A critique of IRAP research

Ian Hussey, Jamie Cummins & Chad Drake

Data & code:

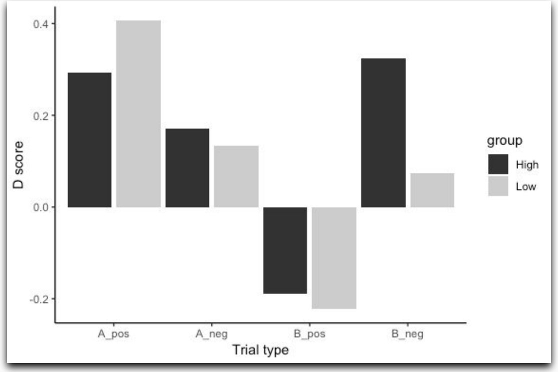
osf.io/ke7zx

3 most common ways to analyse IRAP data

D-IRAP scores differed significantly from zero, $t(20) = 3.85$, $p = .001$

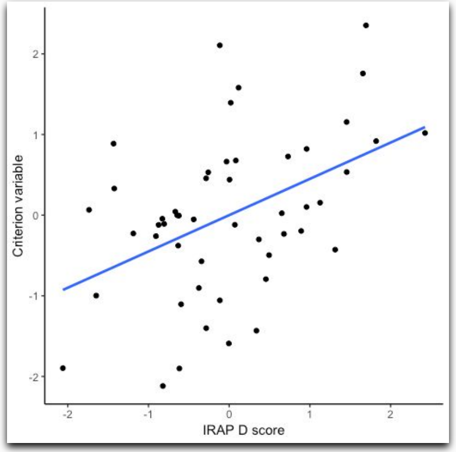
1

Presence of IRAP effects



2

Mean differences in IRAP effects



3

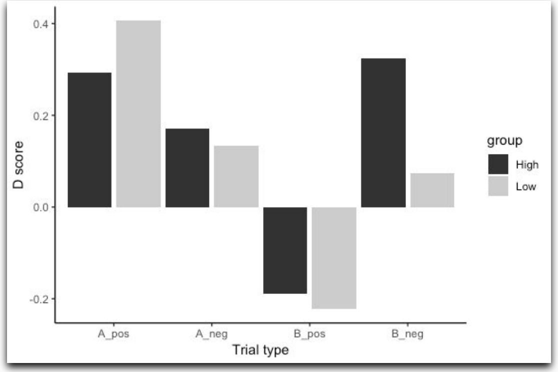
IRAP effects correlating with other variables

3 most common ways to analyse IRAP data

D-IRAP scores differed significantly from zero, $t(20) = 3.85$, $p = .001$

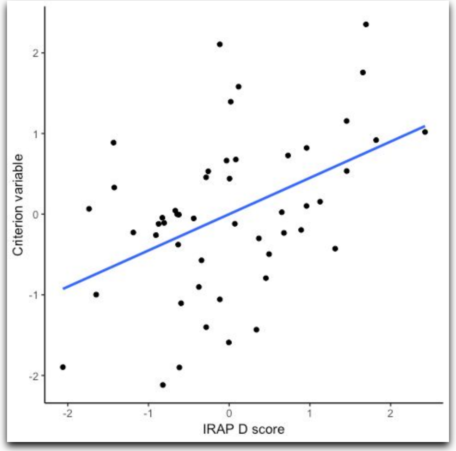
1

Presence of IRAP effects



2

Mean differences in IRAP effects



3

IRAP effects correlating with other variables

IRAP effects are misinterpreted

A large-scale analysis

Generic Pattern

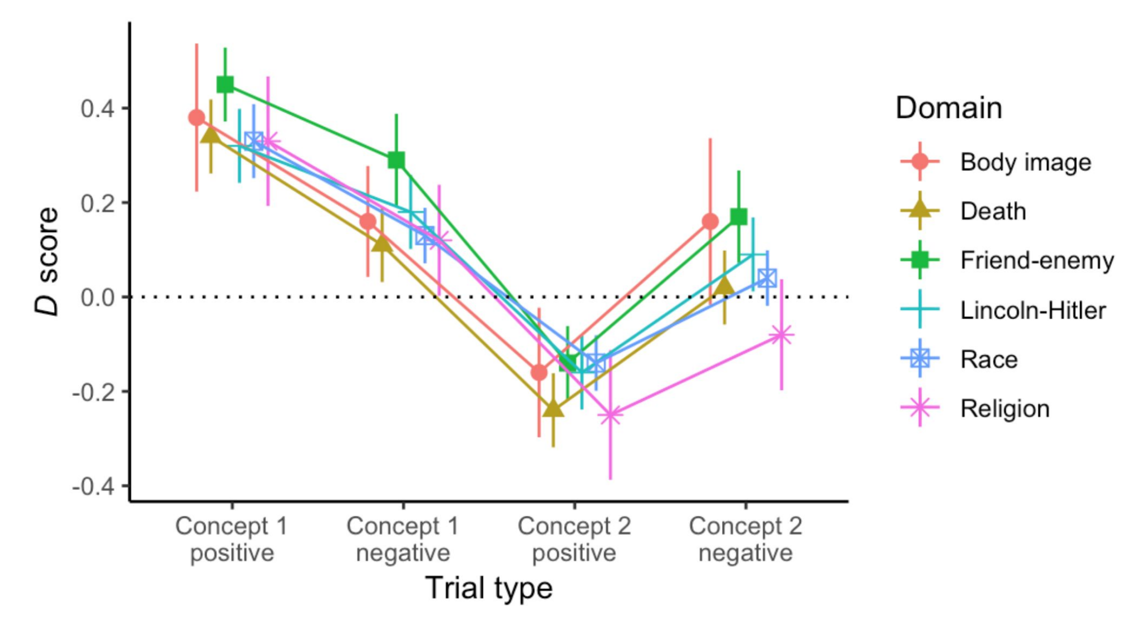
There is a Generic Pattern among IRAP effects

Described in different ways:

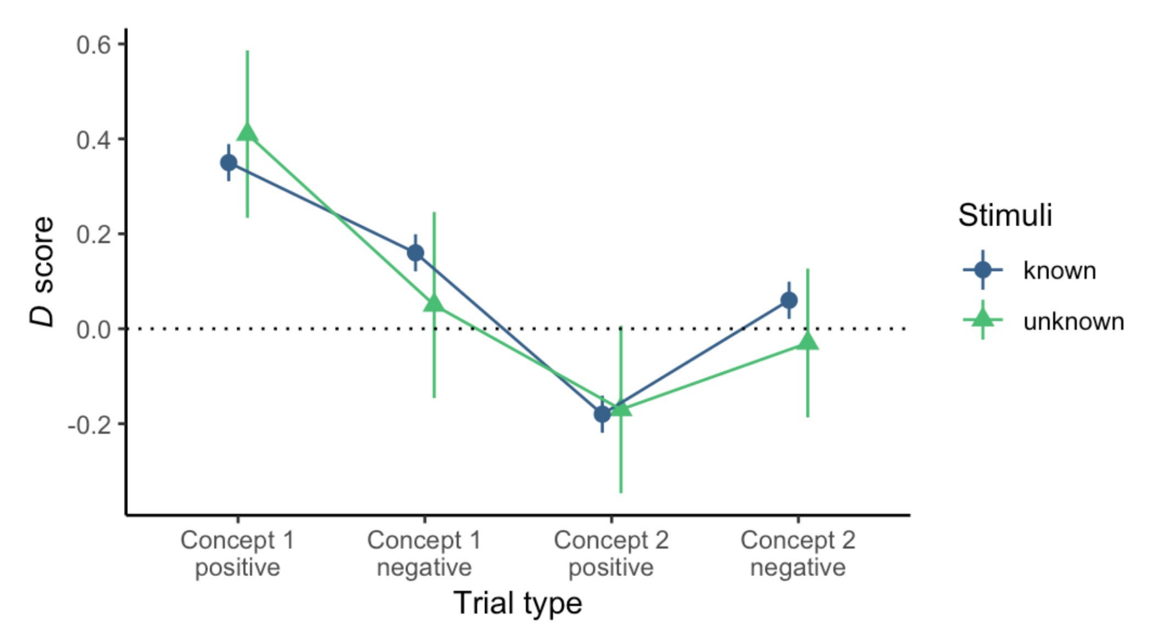
- O'Shea et al. (2015) called it a "positive framing bias"
- Finn, Barnes-Holmes & McEnteggart (2018) called it the "single trial-type dominance effect"

Existence is uncontroversial

Largest IRAP dataset to date:
N = 501 completions of evaluative IRAPs



Generic Pattern (all BF < 0.6)



Generic Pattern even with non-word stimuli (all BF < 0.16)

(Replication of O'Shea et al., 2015)

Generic Pattern

Presence of IRAP effects has
little to do with the domain being assessed

And **everything** to do with the Generic Pattern

	η^2
Main effect for trial type (Generic Pattern)	0.74
Main effect for domain	0.04
Interaction effect	0.02

False conclusions

Non-zero D scores ≠ domain specific biases

Original text

- *"Participants were quicker to endorse the belief that White people are positive (White–Positive), t(19) = 4.12, p = .001"* (Hughes, Hussey, et al., 2016)

Correct interpretation

- *"Participants demonstrated an IRAP effect, t(19) = 4.12, p = .001"*

Less theoretically interesting

False conclusions

Examples are ubiquitous

- ~~Death~~ - positive (Hussey, Daly & Barnes-Holmes, 2015)
- ~~White people~~ - positive (Hughes, Hussey, Barnes-Holmes, 2016)
- ~~Thin~~ - positive
- ~~Attractive~~ - positive
- ~~Clean~~ - positive (Hughes, Hussey, Barnes-Holmes, 2016)

False conclusions

Some studies’ conclusions rely **exclusively** on the presence of IRAP effects

- Finn, Barnes-Holmes, Hussey, & Graddy (2016, Studies 1 & 3)

How prevalent are these false conclusions?

A systematic review

How prevalent are these invalid inferences?

A systematic review of **every published empirical IRAP study**

- Followed PRISMA guidelines
- 102 empirical IRAP articles found (to end of 2018)
- Rated by two independent teams
    - Inter-rater agreement = 98%

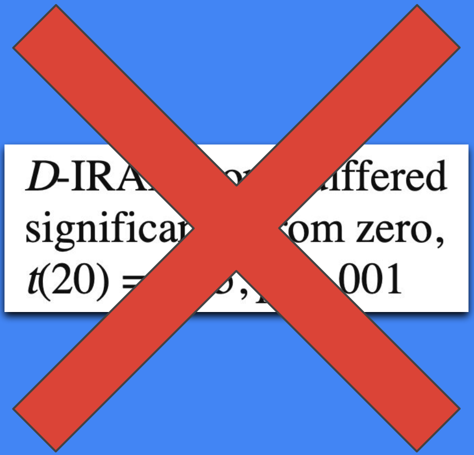
84%

of published IRAP papers
contain false conclusions
from misinterpreting IRAP effects

78%

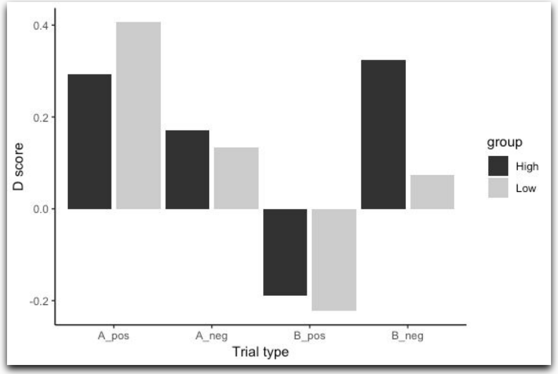
of published IRAP papers
contain false **core conclusions**
from interpreting IRAP effects

3 most common ways to analyse IRAP data



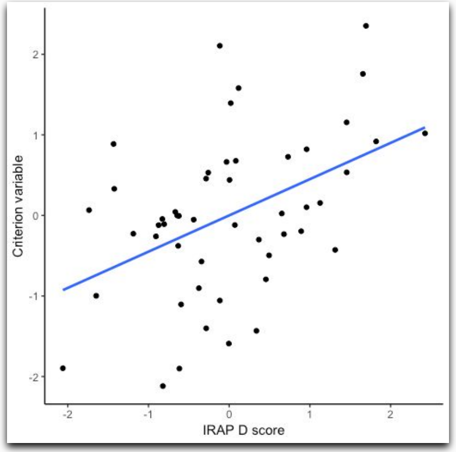
1

Presence of IRAP effects



2

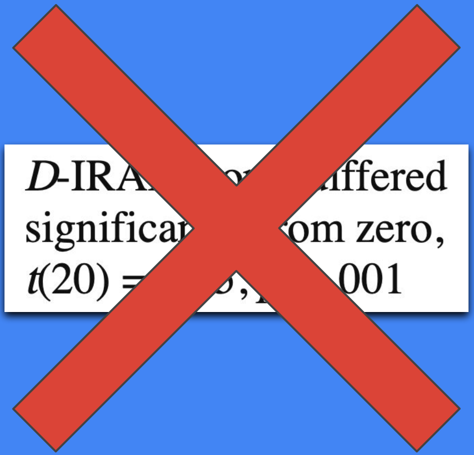
Mean differences in IRAP effects



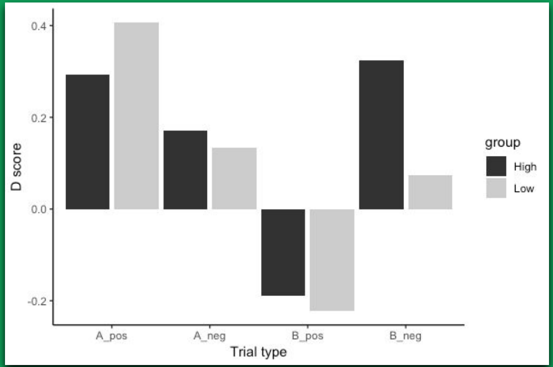
3

IRAP effects correlating with other variables

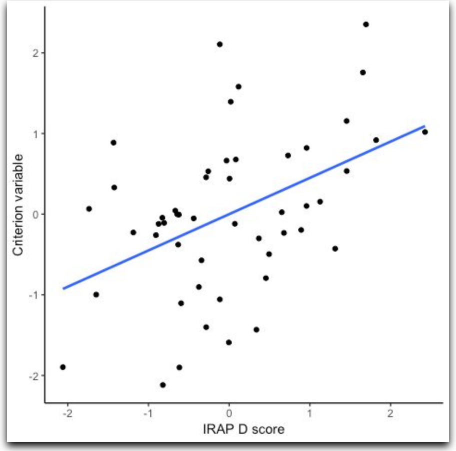
3 most common ways to analyse IRAP data



1
Presence of IRAP effects



2
Mean differences in IRAP effects



3
IRAP effects correlating with other variables

False Positives due to Researcher Degrees of Freedom

Computational simulation studies

Modal research practices

Systematic review shows **Median IRAP study** is:

- A between 2 groups design
- *N* = 18 per group
- Frequentist inferential statistics (*p* values)

Studies are relatively homogeneous, no evidence of change over time (all *ps* > .12)

Large number of “Researcher Degrees of Freedom”

Researcher degrees of freedom

Hidden choices that researchers make, even without meaning to, that lead them to find significant effects where there are none

Ie Allow presenting **anything** as significant

- Simmons et al. (2011) *False-Positive Psychology*

Researcher degrees of freedom

Observed in the IRAP literature

ANOVAs choices

- Between trial-type
- Between timepoints/groups
- Interaction effects
- Significance from zero tests

Inclusion/exclusion criteria

- Mean vs median latency
- Block vs participant

Correlational choices

- Implicit-explicit correlations
- Regressions

Modelling choices

- Trial type collapsing
- Trial type inversions
- Block order
- Implicit/explicit order
- IRAP order

Researcher degrees of freedom

Observed in the IRAP literature

Simulation

ANOVAs choices

- Between trial-type
- Between timepoints/groups
- Interaction effects
- Significance from zero tests

Inclusion/exclusion criteria

- Mean vs median latency
- Block vs participant

Correlational choices

- Implicit-explicit correlations
- Regressions

Modelling choices

- Trial type collapsing
- Trial type inversions
- Block order
- Implicit/explicit order
- IRAP order

Simulation studies

Powerful analytic technique to study false positive rates

Only assumptions are those of the tests they examine (eg ANOVAs)

Procedure:

- Generate null data
- Run test (ANOVA)
- Repeat 10,000 times
- Observe proportion of significant results

Acceptable FPR
implied by alpha = 0.05

5%

False Positive Rate due to
Researcher Degrees of Freedom
in IRAP research

>44%

	Significant results	Non-significant results
True effect exists	True positives	False negatives
True effect Is null	False positives	True negatives

	Significant results	Non-significant results
True effect exists	True positives	False negatives
True effect Is null	False positives	True negatives

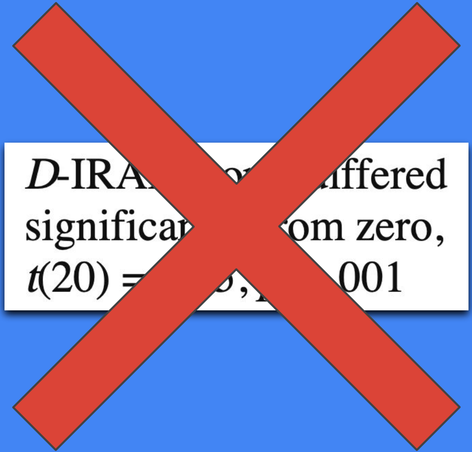
Researcher Degrees of Freedom +
Inferences made from the Generic Pattern

Low statistical power

Implications

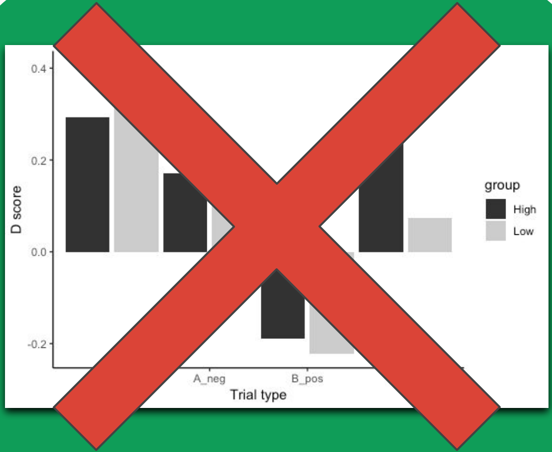
IRAP literature is likely to have many false findings

3 most common ways to analyse IRAP data



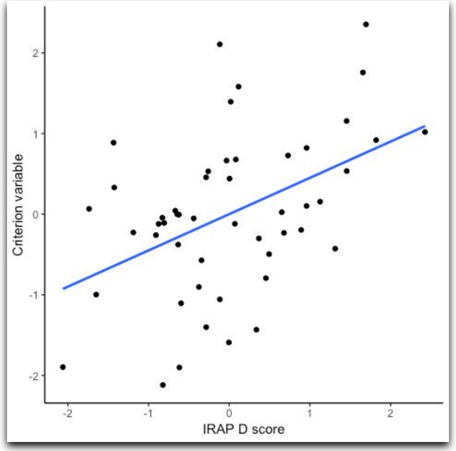
1

Presence of IRAP effects



2

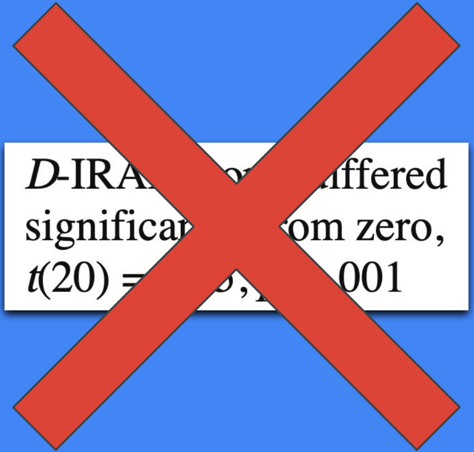
Mean differences in IRAP effects



3

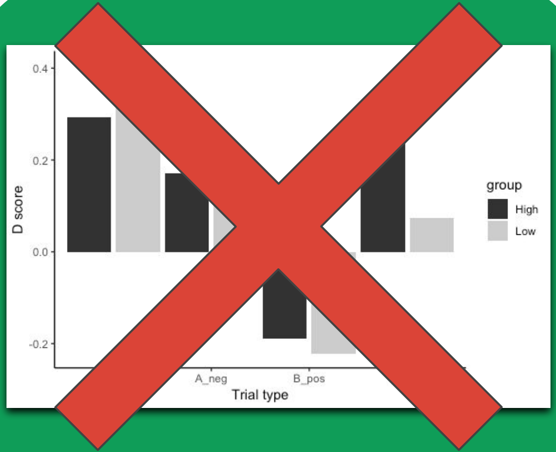
IRAP effects correlating with other variables

3 most common ways to analyse IRAP data



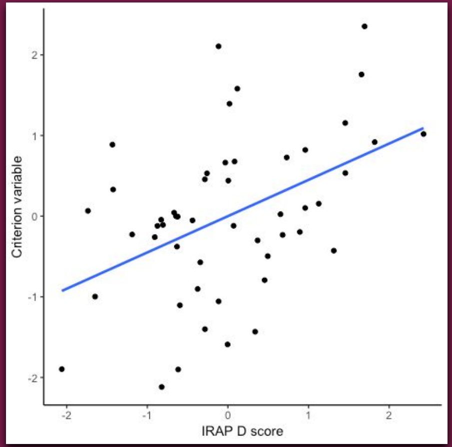
1

Presence of IRAP effects



2

Mean differences in IRAP effects



3

IRAP effects correlating with other variables

IRAP’s Predictive Validity

Updating a recent meta analysis

IRAP’s Predictive Validity

Meta-analysis of association between IRAP & clinically-relevant criterion effects

- Vahey, Nicholson & Barnes-Holmes’ (2015)

Widely-cited for sample size justifications

- 66 citations
    - 39 new IRAP papers in this period

"the Ns involved in the studies ... are often relatively small.

Indeed, it could be argued that this impacts upon on the credibility of IRAP research.

However, in a recent meta-analysis of IRAP studies, it was reported that even small *N* IRAP studies have sufficient statistical power" (McEnteggart, 2015) "the Ns involved in the studies ... are often relatively small.

Indeed, it could be argued that **this impacts upon on the credibility of IRAP research**.

However, in a recent meta-analysis of IRAP studies, it was reported that even small *N* IRAP studies have sufficient statistical power" (McEnteggart, 2015)

Excluded problematic analyses

50% of effect sizes (7 of 15) were excluded

Mere presence of IRAP effects

- Widely misinterpreted due to Generic Pattern
- Not an external criterion (Flake et al., 2017)

IRAP as dependent variable

- No clinical assessment utility (Fried & Kievit, 2016)
- Incompatible with their meta analysis modelling approach
    - Correct multivariate meta: $(Y_1, Y_2) \sim \text{IRAP}$
    - Vahey et al. method: $\text{IRAP} \sim (X_1, X_2)$

IRAP's Predictive Validity

Excluded problematic analyses

Meta analysis via Hunter & Schmidt method

		95% CI	
	r	Lower	Upper
Original	.45	.40	.54
Updated	.39	.27	.51

Sample size recommendations

80% power for a bivariate correlation

	Required *N*
Original	37
Updated	105

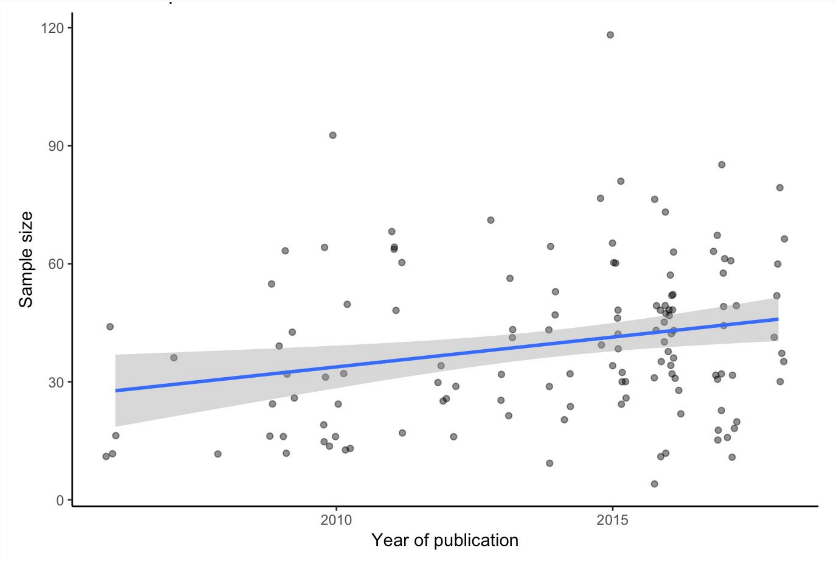
Most IRAP research is under-powered

	% of under-powered published studies
Original	50%
Updated	93%

Most IRAP research is under-powered

If current rate increase in sample sizes continues,

The average study won't be well-powered until the year **2051**



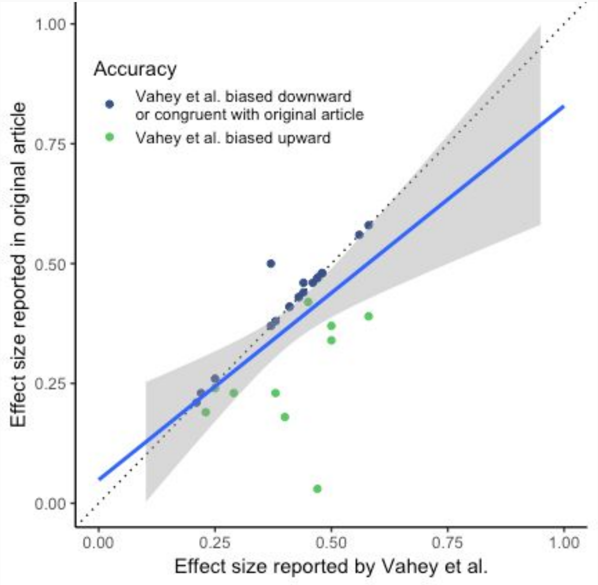
A new meta analysis

Following best practices

Issues with original meta analysis

Effect size extraction errors

- Incongruities in 33% of cases
- Biased upwards



Issues with original meta analysis

Hypothesizing After Results are Known (HARKing)

- Inclusions based on what the meta authors thought **could have been predicted** ahead of time, not what the original authors **actually predicted**

No blinding

- Researchers knew the effect size when choosing them

Issues with original meta analysis

Only relevant to deductive research

- Meta analysis of *predictable effects* can only inform future research that is making *predictions*
- But current this *deductive* meta is now inappropriately cited in *inductive* research to justify sample sizes

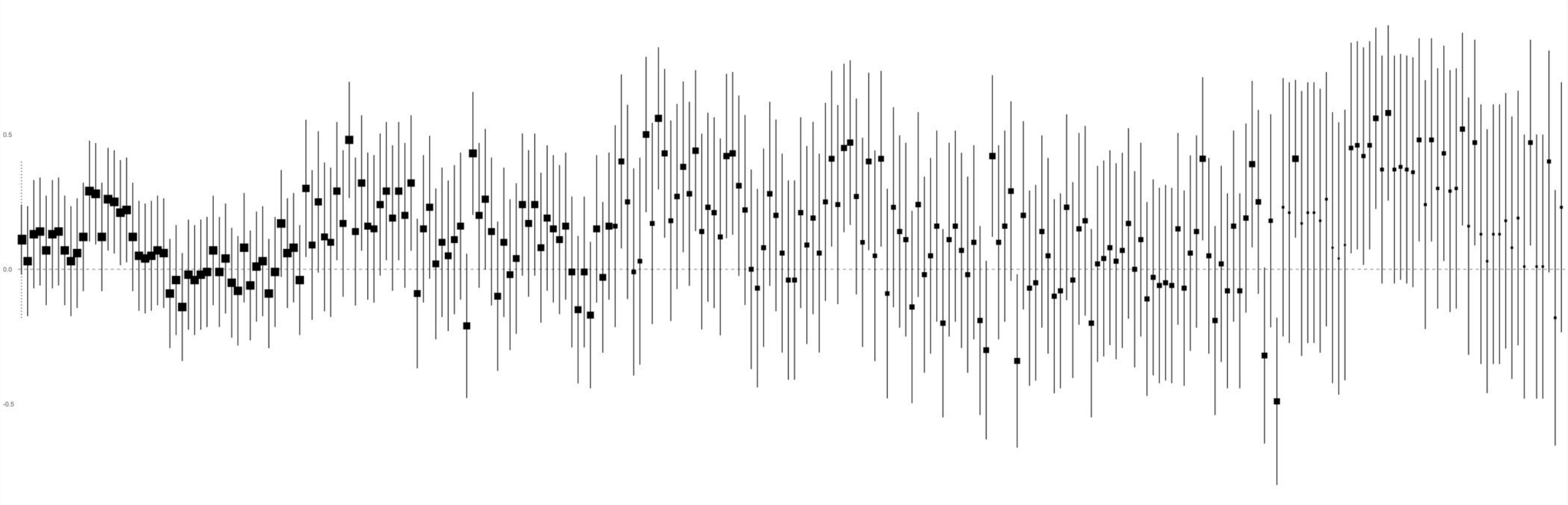
IRAP's Predictive Validity

A meta-analysis for inductive research

Modern meta-analytic best practices

- Multilevel meta analysis
- Restricted Maximum Likelihood estimation & *N* weighting
- Considered same articles as original meta
- Included all 249 effect sizes
    - Other than previously specified problematic analysis types

Sample size recommendations



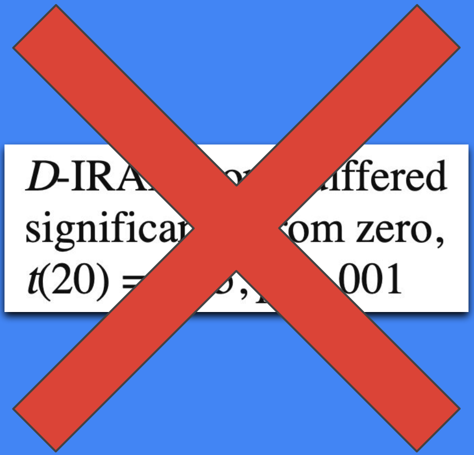
Meta-effect size: $r = .11$, 95% CI [-.02, .24], $p = .10$.

Sample size recommendations

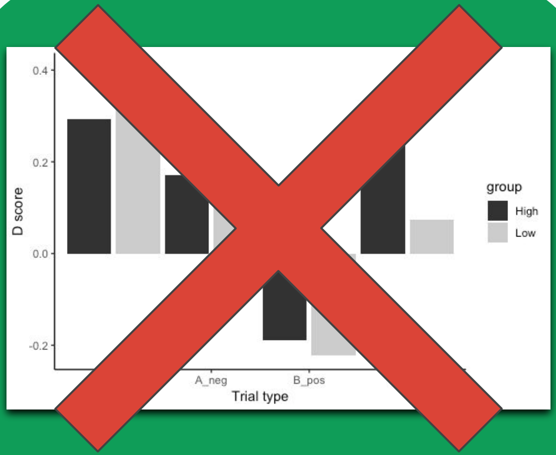
80% power for a bivariate correlation:

	Required *N*
Original	37
Updated	105
New	19,620

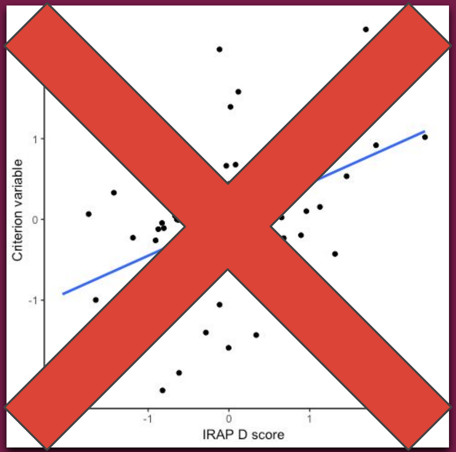
3 most common ways to analyse IRAP data



1
Presence of IRAP effects



2
Mean differences in IRAP effects



3
IRAP effects correlating with other variables

3 most common ways to analyse IRAP data

84% of papers make false inferences

False positive rate due to Researcher Degrees of Freedom is >44%

>50% of literature is underpowered

Multiple issues with existing meta analysis

N > 107 needed for adequate power

New meta suggests low predictive validity

Conclusion

There is a problem with ~~the IRAP~~ our research practices

The way forward

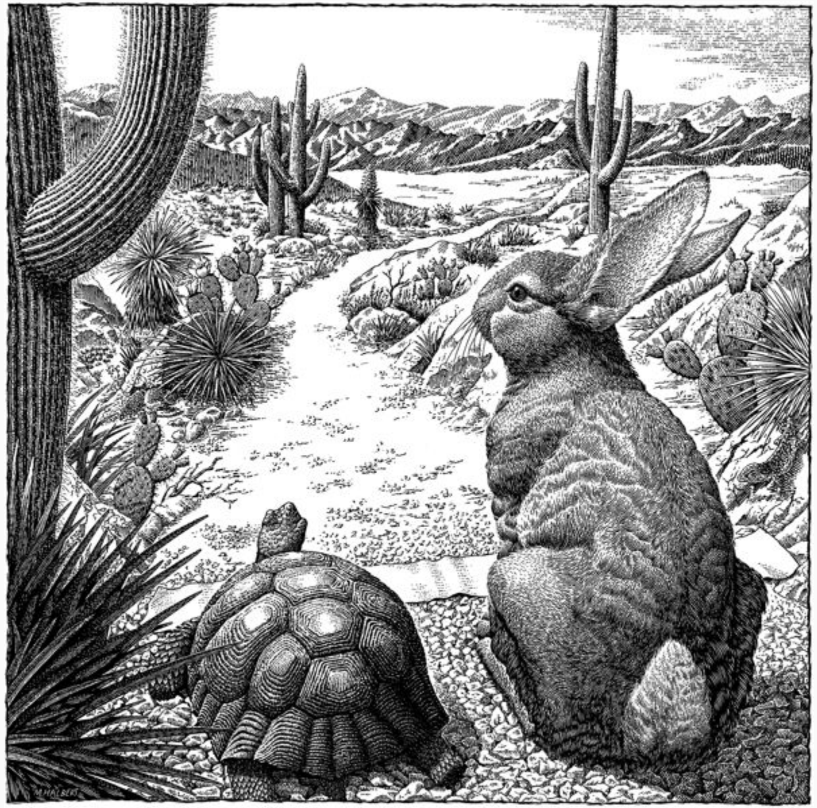
Our research practices are not exceptional

Real issue: we're resistant to change

- Other fields have had their crisis, are now 8 years into recovery

There's still time to fix this!

- More power, better use of statistics, pre-registration, direct replication
    - See Munafò et al. (2017) *A manifesto for reproducible science*



Tortoise vs. Hare approaches to science

Data & code:

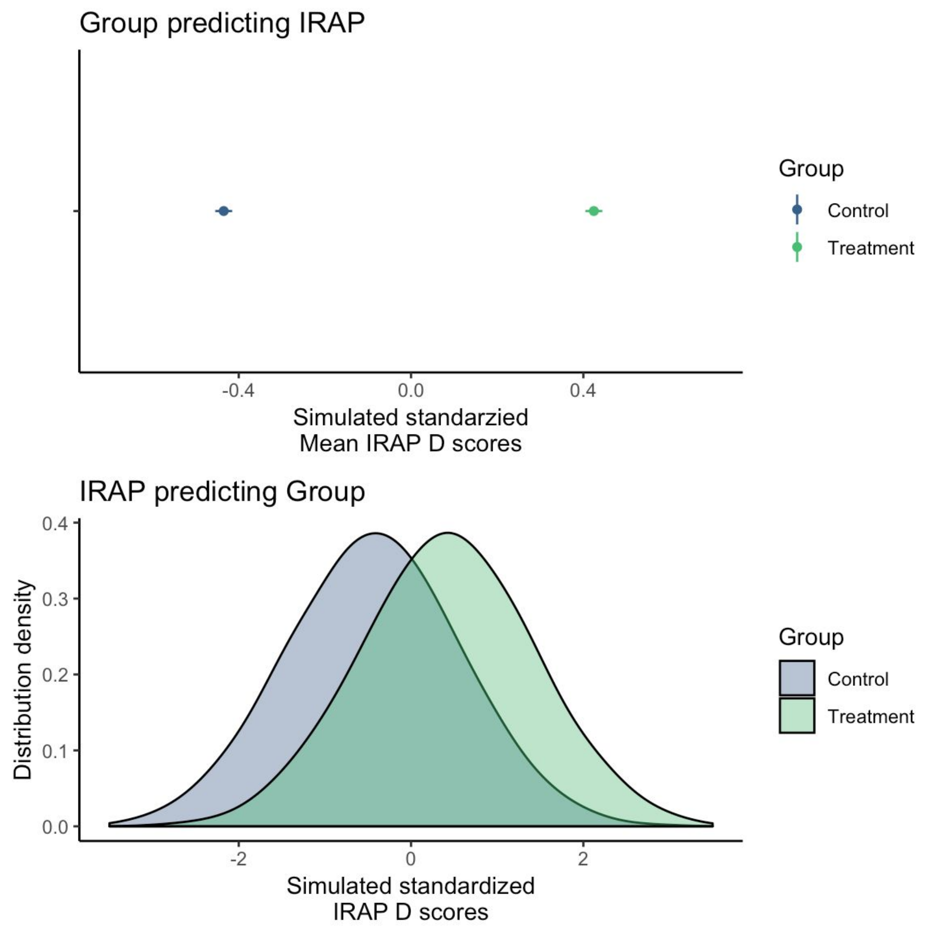
osf.io/ke7zx

Ian Hussey

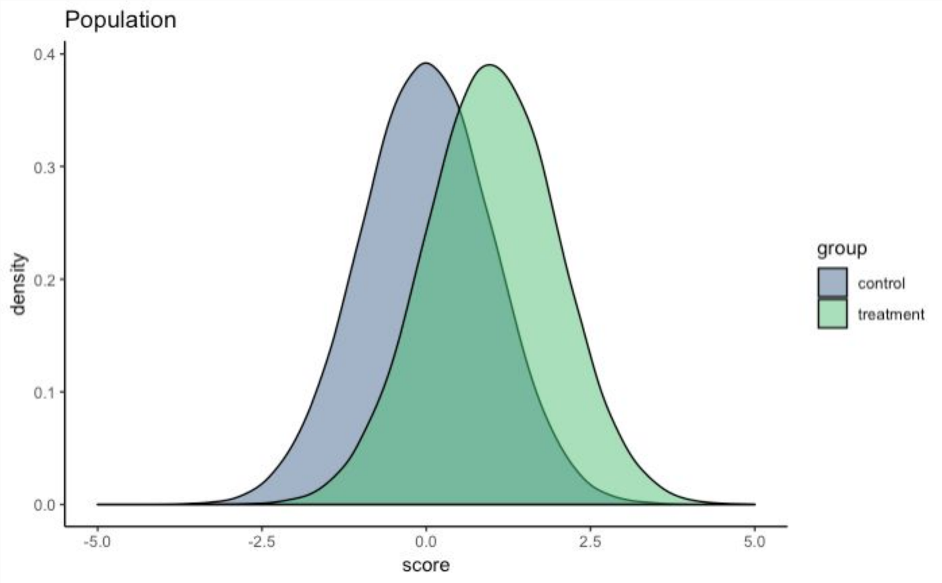
Postdoctoral research fellow
Ghent University
Belgium

ian.hussey@ugent.be
@ianhussey

	η^2	η_G^2	η_p^2
Trial type	0.74	0.19	0.26
Domain	0.04	0.01	0.04
Interaction	0.02	0.01	0.01


Group predicting IRAP
-0.4
0.0
0.4
Simulated standarzied
Mean IRAP D scores
Group
Control
Treatment
IRAP predicting Group
0.4
0.3
0.2
0.1
0.0
Distribution density
-2
0
2
Simulated standardized
IRAP D scores
Group
Control
Treatment

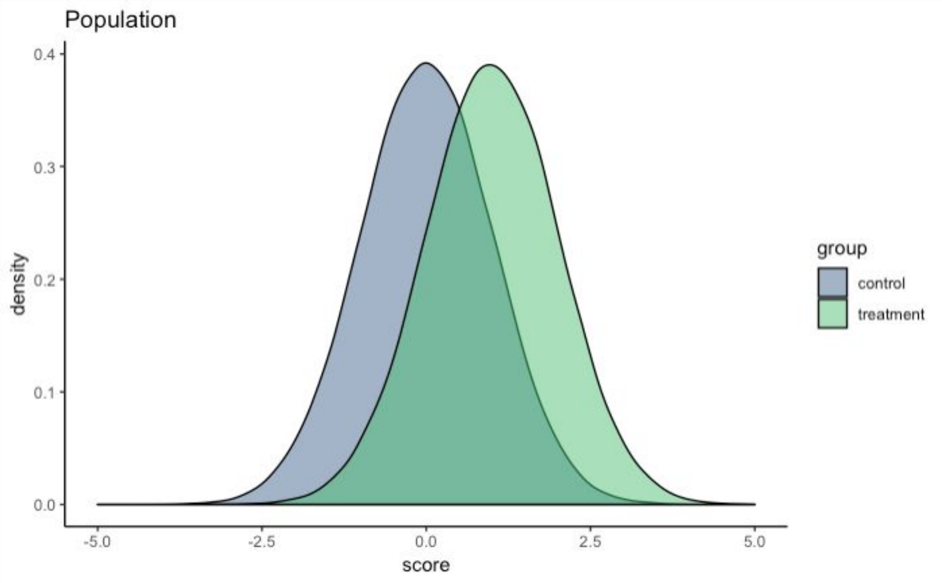
Population

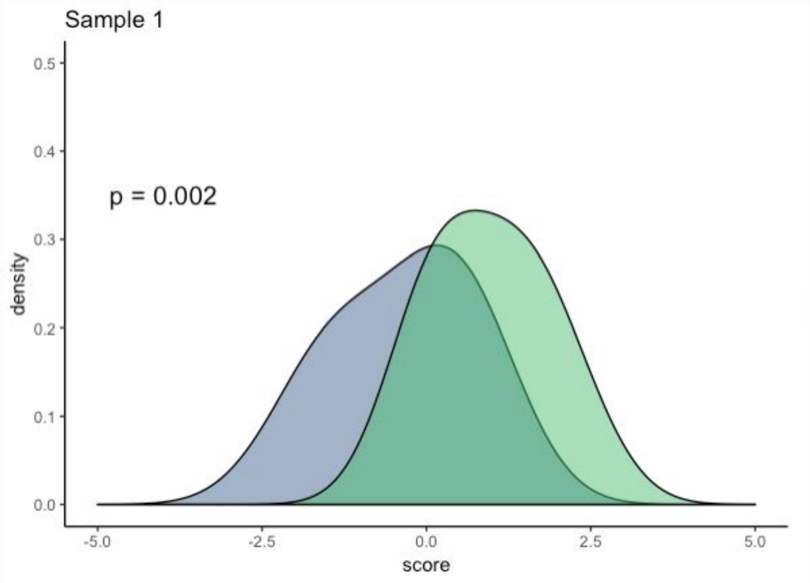


True effect

Simple inferential stats:
the *t* test

Population → Sample 1

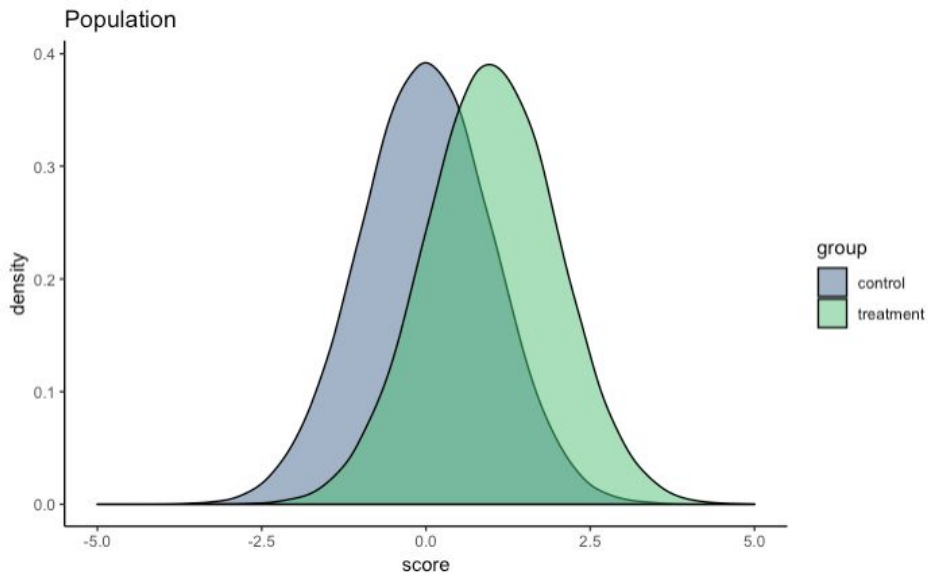


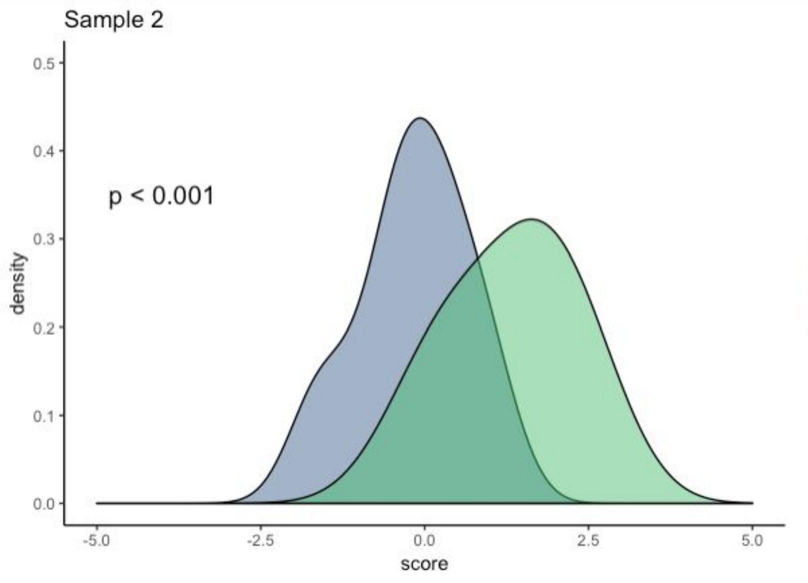


True effect → *True positive*

Simple inferential stats:
the *t* test

Population → Sample 2

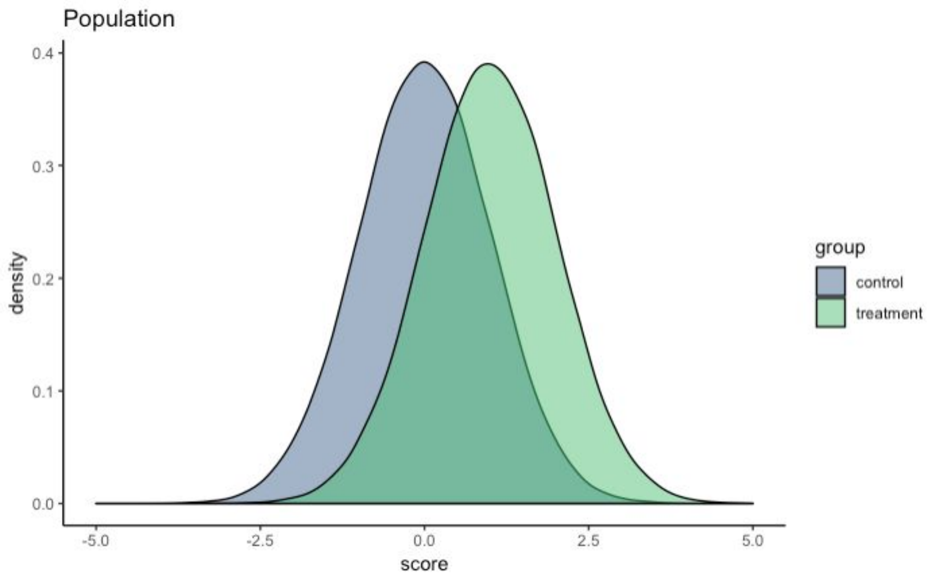


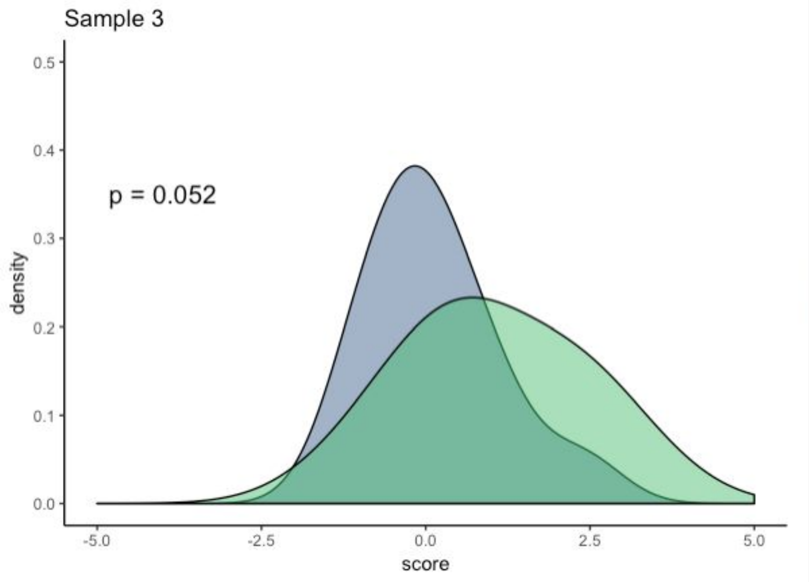


True effect → *True positive*

Simple inferential stats:
the *t* test

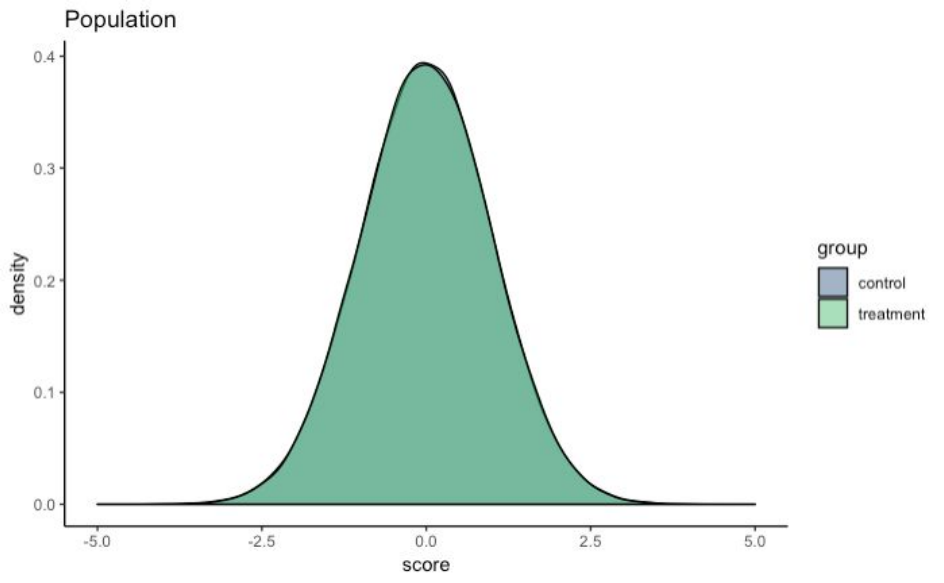
Population → Sample 3





True effect → False negative

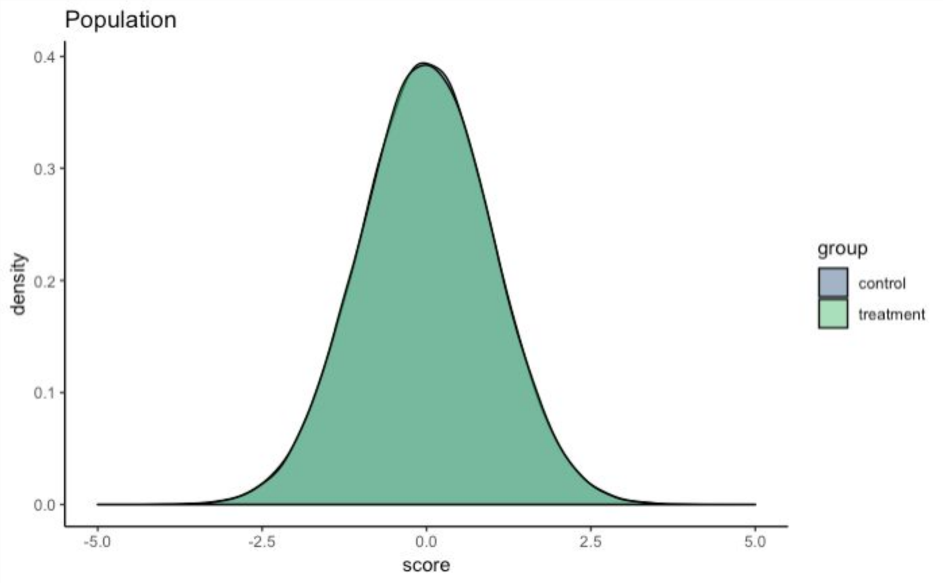
Population

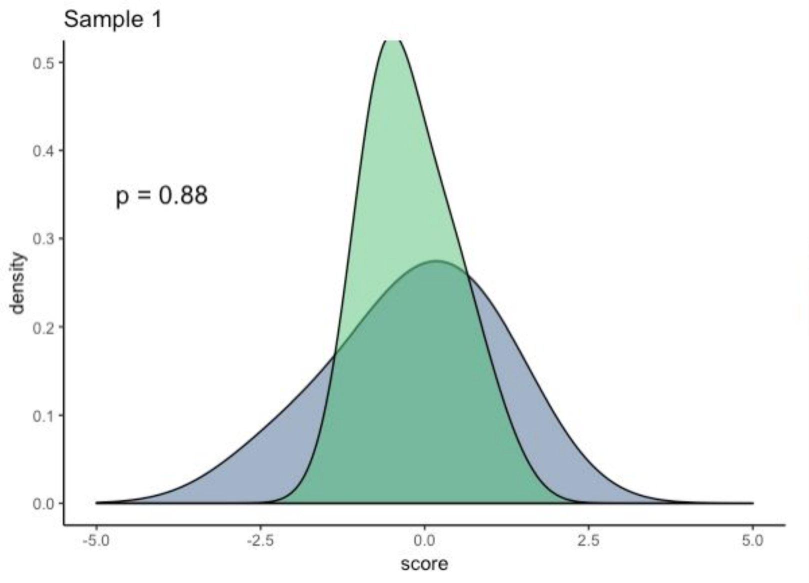


True null

Simple inferential stats:
the *t* test

Population → Sample 1

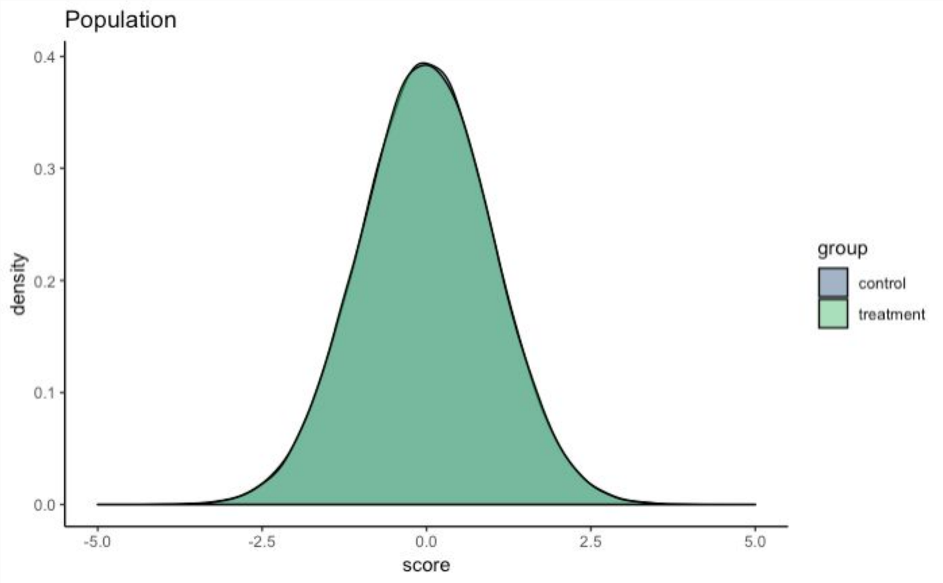


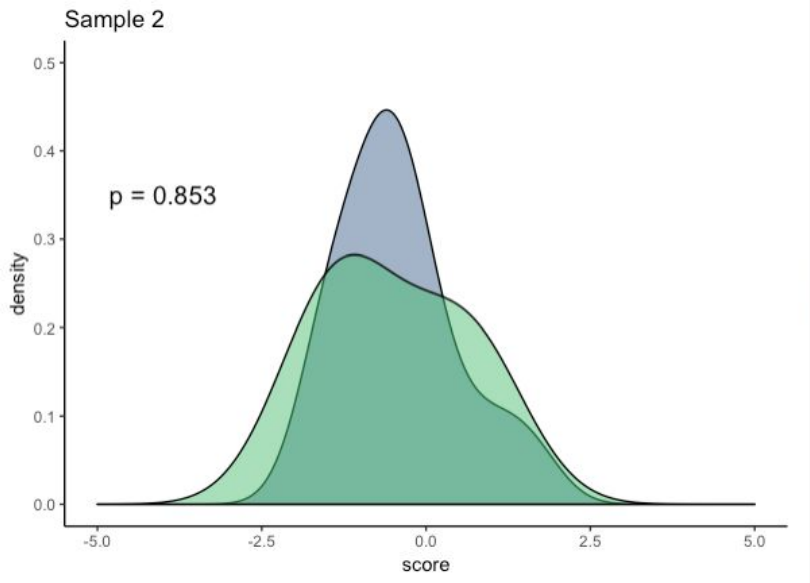


True null → *True negative*

Simple inferential stats:
the *t* test

Population → Sample 2

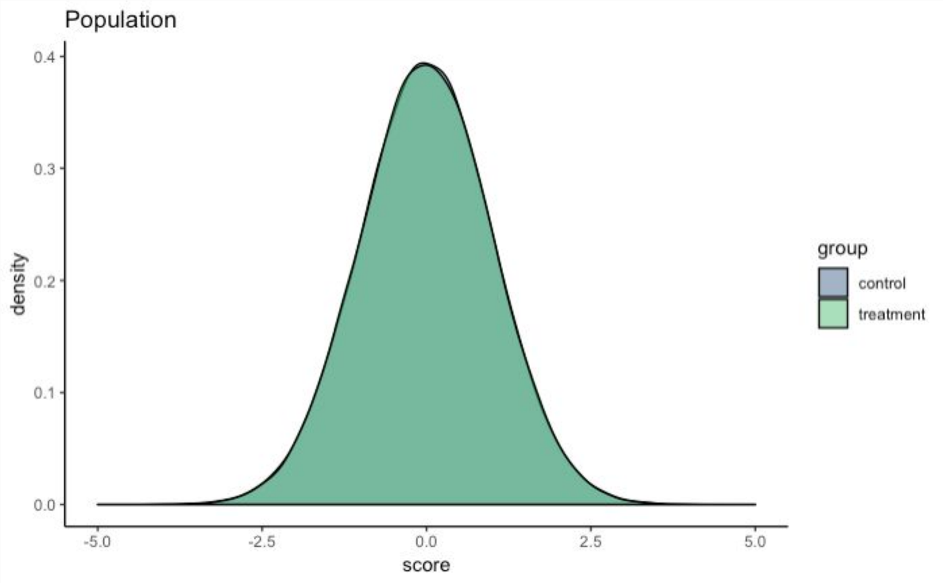


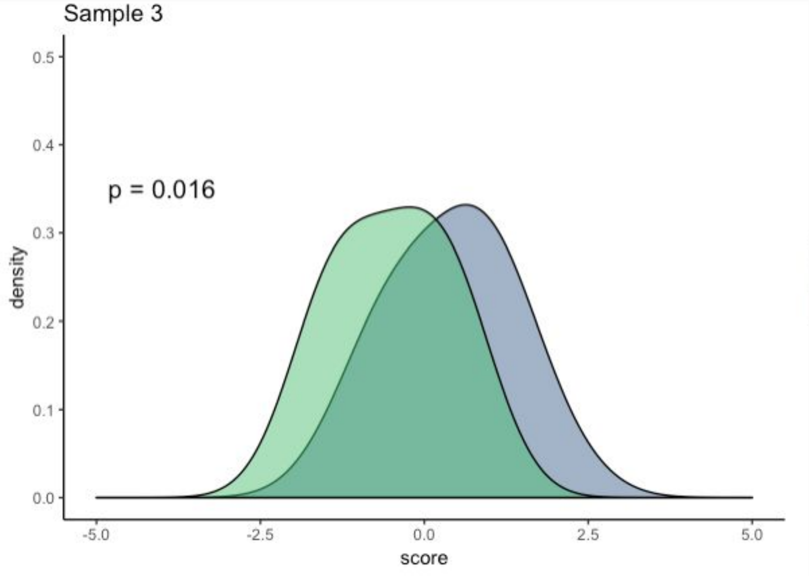


True null → *True negative*

Simple inferential stats:
the *t* test

Population → Sample 3

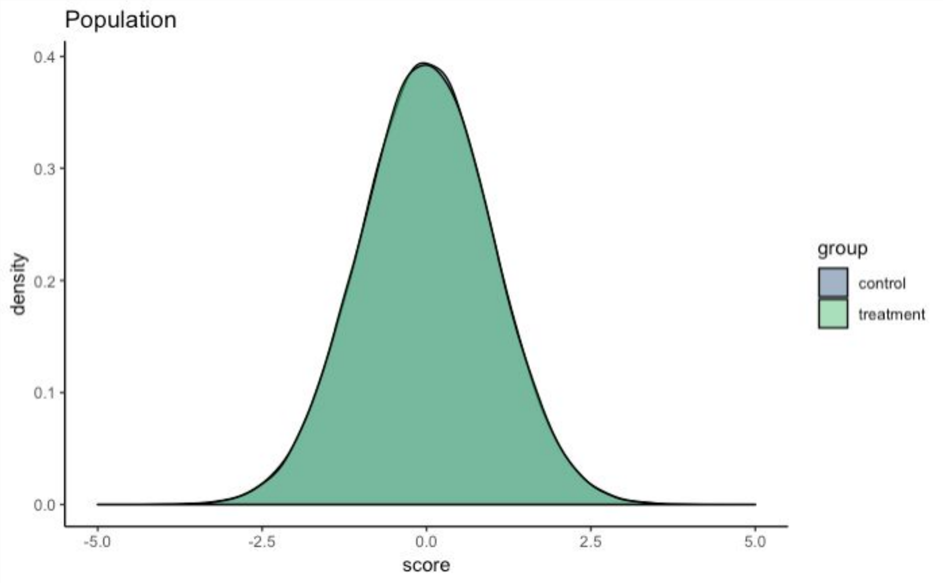




True null → *False positive*

Simulations studies
of the False Positive Rate (FPR)

Population



Sample 1

Sample 2

Sample 3

...

Sample 10,000

True null

5% False Positives

alpha = .05

FPR due to
Generic IRAP pattern:

66%

(median published IRAP study)

FPR due to
Generic IRAP pattern:

100%

(well-powered studies)

HARKing/blinding examples

Vahey et al. (2009)

- Students vs. main-block prisoners (*r* = .46, included)
- Students vs. open-air prisoners (*r* = -.04, excluded)

Timko et al. (2011)

- Non-dieters vs. dieting-to-lose-weight (*r* = .45, included)
- Non-dieters vs. dieting-to-maintain-weight (*r* = .13, excluded)

Sample size recommendations

80% power for a bivariate correlation

	Required N
Original	29
Updated	49
New	646

continue to monitor the homogeneity of their constituent criterion effects. Third, the meta-effect is an estimate based upon an IRAP literature that is currently still evolving, and so as per its accompanying credibility interval (.23, .67), there is a degree of uncertainty about whether it might be subject to over- and/or under-estimation.

Fortunately, the literature has very recently suggested a number of ways of statistically accounting for the possibility of meta-effect under- and/or over-estimation when calculating the sample size required for a given statistical test at a given level of statistical power. Adopting a conservative approach in favour of controlling for overly optimistic publication biases, the most recent recommendation is to calculate sample size requirements not in terms of a given meta-effect, but rather in terms of the lower bound of its associated confidence interval (Perugini, Gallucci, & Costantini, 2014). Given that we obtained a confidence interval of (.40, .54) around the present meta-effect, Perugini et al.'s approach implies that a sample size of at least $N = 37$ would be required in order to achieve a statistical power of .80 when testing a continuous first-order correlation between a clinically-focused IRAP effect and a given criterion variable (i.e. as opposed to $N = 29$ without Perugini et al.'s correction). Likewise, Perugini et al.'s method implies that Ns

Applying exclusions to inductive meta

Vahey and colleagues included effects only when:

(i) criterion variable was of direct clinical relevance, and

(ii) the IRAP could have been predicted to be related to this criterion.

Two independent raters coded our inductive meta-analysis effects based on these criteria

Meta-effect size: $k = 150$, $r = .13$, 95% CI [.03, .23], 95% CR [-.10, .36], $p = .01$