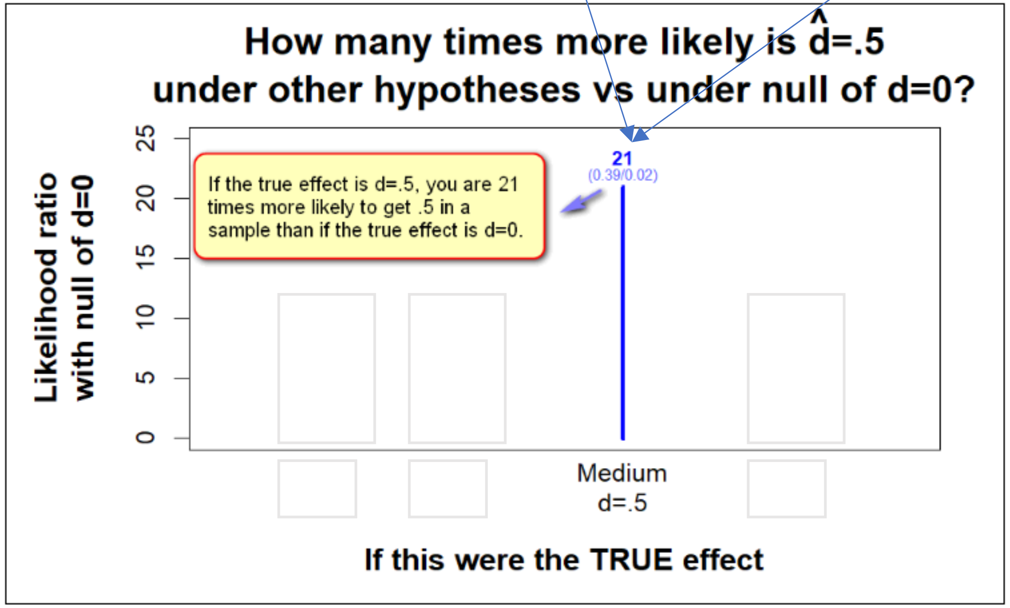
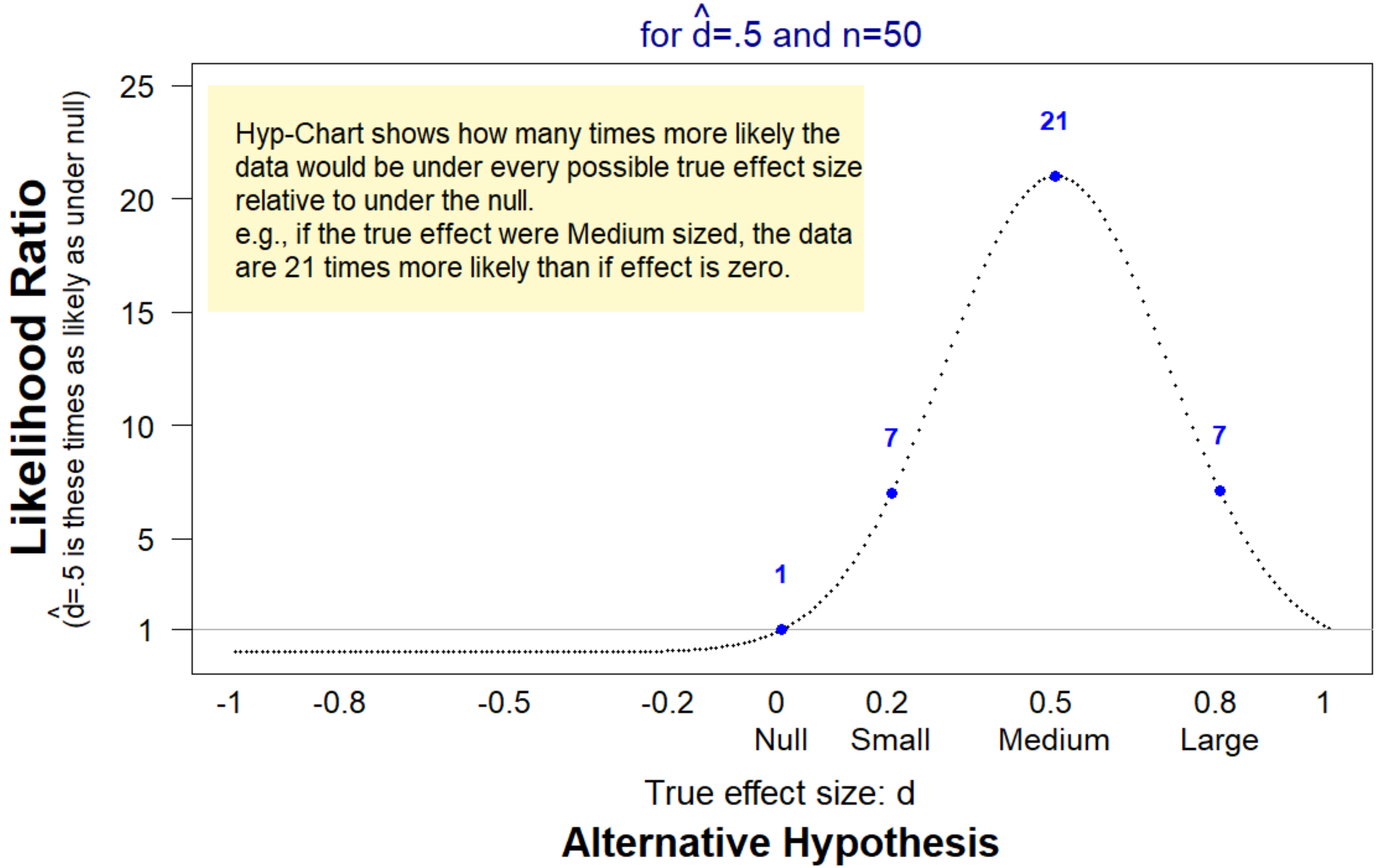



Fig 4. *How likely is $\hat{d}$ =.5 if $d_{true}$=0 vs if it is Small, Medium or Large?*

- **Let's do that for every possible hypothesis**
  - Not just t-shirt sizes

# Hyp-Chart




## Uri's claims

**1) Confidently.**
Many researchers would like this chart and would speak to their question.

**2) Semi confidently**
But probably be persuaded confidence intervals actually have the info they want

**3) Most confident**
***Nobody*** wants the average blue number

Especially not weighted by assumed N(0 , .71)

i.e. Bayes Factor

# Bayes Factors

- **Taking math literally**
  - Assume there is a population of effect size
  - Assume it is centered at 0 and symmetric
  - Assume researchers draw studies at random
  - Assume they wish to know if any particular study is:

  A) more consistent with that family of all possible effects (including 0)

  B) Null of d=0.

What researcher would read that and say *"that's exactly what I want"* ?

# Discussions

**Math literally**

- Ha ha, that's not "*evidence*"
- This or that paradox
- Don't you want to have a principled guide for inference?

**Researchers seriously**

- Does your research question involve an average of hypotheses with these particular weights?

# Shortcomings to my argument

- I am equating my take on researchers being taken seriously
- It is possible to make common-sense arguments against many ideas
- That's OK. We can have *those* arguments.
- The meta-point:
  - → we need methods arguments real researchers can play jury to