Journal of Contextual Behavioral Science

A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using

the Implicit Relational Assessment Procedure

--Manuscript Draft--

Manuscript Number:	JCBS-D-23-00021R1
Article Type:	Review Article
Keywords:	Systematic Review; statistical power; meta-science; implicit relational assessment procedure; relational frame theory; implicit social cognition
Corresponding Author:	Ian Hussey Ruhr University Bochum GERMANY
First Author:	Ian Hussey
Order of Authors:	Ian Hussey
Abstract:	Following recent calls to examine the replicability of behavioral research, I examine two key determinants of replicability, sample sizes, and statistical power, in research using the Implicit Relational Assessment Procedure (IRAP). A systematic review was used to gather all published studies employing the IRAP and extract their designs and sample sizes. The use of Null Hypothesis Significance Testing was found to be nearly ubiquitous, justifying the examination of statistical power. Using an established method, median sample sizes were used to estimate the statistical power to detect the average published effect size in psychological research (r = .20) in each year. Sample sizes and the statistical power they imply were found to be very low in IRAP studies (in 2022, median N = 64, power = 34%). At the current rate of growth, the median study will only reach the recommended minimum power of at least 80% by 2080. The IRAP literature was directly compared with the Social and Personality psychology literature using an existing dataset. Median sample sizes and their implied statistical power were lower in the IRAP literature in all years than they were in Social and Personality psychology at the beginning of the Replication Crisis in 2011, and indeed in all subsequent years. Improvements in sample sizes and statistical power in the Social and Personality psychology literature were significantly and substantially larger than in the IRAP literature. Direct tests of the reproducibility and replicability of claims in the IRAP literature are needed.
Suggested Reviewers:	R. Chris Fraley, PhD Professor, University of Illinois Urbana-Champaign rcfraley@illinois.edu Prof Fraley created the method that is used in the current work, as well as the dataset that is reused in the current article. He therefore has expertise in power analysis and the implications of power.
	Simine Vazire, PhD Prof, The University of Melbourne simine@gmail.com Prof Vazire is a well-known meta-science researcher and co-author with Prof Fraley of the method that the current paper employs. She therefore has expertise in power analysis and the implications of power.
	Jason Luoma, PhD Oregon Health & Science University jbluoma@portlandpsychotherapy.com Dr Luoma has conducted meta-science research within CBS.
	Cody Christopherson, PhD Professor, Southern Oregon University christoc1@sou.edu Prof Christopherson has published both open science and CBS research.

	Donny Newsome PhD, Fit Learning New York DrDonny@FitLearners.com Dr Newsome has published meta-scientific reviews of CBS research in JCBS (Newsome et al, 2018, How Contextual Behavioral Scientists Measure and Report About Behavior: A Review of JCBS)
Opposed Reviewers:	
Response to Reviewers:	

Dear Prof Levin, Prof Rogge, and anonymous peer reviewers,

Thank you for your very helpful reviews of my manuscript.

I have provided point-by-point responses to each of your comments below. To aid your understanding of the original comment and my changes, I have labeled each actionable comment in square brackets, e.g., “[reviewer 1 comment 1]”. My response is then provided in italics below, followed by quotes from the changes made to the manuscript.

I have made a few substantive additions to the manuscript based on your excellent comments, and I think they have improved the manuscript.

Kind regards,
****

Manuscript Number: JCBS-D-23-00021

A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using
the Implicit Relational Assessment Procedure

Dear ****,

Thank you for submitting your manuscript to the Journal of Contextual Behavioral Science. The AE, Dr. Rogge, received 2 reviews of your manuscript and provided his own feedback. Based on these reviews (found below) we have decided to ask you to make minor revisions and resubmit. We are asking that the revised manuscript be submitted by Aug 11, 2023.

When you resubmit your manuscript please take care to note all comments along with how they were addressed or why they were not addressed in a separate Response to Reviewers file. Also, please ensure that no identifying information is included in the Response to Reviewers, as this would unmask the reviewers and delay processing of your manuscript considerably. For example, do not sign the Response to Reviewers or provide it on letterhead.

To submit your revised manuscript, please log in as an author at https://www.editorialmanager.com/jcbs/, and navigate to the "Submissions Needing Revision" folder.

Thank you for the opportunity to consider your work. Please contact me, Michael Levin, if you have any concerns or questions about this decision, the revision process, or about JCBS in general.

Regards,

Michael Levin
Editor-in-Chief
Journal of Contextual Behavioral Science

Associate Editor and Reviewer comments:

Dear authors,

Thank you for submitting such an interesting review of the power and sample sizes within the IRAP literature. I feel that your manuscript is an excellent extension of the work of Fraley and colleagues (2022). More importantly, the comparison of the growth of sample sizes in the social literature and the IRAP literature in the last decade (i.e., since the discovery of the replication crisis) is a bit alarming and provides critical insights toward understanding limitations of that work and how future research on the IRAP needs to be improved. I have received two reviews of your manuscript and the reviewers agree with me on the importance of this work and its potential impact on the field. Both reviewers have provided comments and suggestions to help strengthen your work and I have a suggestion for improvement as well. Thankfully, most of these revisions involve fairly minor additions and revisions to your narrative. I am therefore happy to recommend a revise and resubmit with minor revisions at this point.

Despite my enthusiasm for your manuscript, I want to note that this recommendation does not guarantee the ultimate acceptance of your manuscript. As with all papers in peer-review, the level of review following this revision and the ultimate decision on this manuscript depend heavily on the depth and responsiveness of your revisions to all of the comments raised as well as the level of detail provided in describing and explaining the revisions made within your response letter.

I look forward to reading your revised manuscript and I hope the comments raised in this peer-review are helpful in revising it. Please let me know if you have any questions.

Sincerely,

Ron Rogge
Associate Editor, JCBS

[AE comment 1]

ASSOCIATE EDITOR COMMENT:
The sentence after you first introduce Figure 2, you explain what is in the graphs: “In this and all subsequent figures, the straight line represents the fitted Ordinary Least Squares linear regression line (discussed later) and the shaded region around it represents its 95% Confidence Interval.” The second half of that sentence does not make sense as none of your line graphs have shaded regions. Please delete that portion of the sentence.

Author response: Thank you for noticing this. The pdfs of these plots do indeed have shaded regions, but something about the editorialmanager pdf creation process has *rendered them transparent. In addition to embedding the plots inside the manuscript I have also attached them as separate PDFs, in which you can see the shaded regions.*

REVIEWER 1 COMMENTS:
The purpose of this article was to review the empirical literature using the Implicit Relational Assessment Procedure, with a focus on the typical sample sizes used. The authors argue that, in the wake of the replication crisis in psychology, there has been a greater interest in improving research methods. One such call, which actually goes back several decades, is for greater statistical power. Power is useful in empirical research because, on the one hand, underpowered work cannot detect real effects that exist. But, it is also important because, in literatures where the typical power of studies is low, the false positive rate will be higher and published effect size estimates will be inflated.

The author finds that the typical sample size in work using the IRAP is close to 64 and that the power of a typical study to detect a typical effect size of r = .20 is only 34%. Moreover, although there is some evidence that sample sizes may be increasing, they are increasing at a rate that, frankly, is unacceptable in light of these debate.

I think this paper addresses an important topic and does so in a way that seems systematic, comprehensive, and credible. As such, it is my guess that this paper would make a fantastic contribution to the literature. I don't have any substantive criticisms. I have a couple of constructive suggestions below; use them if you find them helpful.

[reviewer 1 comment 1]

1. The author is reluctant to make specific recommendations for sample sizes. I understand why. Nonetheless, I do think the natural arc for a paper like this is to end with recommendations (i.e., identify a problem, explain why it is important, make some recommendations) that people can easily follow. If the bottom line is "think about power in a way that you think is sensible for the problem you're studying," most researchers (many of whom are, ironically, not adept at statistics and methods) will make poor decisions. They will, for example, assume the effects they are studying are likely to be much larger than that "other" people study. I can't think of any good reason why researchers in 2023 should not--as a bare minimum--power their studies to be capable of detecting an effect of r = .20 or higher. The costs of doing underpowered research are simply too high to try to take short-cuts based on unknown or assumed population parameters.

Author response: I have added worked examples of power analyses for r = .20 (as used elsewhere in the manuscript) as well as Cohen’s d = .46, which represents the average transdiagnostic efficacy of ACT and therefore a potentially useful benchmark for intuitive comparison. Page 26-28 now reads:

"While keeping the above numerous qualifiers in mind as to why power calculations must be done for each study with respect to its design, measurement quality, and specific analysis, some readers may nonetheless wish for some discussion of the sample size needed to detect specific effects in future IRAP studies. I will therefore take two examples, both related to common use cases for the IRAP.

First, imagine a researcher who wishes to detect a correlation between IRAP scores and another continuous variable. Power analyses were again conducted using the R package pwr (Champely, 2016). To detect a correlation that is the magnitude of the average effect size observed in published psychological research (i.e., Pearson's r = 0.20); with an α-level of 0.05 and a two-sided test (typical for correlations and regression); 194 participants would be needed to obtain the minimum recommended power (80%), or 319 participants for high power (95%). Comparing these sample sizes to the sample sizes observed in the published IRAP literature, 2% of published IRAP studies employed a sample size sufficient to detect the average effect size in the broader psychology literature (r = .20) with minimum recommended power (80%), and 0% could detect this with high power (95%). Note that half of all published effects are smaller than the average, and therefore would require even higher sample sizes to detect.

Second, imagine that a researcher wishes to detect a difference in mean IRAP scores between two groups, and they wish to power the study to detect effect sizes that are at least the average transdiagnostic efficacy of Acceptance and Commitment Therapy versus active and inactive control groups as reported in a recent review of meta-analyses (Hedges' g = 0.46: Gloster et al., 2020). This effect size is no more or less meaningful than any other for the sake of illustration but is chosen to leverage intuitions the reader may have to define its Smallest Effect Size of Interest. That is, such a study would be powered to detect mean differences on the IRAP between groups that are no smaller than the general efficacy of ACT. Loosely speaking, I am inviting the reader to think about other effect sizes they have an intuition for to provide a benchmark. For example, if you are interested in between groups differences that are smaller than the average efficacy of ACT, you would need more participants than this again. With an effect size of Cohen's d = .46 (i.e., ignoring the very small difference between Hedges' g and Cohen's d at this sample size), an α-level of 0.05, a two-sided test, and two equally sized groups; 152 participants would be needed to obtain the minimum recommended power (80%), or 248 participants for high power (95%). Comparing these sample sizes to the sample sizes observed in published IRAP research employing between-groups or mixed within-between designs, 2% of such published IRAP studies employed a sample size sufficient to detect an effect size the same size as the average transdiagnostic efficacy of Acceptance and Commitment Therapy (Hedges' g = 0.46: Gloster et al., 2020) with minimum recommended power (80%), and 0% could detect this with high power (95%).

Please note again that the above power analyses are not specific recommendations for future IRAP research, and this article should not be cited as a source of such recommendations, but rather are worked examples of (a) how authors should begin to engage with conducting their own a priori power analyses, and (b) illustrative of how underpowered the published IRAP literature is to detect effect sizes of these magnitudes. I reiterate this caution due to The Law of Lakens' Guidelines, which states that whenever you try to make the point that researchers should not follow certain guidelines, you will nonetheless sometimes be cited as a source of said guidelines (Rohrer, 2023)."

[reviewer 1 comment 2]

2. This is just a preference issue: I think it makes sense to "connect the dots" in a time series graph it no other interpolation is going to be used. But, if one is fitting linear regressions to the data points, I'd rather just see the points and the regression line; the "connect the dots" line isn't really needed in such a situation.

Author response: Thank you for this suggestion. I retraced my thinking on this and this choice to "connect the dots" came down to the issue of accessibility and interpretability. To increase accessibility, I have chosen a color palette that is color-blind friendly and still interpretable when printed in black and white. However, interpreting whether each dot belongs to one group or the other still requires the reader to discriminate the color of individual dots, which can be difficult for some. This is particularly the case in Figure 2 where some of the green dots ('all studies') overlap with the blue dots in the other group ('studies with between-subject comparisons'). I've tried combinations of removing the joining lines and elaborating the legend, but this always produces a plot that is ambiguous under at least some conditions or for some viewers. Given that the addition of the connecting lines is at worst redundant, I've therefore elected to retain them for accessibility and interpretability of the group membership of each data point.

REVIEWER 2 COMMENTS:
This is a very well written and important paper about a central measure used in the CBS literature. It uses well articulated strategy to demonstrate important cultural practices inside the CBS community that are likely very harmful to the science being conducted. These results show that it is likely that many IRAP findings are not replicable and that effect sizes are likely overestimated in the existing literature due to file drawer effects. The paper uses a straightforward approach to assessing the literature, that while imperfect, is strong and has valid conclusions. The data is available for other researchers to verify whether conclusions are warranted. In addition, the author draws well formulated and sober conclusions from the findings and does not exaggerate nor draw conclusions that are overly broad. In all, this is an excellent paper that is a service to the field and much needed to hopefully start to correct these systemic problems.
Below I note a number of minor issues with the paper in order to further strengthen it.

[reviewer 2 comment 1]

--They need to better describe what a multiway ANOVA is. I'm not completely confident I know what the author is saying in using that term and so I suspect other readers may not be either. It would also be useful to demonstrate/explain why it inflates FP rates, which is not obvious from the current description.

[AE NOTE – I think you might be referring to 2-way ANOVAs, 3-way ANOVAs, etc. as a group. Please clarify this and add narrative as requested.]

Author response: Thank you – I agree this point needed fleshing out. Page 6 now reads:

“One specific class of statistical methods, multiway Analyses of Variance (ANOVAs, i.e., those with more than one independent variable such as 2-way ANOVAs, 3-way ANOVAs), are almost ubiquitous in IRAP research. Due to familywise error rates, the use of multiway ANOVA in exploratory or inductive research inflates false positive rates much higher than the 5% rate implied by the standard alpha level of 0.05 (Cramer et al., 2016). In the case of a simple 2X2 between groups ANOVA, this can be illustrated with simple math: if a researcher is willing to accept the result of any of the three p values generated by the ANOVA (i.e., either main effect or the interaction effect) as evidence of an effect, as would be common when applying the ANOVA in a an exploratory or inductive manner, then the false positive rate for the ANOVA as a whole is not equal to the alpha value (e.g., 5%), but a higher value. Specifically, False Positive Rate = $1-(1-alpha)^k$*, where k is the number of p values. Using alpha = 0.05 and k = 3 (i.e., two main effects and one interaction effect), False Positive Rate = 14.3%. Cramer et al. (2016) note that the false positive rate implied by larger ANOVA designs, such as those often employed in IRAP research (e.g., 4X2X2 mixed within-between ANOVAs), are higher again, but would require specific simulation studies to estimate.”*

[reviewer 2 comment 2]

--Typo here: "inductive manner (Lakens, 2021) or in an inductive manner"?
[AE NOTE – from the rest of the paragraph, it would seem that the phrase prior to the Lakens citation was intended to be “deductive manner” – please verify and correct this]

Author response: Thank you for catching this – you are correct. Page 7 now reads:

“Testing in a deductive manner (Lakens, 2021) or in an inductive manner (e.g., to generate new hypotheses rather than test existing ones).”

[reviewer 2 comment 3]

--The writer should describe the rationale for "Variant procedures such as the Mixed-Trials IRAP (MTIRAP: Levin et al., 2010) and the Training IRAP (T-IRAP: Kilroe et al., 2014) were excluded." This is important for readers less familiar with the IRAP literature.

Author response: Thank you for this suggestion. I originally didn’t get into detail here for the sake of brevity. I have explicated this logic now. Pages 9-10 now read:

“Variant procedures such as the Mixed-Trials IRAP (MT-IRAP: Levin et al., 2010) and the Training IRAP (T-IRAP: Kilroe et al., 2014) were excluded on the basis that, although these tasks share similar names with the IRAP, they specifics diverge so substantially from the IRAP as to represent a strong risk of a jingle fallacy: the mistaken assumption that two measures sharing the same name measure the same thing (e.g., Lilienfeld & Strother, 2020). For example, the IRAP requires participants to provide responses that are both notionally consistent and inconsistent with their pre-experimentally established learning history (e.g., to respond to “White people” and “positive” with “true” on some blocks and “false” on others). In contrast, the Training IRAP requires responding consistent with only one of these patterns in order to establish that pattern of responding rather than assess it. Despite its name, the Training IRAP is therefore more closely related to the Relational Evaluation Procedure (e.g., J. Hayes et al., 2016) than the IRAP.
The risk of a jingle fallacy also applies to treating the IRAP and MT-IRAP as if they are meaningfully similar. Whereas the IRAP alternates between response patterns between blocks, the MT-IRAP does it between trials through the inclusion of an additional stimulus that indicates whether participants should tell the "truth" (provide a history-consistent response) or "lie" (provide a history-inconsistent response) on that trial. There are thus parallels between the MT-IRAP and the Recoding Free version of the Implicit Association Test (IAT-RF: Rothermund et al., 2009) as variants of their respective original tasks. To the best of my knowledge, no work to date has assessed the correlation between IRAPs and MT-IRAPs designed to assess the same domain. More broadly, it is important to note that although the IRAP and several other tasks including the Implicit Association Test are collectively labeled "implicit measures", scores on these tasks are typically found to correlate poorly with one another, even when the tasks share some procedural features and are intended to measure the same domain (e.g., Clayton et al., 2023; Schimmack, 2021; for a detailed conceptual critique see Corneille & Hütter, 2020). As such, in the absence of evidence for convergent validity between the IRAP and MT-IRAP, the MT-IRAP was excluded out of an abundance of caution against introducing jingle fallacy into the analysis.”

[reviewer 2 comment 4]

--This statement seems incorrect or at least I am interepreting it to be incorrect, "median sample sizes in IRAP studies are small (range 12 to 64)" as the Figure 1 shows samples in the 200s.

Author response: Thank you for catching this lack of clarity. Figure 1 does indeed illustrate the actual sample sizes observed, but the paragraph you quote is from the section on “Change in sample size per study over time”. The quoted text has been adjusted to correspond to the paragraph’s point more closely:

“As can be seen in Figure 2, across years, median sample sizes in IRAP studies have been small (range of medians 12 to 64).”

[reviewer 2 comment 5]

-- I don't think this statement is correct, "Results demonstrated that the implied statistical power to detect the average published effect size (Cohen's d = 0.408, equivalent to Pearson's r = 0.20) was increasing from an estimated .142, 95% CI [.108, .177] in 2006 (the model intercept) by an average of .009, 95% CI [.005, .012], p < .001 participants per year." I believe the "participants per year" should be deleted?

Author response: Thank you for catching this error. “participants per year” was deleted from this quote on page 18.

[reviewer 2 comment 6]

--There are numerous typos on the manuscript that should be corrected by careful proofreading. I'd recommend the author have someone else read the manuscript with the eye of catching potential typos lest this weakness take away from the perceived intellectual contribution of the manuscript.

Author response: Thank you – I have corrected several dozen typos and suboptimal word choices throughout the manuscript and had it thoroughly proofread.

[reviewer 2 comment 7]

-- One additional weakness that should be noted is that it appears that only the author participated in the coding, lending the possibility of systematic bias or inaccurate coding. I realize other people can check all the coding, but this is an arduous process that is not likely to occur, so this remains a weakness that should be noted.

Author response: Good point. I would have preferred a second coder, but this project evolved quickly and I didn’t find an interested collaborator to do this second coding. I have added this limitation to the limitations section. Page 28 now reads:

“Extraction of sample sizes
Sample sizes were extracted from the original articles by a single coder and no estimates of inter-rater reliability were therefore produced. All data and code for the current manuscript are public, therefore making the accuracy of these extractions testable in principle. However, given this is a non-trivial task, and inaccurate data extractions could lead to bias, the lack of a second scorer must be acknowledged as a limitation.”

- Low sample size and statistical power are key contributors to poor replicability
- Median sample sizes across studies can be used to estimate power in that literature
- Systematic review was used to find all research using the IRAP
- Sample sizes were extracted from each study and implied power was calculated
- Statistical power in IRAP studies is consistently and problematically low

A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using the Implicit Relational Assessment Procedure

Ian Hussey

Author note: Ian Hussey, Ruhr University Bochum, Germany. ian.hussey@rub.de. I would like to thank Jamie Cummins and Chad E. Drake for their feedback on earlier versions of this manuscript. This research was supported by the META-REP Priority Program of the German Research Foundation (#464488178). All code and data available at osf.io/vpwuy.

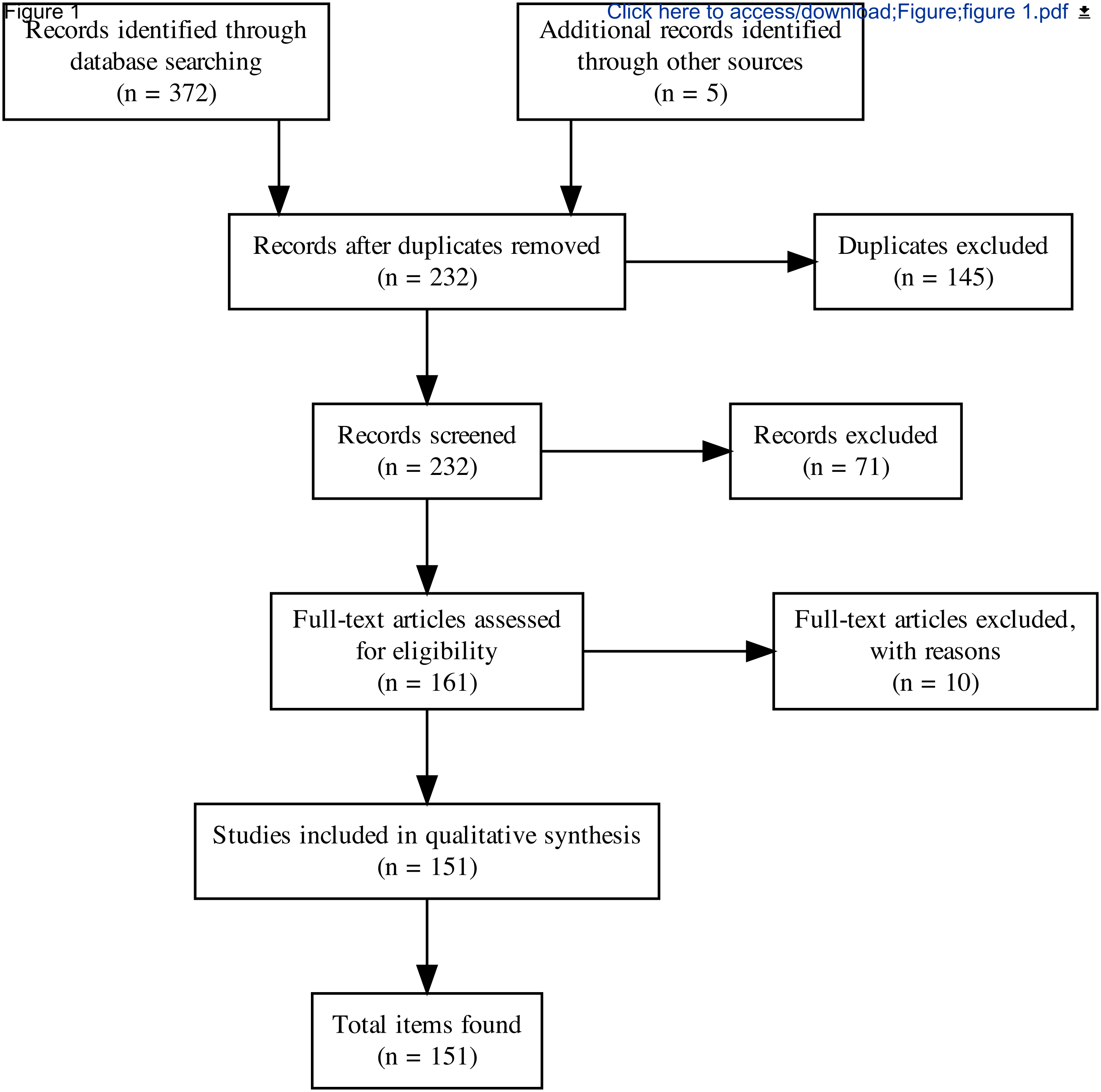

Records identified through database searching (n = 372)
Additional records identified through other sources (n = 5)
Records after duplicates removed (n = 232)
Duplicates excluded (n = 145)
Records screened (n = 232)
Records excluded (n = 71)
Full-text articles assessed for eligibility (n = 161)
Full-text articles excluded, with reasons (n = 10)
Studies included in qualitative synthesis (n = 151)
Total items found (n = 151)

Figure 2

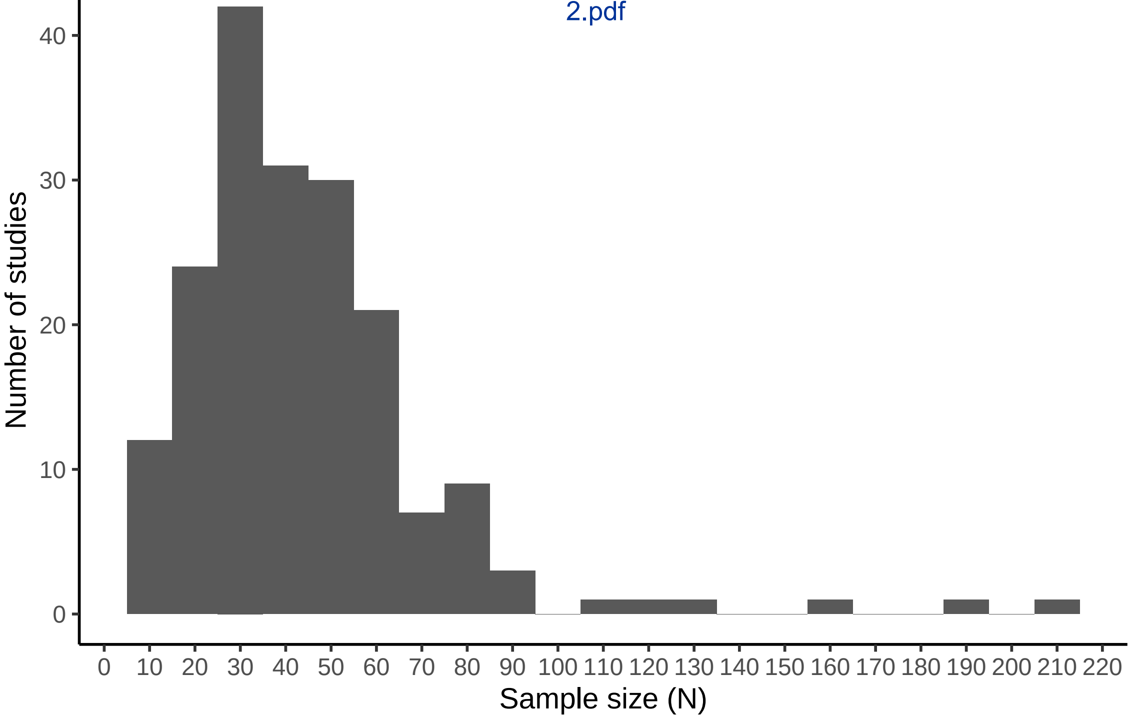

Number of studies
0
10
20
30
40
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
Sample size (N)

Figure 3

Click here to access/download;Figure;figure 3.pdf

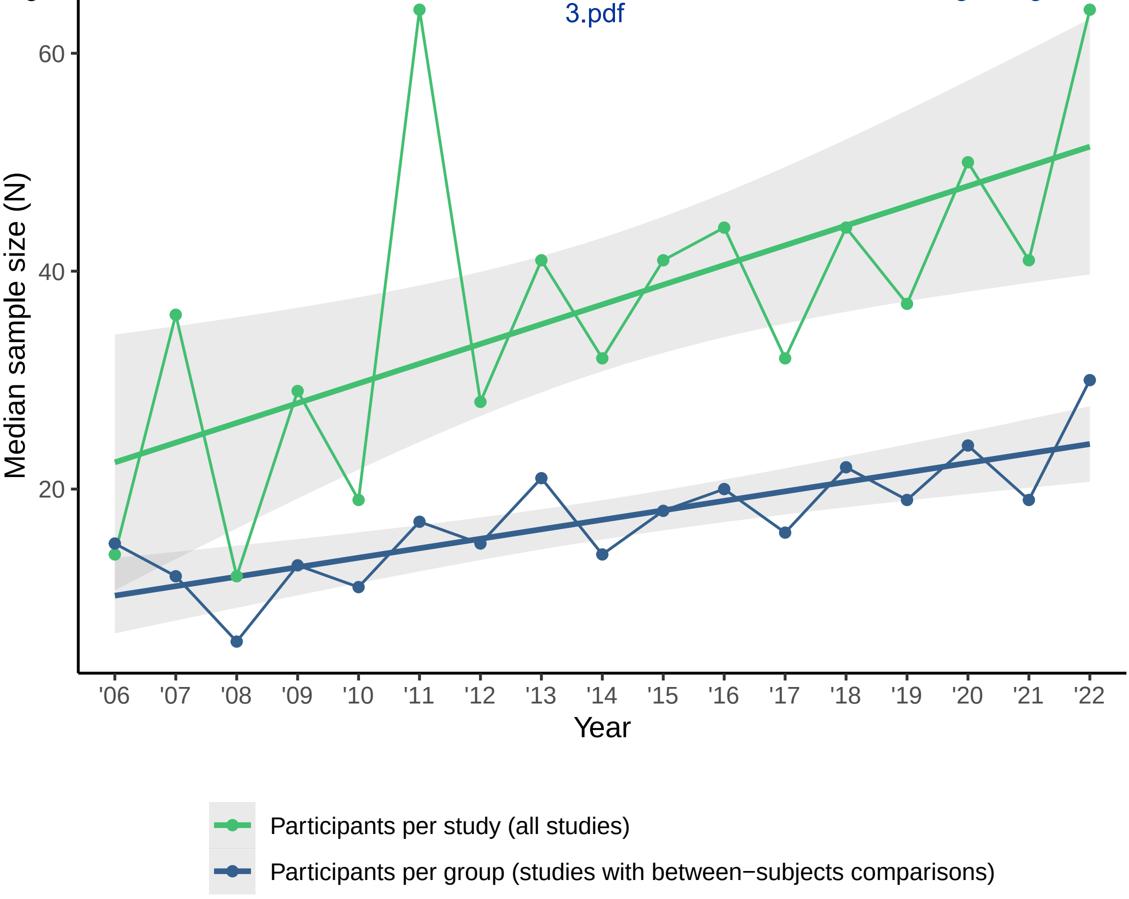


Figure 4

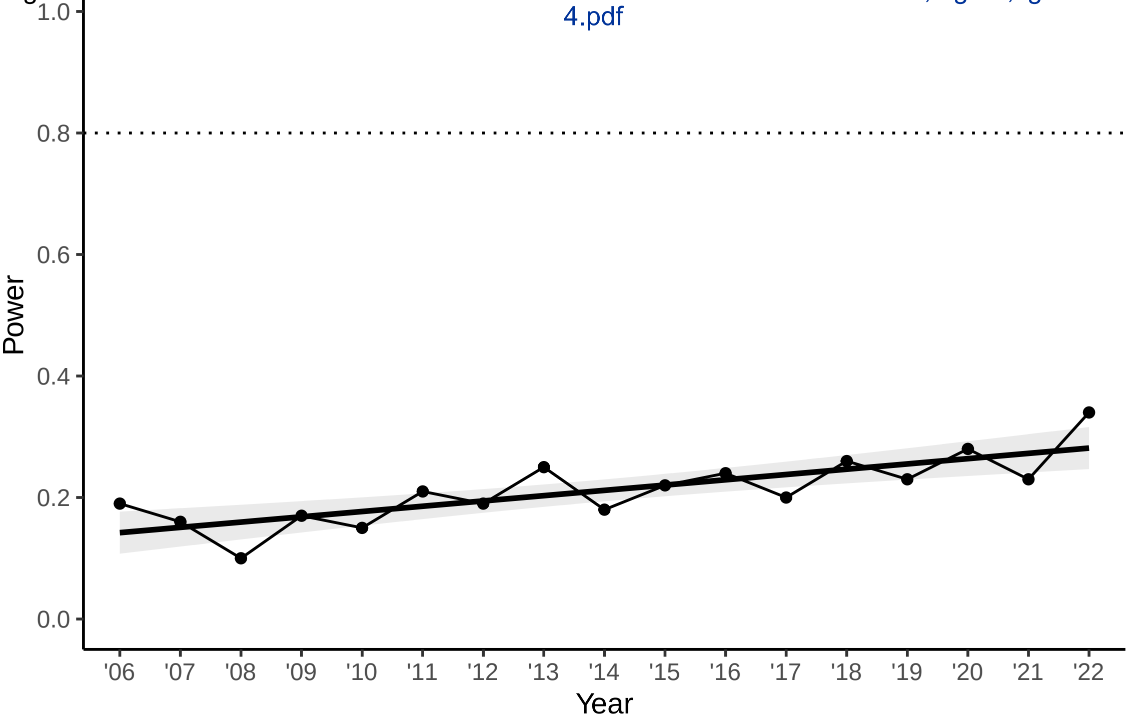

1.0
0.8
0.6
0.4
0.2
0.0
Power
'06
'07
'08
'09
'10
'11
'12
'13
'14
'15
'16
'17
'18
'19
'20
'21
'22
Year

Figure 5

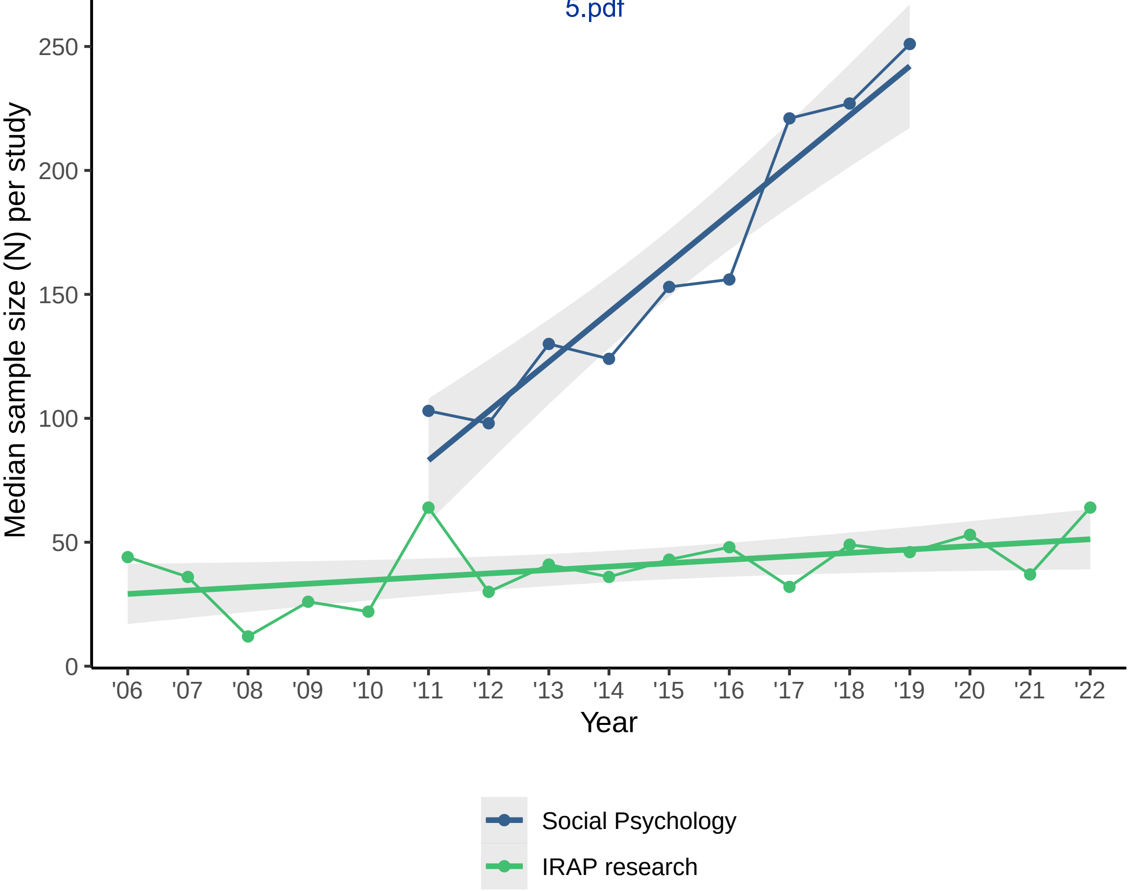
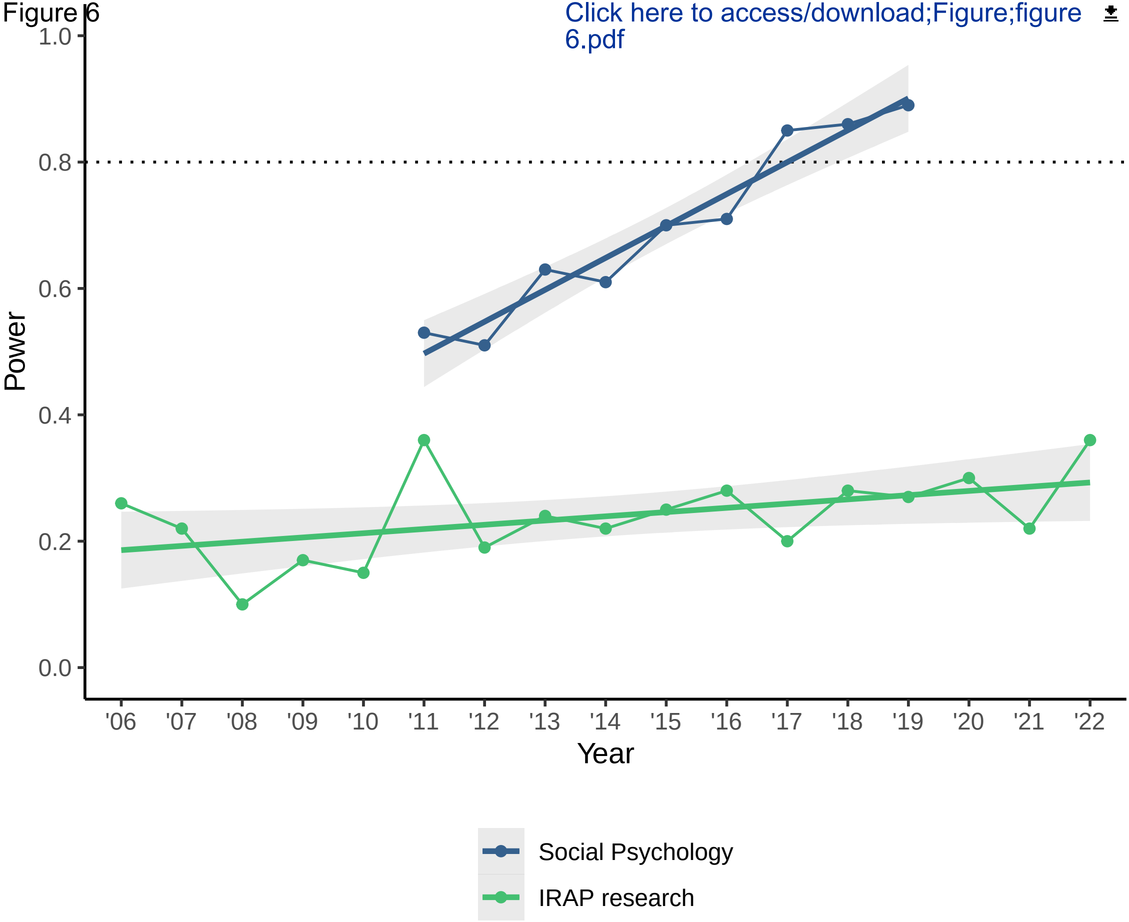

Median sample size (N) per study
250
200
150
100
50
0
'06
'07
'08
'09
'10
'11
'12
'13
'14
'15
'16
'17
'18
'19
'20
'21
'22
Year
Social Psychology
IRAP research

Figure 6


1.0
0.8
0.6
0.4
0.2
0.0
Power
'06
'07
'08
'09
'10
'11
'12
'13
'14
'15
'16
'17
'18
'19
'20
'21
'22
Year
Social Psychology
IRAP research

A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using the Implicit Relational Assessment Procedure

Supplementary Materials

Systematic review full text exclusions with reasons

Publications by Baker et al. (2015), Baker et al. (2017), Smith et al. (2022) and Szarko et al. (2022) each reported employing an IRAP but did not report the procedural features in sufficient detail to determine whether it was a standard IRAP or a variant. Personal correspondence with the authors revealed these studies employed a non-standard IRAP variant. Following the inclusion criteria, these studies were therefore excluded.

Inspection of the full text of Perez et al. (2020) demonstrated that those studies did not employ an IRAP (or IRAP variant).

Inspection of the full texts of four articles by Harte and colleagues (Harte, Barnes-Holmes, et al., 2021; Harte, Barnes-Holmes, et al., 2021; Harte et al., 2018, 2020)(2018, 2020, 2021a, 2021b) demonstrated that those studies employed a Training IRAP (T-IRAP) rather than a standard IRAP. Following the inclusion criteria, these studies were therefore excluded.

Tables

Table S1. Number of sample sizes in the Social and Personality psychology dataset by journal

Journal	*N* sample sizes
European Journal of Personality	113
European Journal of Social Psychology	306
Journal of European Social Psychology	589
Journal of Personality	189
Journal of Personality and Social Psychology	631
Journal Research in Personality	157
Psychological Science (articles coded as relevant to Social Psychology)	269
Personality and Social and Psychology Bulletin	527
Social Psychological and Personality Science	266

Table S2. Median sample size per year in IRAP studies

	Year																
Calculated from	’06	’07	’08	’09	’10	’11	’12	’13	’14	’15	’16	’17	’18	’19	’20	’21	’22
Participants per study	14 (4)	36 (1)	12 (1)	29 (10)	19 (13)	64 (5)	28 (6)	41 (7)	32 (9)	41 (16)	44 (28)	32 (17)	44 (15)	37 (19)	50 (16)	41 (11)	64 (7)
Participants per group in studies with between-subjects comparisons	15 (1)	12 (1)	6 (1)	13 (9)	11 (6)	17 (3)	15 (3)	21 (5)	14 (8)	18 (12)	20 (22)	16 (11)	22 (9)	19 (10)	24 (12)	19 (7)	30 (5)

This table corresponds with Figure 2. The number of studies each median is calculated from is listed in brackets.

Table S3. The statistical power to detect the average published effect size (Cohen’s $d = 0.408$ or Pearson’s $r = 0.20$) implied by median sample size per group per year in IRAP studies

Year																
’06	’07	’08	’09	’10	’11	’12	’13	’14	’15	’16	’17	’18	’19	’20	’21	’22
19	16	10	17	15	21	19	25	18	22	24	20	26	23	28	23	34

This table corresponds with Figure 3.

Table S4. Median sample size per year in IRAP studies compared to Social and Personality Psychology studies

	Year																
	’06	’07	’08	’09	’10	’11	’12	’13	’14	’15	’16	’17	’18	’19	’20	’21	’22
Social and Personality research	-	-	-	-	-	103 (314)	98 (304)	130 (316)	124 (297)	153 (390)	156 (327)	221 (344)	227 (402)	251 (353)	-	-	-
IRAP research	44 (1)	36 (1)	12 (1)	26 (9)	22 (6)	64 (3)	30 (3)	41 (5)	36 (8)	43 (12)	48 (22)	32 (11)	49 (9)	46 (10)	53 (12)	37 (7)	64 (5)

This table corresponds with Figure 4. The number of studies each median is calculated from is listed in brackets.

Table S5. The statistical power to detect the average published effect size (Cohen’s $d = 0.408$ or Pearson’s $r = 0.20$) implied by median sample size per study per year in IRAP studies compared to Social and Personality Psychology studies

	Year																
	’06	’07	’08	’09	’10	’11	’12	’13	’14	’15	’16	’17	’18	’19	’20	’21	’22
Social and Personality research	-	-	-	-	-	53	51	63	61	70	71	85	86	89	-	-	-
IRAP research	26	22	10	17	15	36	19	24	22	25	28	20	28	27	30	22	36

This table corresponds with Figure 5.

References

Baker, T. K., Schwenk, T., Piasecki, M., Smith, G. S., Reimer, D., Jacobs, N., Shonkwiler, G., Hagen, J., & Houmanfar, R. A. (2015). Cultural Change in a Medical School: A Data-Driven Management of Entropy. *Journal of Organizational Behavior Management*, *35*(1–2), 95–122. https://doi.org/10.1080/01608061.2015.1035826

Baker, T. K., Smith, G. S., Jacobs, N. N., Houmanfar, R., Tolles, R., Kuhls, D., & Piasecki, M. (2017). A deeper look at implicit weight bias in medical students. *Advances in Health Sciences Education*, *22*(4), 889–900. https://doi.org/10.1007/s10459-016-9718-1

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2018). The impact of high versus low levels of derivation for mutually and combinatorially entailed relations on persistent rule-following. *Behavioural Processes*, *157*, 36–46. APA PsycInfo. https://doi.org/10.1016/j.beproc.2018.08.005

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2021). Exploring the impact of coherence (through the presence versus absence of feedback) and levels of derivation on persistent rule-following. *Learning & Behavior*, *49*(2), 222–239. APA PsycInfo. https://doi.org/10.3758/s13420-020-00438-1

Harte, C., Barnes-Holmes, D., Barnes-Holmes, Y., McEnteggart, C., Gys, J., & Hasler, C. (2020). Exploring the potential impact of relational coherence on persistent rule-following: The first study. *Learning & Behavior*, *48*(3), 373–391. APA PsycInfo. https://doi.org/10.3758/s13420-019-00399-0

Harte, C., Barnes-Holmes, D., Moreira, M., de Almeida, J. H., Passarelli, D., & de Rose, J. C. (2021). Exploring a training IRAP as a single participant context for analyzing reversed derived relations and persistent rule-following. *Journal of the Experimental Analysis of Behavior*, *115*(2), 460–480. APA PsycInfo. https://doi.org/10.1002/jeab.671

Perez, W. F., de Almeida, J. H., Soares, L. C. C. S., Wang, T. F. L., de Morais, T. E. D. G., Mascarenhas, A. V., & de Rose, J. C. (2020). Fearful Faces and the Derived Transfer of Aversive Functions. *The Psychological Record*. https://doi.org/10.1007/s40732-020-00390-6

Smith, G. S., Houmanfar, R. A., Jacobs, N. N., Froehlich, M., Szarko, A. J., Smith, B. M., Kemmelmeier, M., Baker, T. K., Piasecki, M., & Schwenk, T. L. (2022). Assessment of medical student burnout: Toward an implicit measure to address current issues. *Advances in Health Sciences Education*, *27*(2), 375–386. https://doi.org/10.1007/s10459-021-10089-0

Szarko, A. J., Houmanfar, R. A., Smith, G. S., Jacobs, N. N., Smith, B. M., Assemi, K., Piasecki, M., & Baker, T. K. (2022). Impact of Acceptance and Commitment Training on psychological flexibility and burnout in medical education. *Journal of Contextual Behavioral Science*, *23*, 190–199. APA PsycInfo. https://doi.org/10.1016/j.jcbs.2022.02.004

Declaration of conflicts of interest: none

Abstract

Following recent calls to examine the replicability of behavioral research, I examine two key determinants of replicability, sample sizes, and statistical power, in research using the Implicit Relational Assessment Procedure (IRAP). A systematic review was used to gather all published studies employing the IRAP and extract their designs and sample sizes. The use of Null Hypothesis Significance Testing was found to be nearly ubiquitous, justifying the examination of statistical power. Using an established method, median sample sizes were used to estimate the statistical power to detect the average published effect size in psychological research (r = .20) in each year. Sample sizes and the statistical power they imply were found to be very low in IRAP studies (in 2022, median N = 64, power = 34%). At the current rate of growth, the median study will only reach the recommended minimum power of at least 80% by 2080. The IRAP literature was directly compared with the Social and Personality psychology literature using an existing dataset. Median sample sizes and their implied statistical power were lower in the IRAP literature in all years than they were in Social and Personality psychology at the beginning of the Replication Crisis in 2011, and indeed in all subsequent years. Improvements in sample sizes and statistical power in the Social and Personality psychology literature were significantly and substantially larger than in the IRAP literature. Direct tests of the reproducibility and replicability of claims in the IRAP literature are needed.

Two seminal articles were published in 2011 whose implications the field of psychology is still grappling with. The first, by Daryl Bem (2011), contained literally impossible results about a supposed human ability to predict the future. It was remarkable not merely in its claims but because it employed modal research practices for the field to substantiate these conclusions. A second article by Simmons and colleagues (2011), coincidentally published around the same time as Bem (2011) but without any knowledge of his paper, demonstrated how our modal research practices can easily and routinely generate statistically significant results from what is actually just noise.

The fallout from this pair of papers and many more before and since is now a matter of history for some (e.g., for personal accounts see Gelman, 2016; Spellman, 2015). But their impact has been uneven and recognition of the Replication Crisis as a serious issue to be reckoned with has been heterogenous, both within and between fields. It would be too easy to dismiss the Replication Crisis as specific to Social and Personality psychology when other fields employ similar research practices. Over time, recognition of these issues has spread to other areas of psychology (e.g., clinical psychology: Tackett et al., 2019) as well as a diverse range of other fields including cancer biology, economics, methodology, sociology, and philosophy (Baker & Dolgin, 2017; Boulesteix et al., 2020; Buckwalter, 2022; Gordon et al., 2020; Page et al., 2021).

These calls to take seriously the question of replicability have recently been echoed in the behavioral research communities. A recent editorial for Perspectives on Behavior Science characterized the situation well:

“Despite certain metatheoretical disputes (Burgos & Killeen, 2019), behavior science, behavior analysis, and psychology have much more in common than differences. Hence the ‘replication crisis’ in psychology could well be repeated in behavior science and behavior analysis. Even if it is not, it may hold some important lessons for both scientists and practitioners." (Hantula, 2019, pp. 4-5)

Similarly, the Association for Contextual Behavioral Science's Taskforce on the Strategies and Tactics of Contextual Behavioral Science Research (2021) recently announced its explicit support for Open Science principles, including data transparency and a focus on replication. As such, there is now support for the idea that behavioral science, including component fields such as Behavior Analysis and Contextual Behavioral Science, would be enhanced by examining and enhancing the transparency and replicability of its findings.

In some ways, this could be seen as an appropriate return to our roots, from which we should never have departed. To take some positive examples of our field's history here, whereas general psychology journals have until recently rarely published replication studies (Makel et al., 2012), behavioral journals such as the Journal of Experimental Analysis of Behavior and Journal of Applied Behavior Analysis have a long history of publishing them. Behavioral research also has a long history of sharing research data (i.e., trial-level single-case experimental design data being presented in tables and plots). Lastly, despite being written over 60 years ago, fully one-quarter of Murray Sidman's seminal behavioral textbook *Tactics of Scientific Research* (1960) is dedicated to discussion of the need for replication studies, their taxonomy and function, and links between replication and generalization.

Replicability, sample size, and statistical power

Statistical power is the probability of detecting a true effect, and is synonymous with the sensitivity of a test and its False Negative Rate (Cohen, 1977). Low statistical power in original studies is a key contributor to the Replication Crisis in psychology (e.g., Asendorpf et al., 2013; Button et al., 2013; Munafò et al., 2017), with highly powered replications only obtaining the original finding in around one third to one-half of studies depending on the definition of successful replication, and effect sizes observed in replication studies are typically only one-third the size of those in original studies (e.g., Ebersole et al., 2020; Klein et al., 2018; Open Science Collaboration, 2015). Journals that publish underpowered studies are likely to publish a greater proportion of conclusions that are false positives (Bakker et al., 2012; Ioannidis, 2005). As such, in reaction to the Replication Crisis in psychology, many have called for psychology research to employ more highly-powered tests and therefore larger sample sizes (e.g., Asendorpf et al., 2013; Button et al., 2013; Munafò et al., 2017; Wagenmakers et al., 2012).

Along with the False Positive Rate (i.e., α-level, typically < .05), statistical power is one of two key properties of inference via NHST that defines the long-run error rates of the inferences we make from data. Power is generally a less familiar concept than α-level for many researchers, but it is so central to our ability to make inferences from data (Cohen, 1992). Nonetheless, for decades, statistical power remained very low in the behavioral sciences (i.e., around .46: Cohen, 1990). Additionally, research has shown that researchers' intuitions about the statistical power implied by rules-of-thumb sample sizes are inaccurate and overestimate power (Bakker et al., 2016).

To assess the efficacy of this more recent call for higher power motivated by the Replication Crisis, on the sample sizes employed in published research, Fraley and colleagues (Fraley et al., 2022; Fraley & Vazire, 2014) quantified the median sample size employed in articles published in nine personality and social psychology journals between 2011 (arguably the start of the Replication Crisis) and 2019. Fraley and colleagues (2022) observed that median sample sizes, and therefore implied statistical power, have indeed increased over the last decade in social and personality psychology research, from very poor (circa .50 in 2011) to acceptable (circa .90 in 2019). In doing so, Fraley and colleagues (Fraley et al., 2022; Fraley & Vazire, 2014) provided both a relatively simple method to assess the implied power across a body of work and a useful dataset to compare other fields against.

The current study employs this method to provide one of the first examinations of a key determinant of the replicability of research in an area of work relevant to both behaviorism and Contextual Behavioural Science. The stated goal of CBS is “creating a science more adequate to the challenge of the human condition” (S. C. Hayes et al., 2012). Efforts to meet this noble goal would be aided by scientific findings that are reliable, reproducible, and replicable (Munafò et al., 2017). This fact was recently recognized by leadership within the CBS community (Task Force on the Strategies and Tactics of Contextual Behavioral Science Research, 2021), which served as motivation for the current work.

The Implicit Relational Assessment Procedure (IRAP)

The Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2006) is a reaction-time-based task used variously as a measure of implicit attitudes in implicit social cognition research and a measure of the strength of relational responding within Contextual Behavioral Science research (Hughes et al., 2012). One meta-analysis suggested that the IRAP demonstrates relatively high criterion validity compared to other implicit measures such as the Implicit Association Test (Vahey et al., 2015). However, multiple other meta-analyses have also suggested that the IRAP has poor internal consistency (estimates of α from .51 to .60) and unacceptably low test-retest reliability (estimates of *r* from .13 to .43: Greenwald & Lai, 2020; Hussey & Drake, 2020). This presents somewhat of a conundrum, as the reliability places an upper limit on validity (i.e., through attenuation of observable correlations: Revelle, 2009).

One explanation for these seemingly irreconcilable results is that the IRAP literature may suffer from poor replicability, such as inflated effect sizes and false positive rates. This is not implausible. Although the IRAP grew out of the behaviorist tradition (Barnes-Holmes et al., 2010), IRAP studies typically employ the same research designs and inference methods as Social and Personality psychology (e.g., Null Hypothesis Significance Testing, NHST), and are therefore subject to the same concerns as any research employing this inference approach.

One specific class of statistical methods, multiway Analyses of Variance (ANOVAs, i.e., those with more than one independent variable such as 2-way ANOVAs and 3-way ANOVAs), are almost ubiquitous in IRAP research. Due to familywise error rates, the use of multiway ANOVA in exploratory or inductive research inflates false positive rates much higher than the 5% rate implied by the standard alpha level of 0.05 (Cramer et al., 2016). In the case of a simple 2 X 2 between groups ANOVA, this can be illustrated with simple math: if a researcher is willing to accept the result of any of the three p values generated by the ANOVA (i.e., either of the main effects or the interaction effect) as evidence of an effect, as would be common when applying the ANOVA in an exploratory or inductive manner, then the false positive rate for the ANOVA as a whole is not equal to the alpha value (e.g., 5%), but a higher value. Specifically, False Positive Rate = $1 - (1 - alpha)^k$, where k is the number of p values (Cramer et al., 2016). Using alpha = 0.05 and k = 3 (i.e., two main effects and one interaction effect), the False Positive Rate = 14.3%. Cramer et al. (2016) note that the false positive rate implied by larger ANOVA designs, such as those often employed in IRAP research (e.g., 4 X 2 X 2 mixed within-between ANOVAs), are higher again, but would require specific simulation studies to estimate. The unavoidable implication is that if the IRAP literature employs a statistical method, which is known to have both inflated false positive rates under modal use and inflated false negative rates under low statistical power, then the published IRAP literature will have inflated rates of false conclusions (i.e., low replicability). At a minimum, there is no sound statistical basis by which the IRAP literature could be judged to be a priori immune from such concerns. Rather, the replicability of conclusions in the published IRAP literature must be assessed empirically, via both direct replication studies and assessment of indicators of replicability, such as sample size and statistical power.

It is also worth noting that, given that it is the probability of detecting effects that exist, high statistical power is a desirable property regardless of whether a researcher is employing

Null Hypothesis Significance Testing in a deductive manner (Lakens, 2021) or in an inductive manner (e.g., to generate new hypotheses rather than test existing ones). Some IRAP researchers have stated they do the latter (Kavanagh, Matthyssen, et al., 2019), although as an aside it should be recognized that this risks representing a form of Hypothesizing After Results Are Known (HARKing: Kerr, 1998), which lowers the replicability of findings (Munafò et al., 2017). Regardless of a researcher's self-identified approach as deductive versus inductive, it should be recognized that a smaller number of high-powered studies generally generates a larger number of true conclusions than a larger number of low-powered studies (LeBel et al., 2017).

The current study therefore represents a first effort toward quantifying two related indicators of replicability in the IRAP literature. I performed a systematic review of published research using the IRAP and then applied Fraley et al.'s (2022; see also Fraley & Vazire, 2014) approach to estimating median sample size over time and the implied statistical power in this literature. As noted in those publications, median sample sizes are more appropriate than other metrics of central tendency (e.g., the mean) due to the strong skew in sample sizes. The median is also highly interpretable as it tells you that half of the studies had larger sample sizes than this value and half of the studies had smaller sample sizes than this.

Sample sizes in the IRAP literature were then contrasted with sample sizes employed elsewhere. Notionally, a comparison with closely related literatures might seem appropriate, such as studies employing other implicit measures such as the Implicit Association Test (Greenwald et al., 1998), Affect Misattribution Procedure (Payne et al., 2005), or Evaluative Priming Task (Fazio et al., 1995). However, this comparison is quite extreme: thanks in part to the popularity of the Project Implicit website (implicit.harvard.edu), studies employing other implicit measures often contain thousands of participants (e.g., Hughes et al., 2022), frequently contain tens of thousands (Bar-Anan & Nosek, 2014; Nosek et al., 2007), sometimes contain hundreds of thousands (Hussey et al., 2019), and occasionally even millions of participants (Xu et al., 2014). Of course, the sample sizes employed in the field of implicit social cognition are apparently large not only in comparison to the IRAP but also in comparison to other areas of Social and Personality psychology. As such, it is perhaps more informative to compare sample sizes in the IRAP literature with a more diverse sample as a reference, such as the Social and Personality psychology literature as a whole. I, therefore, made use of the openly available dataset created by Fraley et al. (2022), which covers the Social and Personality psychology literature.

Method

Data was obtained from two separate sources. Research designs and sample sizes within the published IRAP literature were obtained via a systematic review. To provide a comparison for this literature, existing data on the research designs and sample sizes reported in articles published in nine Social and Personality Psychology journals was taken from a recent openly-available dataset (Fraley et al., 2022). The data extraction method for the IRAP literature is based on the example provided by Fraley et al. (2022).

A systematic review of research designs in the IRAP research (2006-2022)

Results from both searches were integrated. Results of each stage of this review are computationally reproducible: BibTeX files for all articles at each stage of the search and exclusion process are available in the supplementary materials (blinded URL for peer review: https://osf.io/vpwuy/?view_only=21905097cd054e9497d1c5574796e86b) and can be updated by others or used for other evidence synthesis or meta-science purposes.

Both the Web of Science and PsycINFO databases were searched. Boolean search terms for the Web of Science database were “implicit relational assessment procedure” OR “IRAP” in the title, abstract, or keywords. Search constraints were publication date between 2006 and

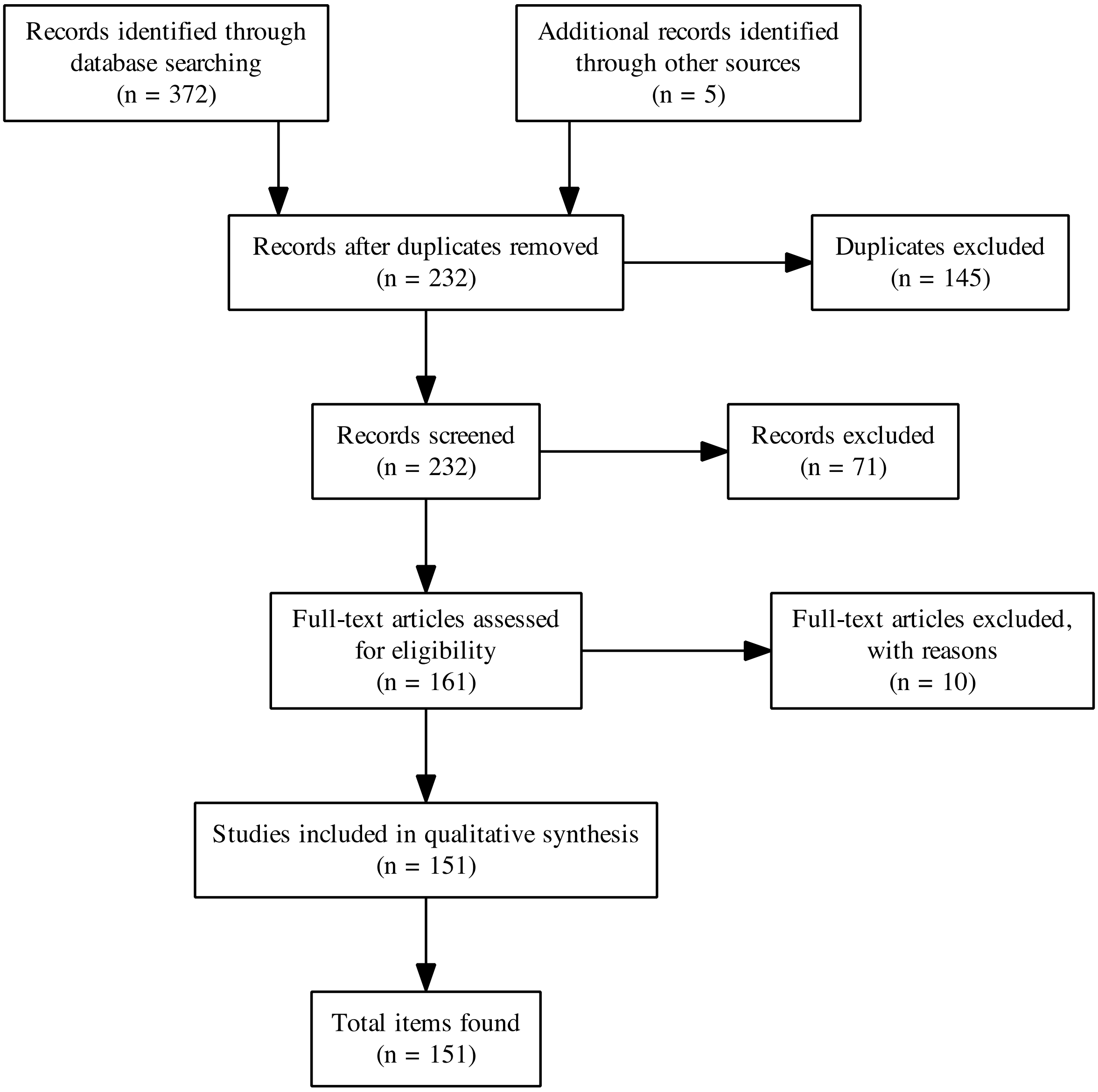
2022, limited to publications in English. The search was run on 23 December 2018. The systematic review was updated with a second search run on 11 September 2022.

A PRISMA flow chart detailing all exclusions can be found in Figure 1 (Moher et al., 2009). 372 records were obtained from the database searches and 5 from other sources. After duplicates were removed, 232 records remained. The retained records were screened based on their title and abstract. The inclusion criterion was the use of the Implicit Relational Assessment Procedure (IRAP) within the study. Variant procedures such as the Mixed-Trials IRAP (MT-IRAP: Levin et al., 2010) and the Training IRAP (T-IRAP: Kilroe et al., 2014) were excluded on the basis that, although these tasks share similar names with the IRAP, their specifics diverge so substantially from the IRAP as to represent a strong risk of a jingle fallacy: the mistaken assumption that two measures sharing the same name measure the same thing (e.g., Lilienfeld & Strother, 2020). For example, the IRAP requires participants to provide responses that are both notionally consistent and inconsistent with their pre-experimentally established learning history (e.g., to respond to “White people” and “positive” with “true” on some blocks and “false” on others). In contrast, the Training IRAP requires responding consistent with only one of these patterns in order to establish that pattern of responding rather than assess it. Despite its name, the Training IRAP is therefore more closely related to the Relational Evaluation Procedure (e.g., J. Hayes et al., 2016) than the IRAP.

The risk of a jingle fallacy also applies to treating the IRAP and MT-IRAP as being meaningfully similar. Whereas the IRAP alternates between response patterns between blocks, the MT-IRAP does it between trials through the inclusion of an additional stimulus that indicates whether participants should tell the “truth” (provide a history-consistent response) or “lie” (provide a history-inconsistent response) on that trial. There are thus parallels between the MT-IRAP and the Recoding Free version of the Implicit Association Test (IAT-RF: Rothermund et al., 2009) as variants of their respective original tasks. To the best of my knowledge, no work to date has assessed the correlation between IRAPs and MT-IRAPs designed to assess the same domain. More broadly, it is important to note that although the IRAP and several other tasks including the Implicit Association Test are collectively labeled "implicit measures", scores on these tasks are typically found to correlate poorly with one another, even when the tasks share some procedural features and are intended to measure the same domain (e.g., Clayton et al., 2023; Schimmack, 2021; for a detailed conceptual critique see Corneille & Hütter, 2020). As such, in the absence of evidence for convergent validity between the IRAP and MT-IRAP, the MT-IRAP was excluded out of an abundance of caution against introducing jingle fallacy into the analysis.

161 records remained after the title and abstract exclusions. The full texts of these articles were then screened using the same inclusion criterion. Ten articles were excluded based on this full-text search. In each case, this was because they did not employ an IRAP (or IRAP variant) at all, or because they employed an IRAP variant such as a Training IRAP or Mixed-Trials IRAP. A list of these exclusions and their individual reasons is available in the supplementary materials. After all exclusions, 151 published articles and book chapters using the IRAP were found that met the inclusion criteria.

Figure 1. PRISMA flow chart for systematic review


Records identified through database searching (n = 372)
Additional records identified through other sources (n = 5)
Records after duplicates removed (n = 232)
Duplicates excluded (n = 145)
Records screened (n = 232)
Records excluded (n = 71)
Full-text articles assessed for eligibility (n = 161)
Full-text articles excluded, with reasons (n = 10)
Studies included in qualitative synthesis (n = 151)
Total items found (n = 151)

The full text for each record was then inspected to extract the following information for each study described: the sample size after exclusions (*N*); study design (between, within, or mixed); the number of between group conditions; and whether the study reported employing Null Hypothesis Significance Testing (NHST). The sample size after exclusions was extracted rather than the sample prior to exclusions given the IRAP's established high attrition rate

(Hussey et al., 2015). Note that comparisons among multiple IRAP trial types were excluded from consideration when labeling a given study as including a within-subjects element, given that this feature is so common in the literature. Where a study employed multiple designs (e.g., both correlating the IRAP with a criterion variable and examining the pattern of IRAP effects between groups) it was labeled "mixed". As such, "mixed" refers not only to mixed within-between research designs but also to articles that reported both within and between designs. This was suitable for the current analytic purposes, which required excluding the purely within-subject studies from the analyses to estimate statistical power correctly (i.e., using those studies employing at least one between groups analysis).

Review of research practices in Social and Personality Psychology journals (2011-2019)

Fraley et al. (2022) recently reviewed the sample sizes employed in nine Social and Personality Psychology journals (European Journal of Social Psychology, European Journal of Personality, Journal of Experimental Social Psychology, Journal of Personality, Journal of Personality and Social Psychology, Journal of Research in Personality, Personality and Social Psychology Bulletin, Psychological Science, and Social and Personality Psychology Science). The authors extracted data from a random 20% of the empirical studies published in each journal in each year between 2011 and 2019. According to the authors, their chosen start date corresponded to the beginning of the Replication Crisis in psychology, which many would place at the publication of impactful papers by Bem (2011) and Simmons et al. (2011). As in Fraley et al. (2022), (a) only data from studies that employed between group comparisons were employed for the below analyses; and (b) only studies in Social and Personality psychology were included. Studies published in Psychological Science, which is a general psychology journal, were individually screened by Fraley et al. (2022) for their relevance to Social or Personality psychology and excluded appropriately. Their openly available dataset was obtained from their supplementary materials (i.e., osf.io/rvbxp). The analytic dataset included sample sizes from 3047 studies (range 113 to 631 studies per journal).

Results

Prevalence of Null Hypothesis Significance Testing in the IRAP literature

Given that the IRAP emerged from the behavioral tradition (Barnes-Holmes et al., 2010), I considered it possible that some IRAP studies may employ inference methods other than Null Hypothesis Significance Testing (NHST), including Single Case Experimental Design methods. Such studies would be both likely to employ smaller sample sizes and would not be susceptible to issues of statistical power in quite the same way as those which explicitly employed NHST. As such, before applying any critique that was relevant only to studies employing NHST, I first began by quantifying the proportion of IRAP publications that employed NHST.

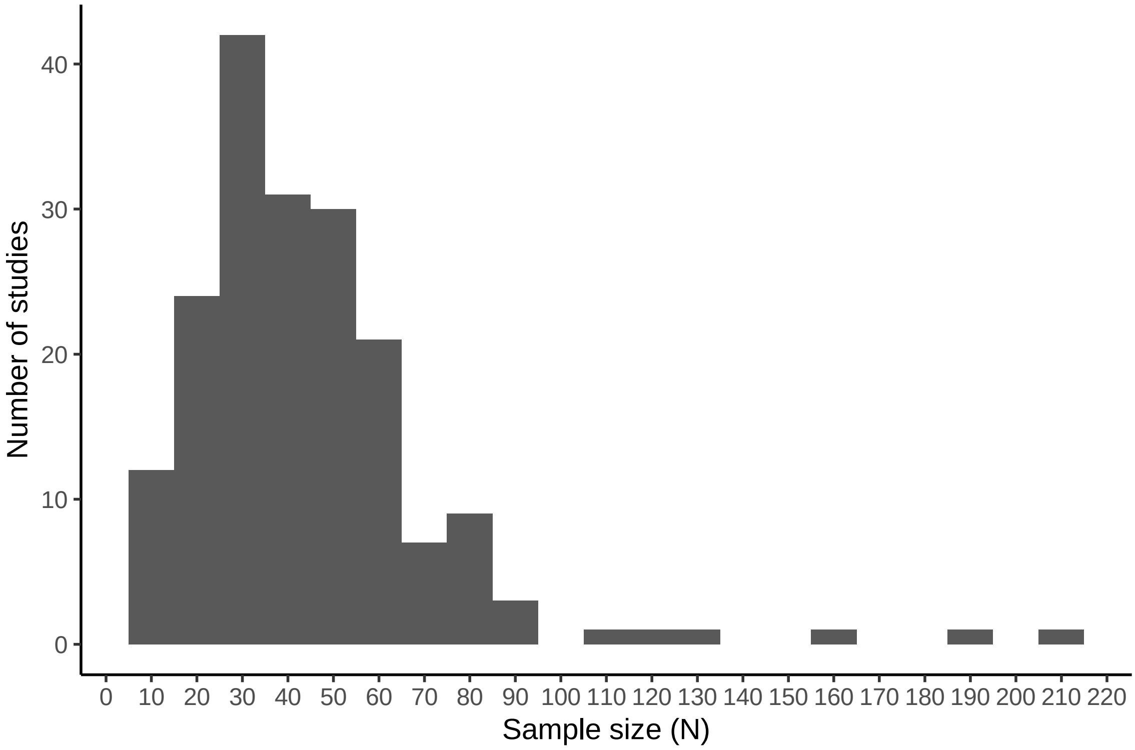
One publication did not report the sample size (Cullen & Barnes-Holmes, 2008). The remaining 150 publications contained 188 studies. Just 3 studies in two articles (1.6%) did not employ NHST (i.e., Jackson et al., 2016; Rafacz et al., 2019). The overwhelming majority of published IRAP studies have employed NHST (185 studies, 97.9%). As such, the constraints of inference via NHST necessarily apply to these studies.

Sample size in the IRAP literature

Distribution of sample sizes

In the 185 studies that both employed NHST and reported sample sizes, a total of 8384 participants were reported. Sample sizes ranged from 9 to 210 participants, Median = 41, Median Absolute Deviation (MAD) = 17.8. A histogram of the distribution of sample sizes in IRAP research can be found in Figure 1.

Figure 2. Histogram of the distribution of sample sizes in IRAP studies



Change in sample size per study over time

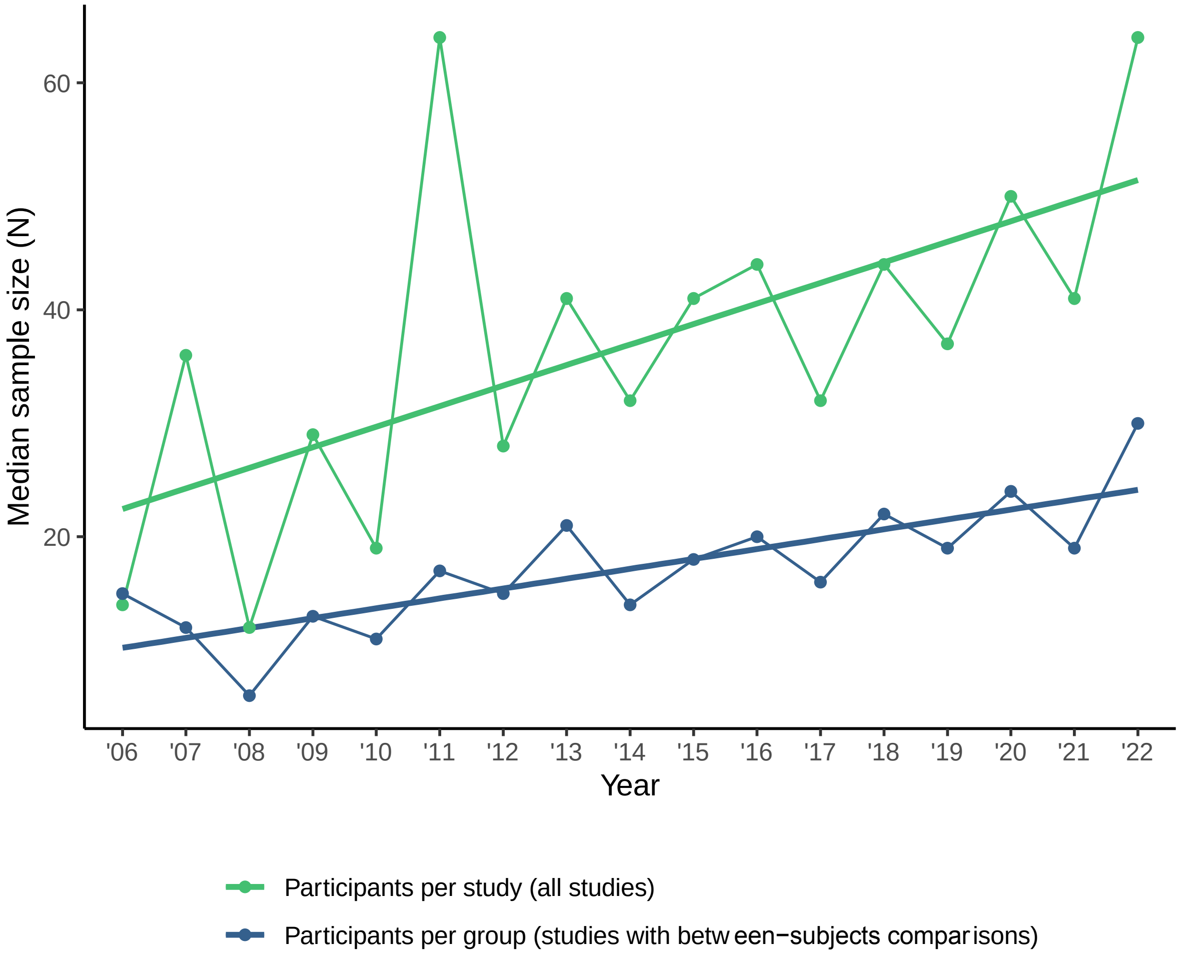
Change in sample sizes over time was quantified by calculating the median sample size for each year. In order to illustrate the change in sample sizes used in all IRAP studies over time, the median sample size per study per year using all studies was calculated (see Figure 3, green line). In this and all subsequent figures, the straight line represents the fitted Ordinary Least Squares linear regression line (discussed later) and the shaded region around it represents its 95% Confidence Interval. As can be seen in Figure 3, across years, median sample sizes in IRAP studies have been small (range of medians 12 to 64). Notably, given that these are medians, this also implies that half of the studies in each year employ samples that are even smaller than this. Data from all plots are available in table format in the supplementary materials [URL removed for peer review blinding].

To quantify any changes in an accessible manner, an Ordinary Least Squares linear regression was fit to the data with median sample size as the dependent variable and year as the independent variable. Year was rescored so that 2006 was the intercept. Results demonstrated that the estimated median sample size per study increased from an estimated 22.5, 95% CI [10.7, 34.2] in 2006 (the model intercept) by an average of 1.8, 95% CI [0.6, 3.1], *p* = .008 participants per year (see Figure 3, green line).

Change in sample size per group over time

While the previous quantification benefits from including data from all studies and therefore providing an overview of the literature, it does not necessarily compare like with like over time in a way that facilitates understanding statistical power. For example, imagine two studies: the first has a sample size of 100 in two between-group conditions, and the second has a sample size of 150 in three between-group conditions. The studies have different sample sizes, but this does not translate to them having higher statistical power for their pairwise group comparisons: both have an average of 50 participants per group. To compare like with like, it is useful to also plot the median sample per group (aka cell within the factorial design). The median sample size per experimental group per year was therefore calculated by dividing the study sample size by the number of between-group conditions employed within each study. Similarly, to compare like with like, only studies employing between groups comparisons were included (69.5% of all studies). This rationale follows that employed by Fraley et al. (2022). Figure 3 therefore also plots the median sample size per study per year using all studies (see Figure 3, blue line).

Figure 3. Median sample size per year in IRAP studies

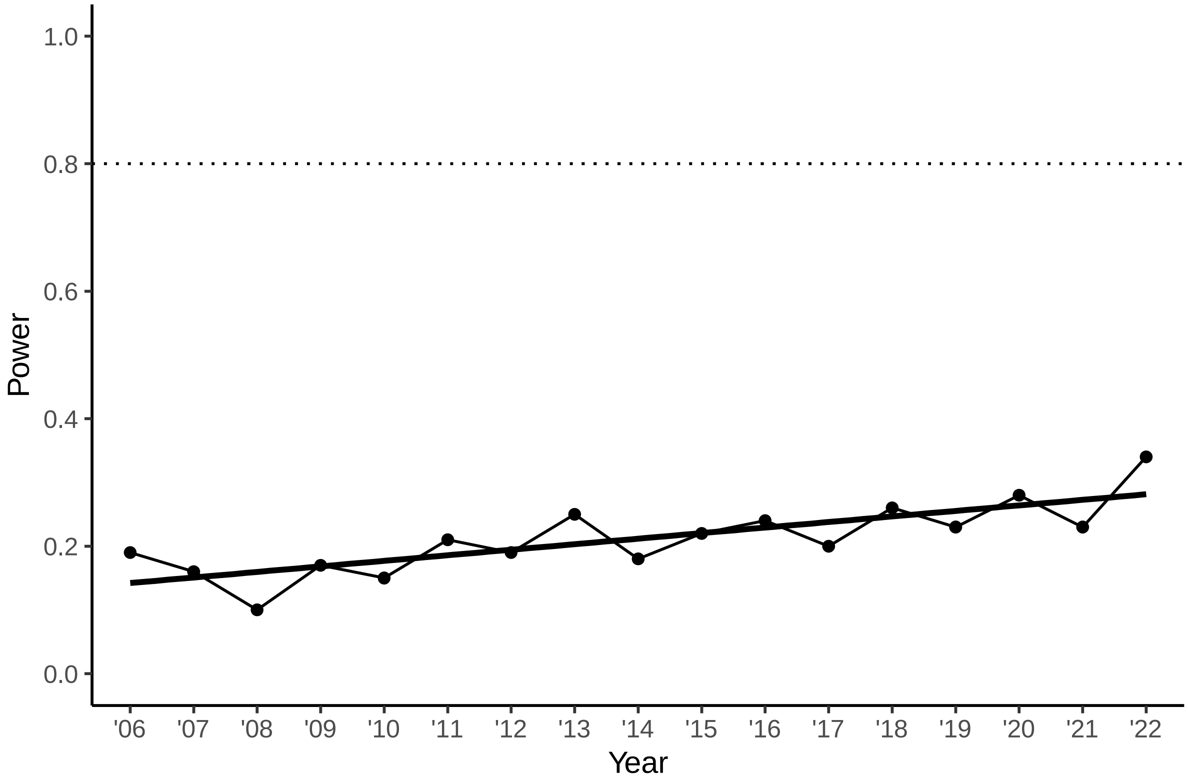


A similar regression was fitted this time using median sample sizes per group as the dependent variable. Results demonstrated that the estimated median sample size per group increased from an estimated 10.2, 95% CI [6.8, 13.7] in 2006 (the model intercept) by an average of 0.9, 95% CI [0.5, 1.2], *p* < .001 participants per year (see Figure 3, blue line). Caveats about robustness notwithstanding, because this analysis compares like with like better than the previous one, these results likely represent a more appropriate estimation of the change in sample sizes over time with regard to implications for statistical power.

Statistical power in the IRAP literature

I then quantified the median statistical power that these median sample sizes imply. It is important to recall that statistical power is a function of multiple variables other than sample size, and power and sample size should not be treated as synonymous. Power is a function of (1) a specific type of test, (2) its alpha level, (3) whether one-tailed or two-tailed hypotheses are employed, (4) the sample size estimate, and (5) the effect size of interest. Choices must be made for each of these in order to estimate power. As in Fraley et al. (2022), I therefore (1) limited my consideration to specific analyses (i.e., independent *t*-tests or Pearson's *r* correlations, using equivalent effect sizes for each); (2) employed the standard alpha level of .05; (3) employed modal two-tailed comparisons; (4) estimated the median sample size from the literature that used broadly consistent designs (i.e., median *N* estimated from studies that reported at least one between-group comparison, excluding exclusively within-sample designs); and (5) estimated the ability to detect an effect size of Cohen's $d = .408$. This effect size is equivalent to a Pearson's $r = .20$ (as used in as used in Fraley et al., 2022), which has been shown in multiple meta-analyses to be approximately the average size effect found across the psychology research literature (Gignac & Szodorai, 2016; Hemphill, 2003; Richard et al., 2003).

Figure 4. The statistical power to detect the average published effect size (Cohen's *d* = 0.408 or Pearson's *r* = 0.20) implied by the median sample size per group per year in IRAP studies



Implied statistical power was calculated using the above parameters for each year using the R package pwr (Champely, 2016). Results can be found in Figure 4. The dotted line represents Cohen's (1988) commonly accepted guideline for a minimum of at least 80% power. As can be seen in the plot, the implied statistical power to detect an average effect size in the IRAP literature is very low (range 10% to 34%).

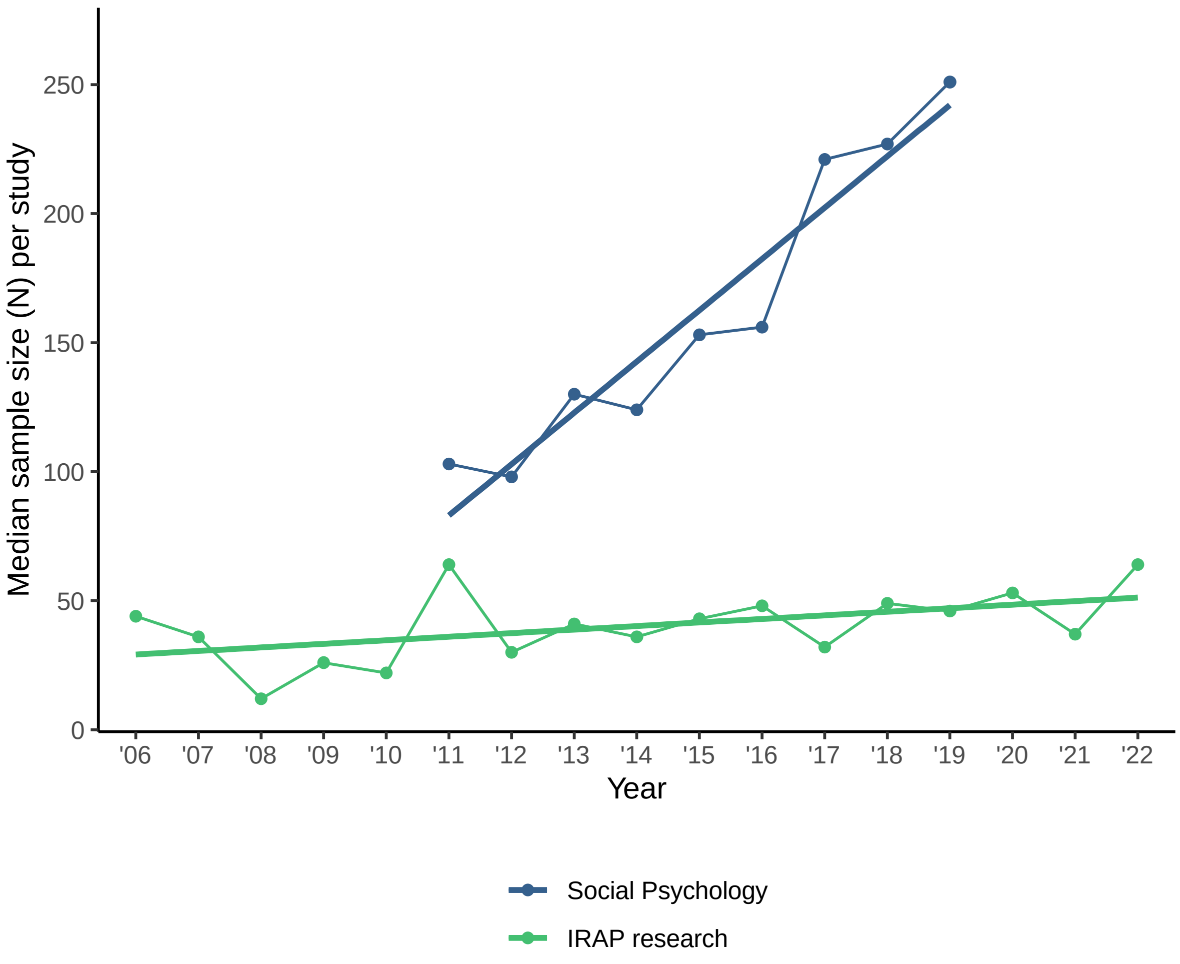
To illustrate the magnitude of change in power over time, power and median sample size were entered into a regression as the dependent variable, otherwise similar to the previous analyses. Results demonstrated that the implied statistical power to detect the average published effect size (Cohen's *d* = 0.408, equivalent to Pearson's *r* = 0.20) was increasing from an estimated 14.2%, 95% CI [10.8, 17.7] in 2006 (the model intercept) by an average of 0.9, 95% CI [0.5, 1.2], $p < .001$. This estimate was used to calculate how long it would take to achieve Cohen's (1988) recommendation of power of at least 80%. If this linear rate of growth continued, the median IRAP sample size would take another 58 years to reach this commonly accepted minimum for statistical power (i.e., in 2080). Note that this is not a prediction that the rate of growth will be linear or stable, but a prediction based on these assumptions, in a world where IRAP researchers' behavior stays on their current trajectory.

Comparing the IRAP literature with Social and Personality Psychology

The analyses reported in the previous section demonstrate that median sample sizes and implied statistical power in the IRAP literature are low in absolute terms. It is useful to supplement this with a relative comparison, i.e., to research in other areas, using the dataset provided by Fraley et al. (2022). Whereas Fraley et al. (2022) calculated median sample size by journal, I calculate a single overall median for Social and Personality and psychology in order to make a simple comparison between these two literatures.

However, Fraley et al. (2022) did not extract the number of between-group conditions employed in each study, only the design (between, within, or mixed within-between) and sample size. The most direct and informative comparison possible with the IRAP literature is, therefore, the comparison of median sample sizes by study in studies that employed between-group comparisons (i.e., where the analyses in the previous section compared medians for each group rather than for each study). Medians and power will therefore differ between the analyses reported in these analyses and those reported in the previous section. Analyses in the previous section represent more appropriate absolute estimates, whereas those reported here are more useful for comparisons.

Figure 5. Median sample size per year in IRAP studies compared to Social and Personality Psychology studies



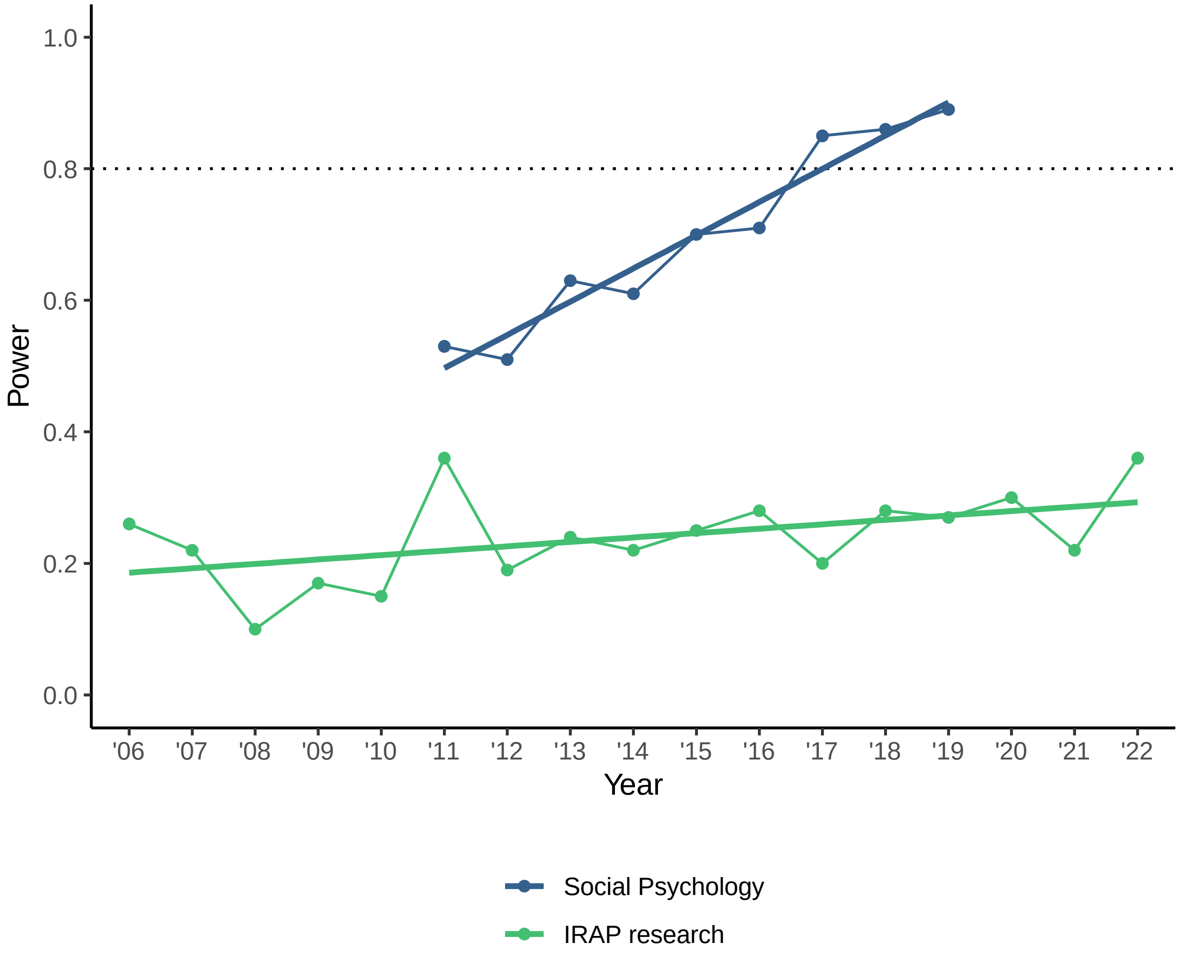
Comparing median sample sizes

Figure 5 illustrates the median sample sizes per study per year for IRAP studies (green line) compared to the Social and Psychology literature (blue line). A regression was fitted to the data, with median sample size as the dependent variable; and year, literature (IRAP literature vs. Social and Personality psychology literature), and their interaction as independent variables. The estimate of the interaction effect was used to test the hypothesis that the change in median sample size over time was larger in Social and Personality psychology studies than in IRAP studies. Results demonstrated that this difference was statistically significant, substantive in size, and in the predicted direction, $B = 18.5$, 95% CI [14.5, 22.5], $p < .001$.

Comparing statistical power

The statistical power implied by each median sample size was then calculated using the same manner as previously. See Figure 6, in which the dotted line represents Cohen's (1988) commonly accepted guideline for a minimum of at least 80% power. Power estimates were then entered into a regression as the dependent variable. The independent variables were identical to the previous regression model. The estimate of the interaction effect was used to test the hypothesis that the change in implied statistical power to detect the average published effect size (Cohen's $d = 0.408$, equivalent to Pearson's $r = 0.20$) over time was larger in Social and Personality psychology studies than IRAP studies. Results demonstrated that this difference was statistically significant, substantive in size, and in the predicted direction, $B = .04$, 95% CI [.03, .06], $p < .001$.

Figure 6. The statistical power to detect the average published effect size (Cohen's *d* = 0.408 or Pearson's *r* = 0.20) implied by the median sample size per study per year in IRAP studies compared to Social and Personality Psychology studies



Discussion

Results demonstrated the sample sizes employed in IRAP literature have been problematically small in both absolute terms and relative to the sample sizes employed in Social and Personality psychology studies. The statistical power to detect the average effect size in published psychology research (i.e., Cohen's *d* = 0.408, equivalent to Pearson's *r* = .20: Gignac & Szodorai, 2016; Hemphill, 2003; Richard et al., 2003) implied by the median sample sizes in IRAP research are also problematically low in both absolute terms and relative to that in the Social and Personality psychology literature. Implied statistical power was < 35% in all years, where Cohen (1988) recommends a minimum power of 80%.

Although the sample sizes employed in IRAP studies (and therefore implied power) are increasing detectibly over time, the rate of this increase is too low to make meaningful progress. Results suggest that if the current rate of linear increase in median sample sizes continued it would take 58 years for just half of IRAP studies to meet Cohen's (1988) recommendation of power of a minimum of least 80% (i.e., in 2080).

When compared with the sample sizes employed in 3047 studies published across nine Social and Personality psychology journals (Fraley et al., 2022), results demonstrated that median samples sizes (and therefore implied statistical power) increased at a much greater rate in Social and Personality psychology between 2011 and 2019 than they did in IRAP studies between 2006 and 2022. Furthermore, median sample sizes and their implied statistical power were lower in the IRAP literature in all years than they were at the beginning of the Replication Crisis in Social and Personality Psychology in 2011, and indeed in all subsequent years (see Figures 5 and 6).

The necessary implications of this low power must be appreciated: they imply that of all effects of average size in the population that have been studied in IRAP publications, those effects that truly exist in the population were not detected in nearly two-thirds of cases. Even for those researchers who self-identify as employing NHST in an inductive manner (e.g., Kavanagh, Matthyssen, et al., 2019), this would imply most opportunities to induce hypotheses from data are missed because there was insufficient power to generate significant results. Simulation studies such as those by Bakker and colleagues (2012) and LeBel and colleagues (2017) demonstrate that a smaller number of better-powered studies produces a larger number of true inferences than a larger number of less well-powered studies. As such, even researchers with very limited resources available to them would be better to allocate those resources to larger studies.

Recommendations for power analyses and sample sizes

Readers might reasonably seek concrete recommendations for sample sizes in future IRAP studies. Unfortunately, my answer may be unsatisfying: (1) it depends, and (2) it should probably be much larger than you think. This position is drawn from a few sources. First, the authors of the seminal article "False Positive Psychology" paper (2011) have since stated that one of their biggest regrets in that paper was to specify a minimum sample size, due to the subsequent misuse of that recommendation (Simmons et al., 2018). It's worth noting that similar misuses of sample size recommendations are already visible within the IRAP literature: Vahey and colleagues (2015) reported an analysis of the IRAP's clinical criterion validity and the results of multiple power analyses based on their effect size estimate. Citations of Vahey et al. (2015) often inappropriately or inaccurately cite these recommendations. For example, many papers have cited the sample size recommendations reported in Vahey et al.'s (2015) abstract (i.e., "*N* = 29 to 37") even when the authors employed a completely different analysis (e.g., other than a one-tailed Pearson's *r* correlation with α-level = .05) and/or are conducting research outside of the clinical domain to which Vahey et al. (2015) limited the scope of their meta-analysis (e.g., Farrell et al., 2015; Kavanagh et al., 2016; Kavanagh, Roelandt, et al., 2019; Leech & Barnes-Holmes, 2020; Maloney & Barnes-Holmes, 2016).

Second, research has demonstrated that researchers' intuitions about the relationship between statistical power and sample size are inaccurate and tend to greatly overestimate power (Bakker et al., 2016). Deeper engagement with sample size choices and their power implications is therefore warranted.

In light of the above, I am reluctant to provide sample size recommendations that risk being cited or followed unthinkingly, absent of context or specifics. Instead, I encourage readers to think more deeply about their inference method, engage with the concept and calculation of statistical power, and plan their studies accordingly. In general, power analyses should be conducted and reported in a reproducible manner, for example using the pwr R package (Champely, 2016) or Superpower (Caldwell et al., 2022). For a beginner introduction to statistical power using interactive visualizations, see Magnusson (2023). For a seminal book-length treatment see Cohen (1988). It is also worth considering that tests of interaction effects are often reported in the IRAP literature, typically via the interaction term in multiway ANOVAs. Determining statical power for interaction terms is more complex than implied by some power analysis software such as G*Power (Faul et al., 2007), and I recommend that researchers should base power analyses on specific forms of expected or plausible interactions (e.g., reversed, fully attenuated, partially attenuated) and their post hoc tests rather than the interaction term in an ANOVA alone (see Caldwell et al., 2022; Sommet et al., 2022).

Separately, sample size determination also involves additional considerations beyond power analysis, such as the availability of resources and desired precision (Lakens, 2022). Researchers should consider that sample sizes employed in power analyses do not have to be based on as-yet-unknown estimates of the effect size they are studying, but can instead be based on the researchers' Smallest Effect Size of Interest (SESOI: Lakens, Scheel, et al., 2018). Previous work has also pointed out ways in which standard power analyses may still under-power studies due to between-study heterogeneity. This may require that sample sizes be increased further. For a discussion of this issue as well as materials for performing power analyses that can account for this see McShane and Böckenholt (2014).

Of course, power is not the only factor that influences the quality of inferences when using NHST. Exploratory multiway ANOVAs are commonly reported in IRAP publications, and yet corrections for multiple testing (i.e., to control the family-wise error rate) are rarely applied. Cramer and colleagues (2016) demonstrated that this factor alone gives rise to false positive rates that are greatly inflated beyond the 5% implied by an α-level of 0.05 (e.g., 14% in 2 X 2 between group ANOVAs). Exactly how high false positive rates are inflated for the specific types of multiway ANOVAs employed in IRAP research (e.g., 4 X 2 mixed within-between repeated measures ANOVAs) remains unclear. Future research may wish to examine this in simulation studies.

Many other factors influence the replicability of findings, providing many opportunities for researchers to improve the strength of the evidence provided by their studies (Nosek et al., 2018). Researchers should therefore also give serious consideration to preregistering their sample size planning justifications, their chosen sample size, stopping rule, precise analysis plan, and other elements of their research, such as their chosen α-level (Lakens, Adolfi, et al., 2018).

While keeping the above numerous qualifiers in mind as to why power calculations must be done for each study with respect to its design, measurement quality, and specific analysis, some readers may nonetheless wish for some discussion of the sample size needed to detect specific effects in future IRAP studies. I will therefore take two examples, both related to common use cases for the IRAP.

First, imagine a researcher who wishes to detect a correlation between IRAP scores and another continuous variable. Power analyses were again conducted using the R package pwr (Champely, 2016). To detect a correlation that is the magnitude of the average effect size observed in published psychological research (i.e., Pearson's $r = 0.20$); with an α-level of 0.05 and a two-sided test (typical for correlations and regression); 194 participants would be needed to obtain the minimum recommended power (80%), or 319 participants for high power (95%). Comparing these sample sizes to the sample sizes observed in the published IRAP literature, 1% of published IRAP studies employed a sample size sufficient to detect the average effect size in the broader psychology literature ($r = .20$) with minimum recommended power (80%), and 0% could detect this with high power (95%). Note that half of all published effects are smaller than the average, and therefore would require even higher sample sizes to detect.

Second, imagine that a researcher wishes to detect a difference in mean IRAP scores between two groups, and they wish to power the study to detect effect sizes that are at least the average transdiagnostic efficacy of Acceptance and Commitment Therapy versus active and inactive control groups as reported in a recent review of meta-analyses (Hedges' *g* = 0.46: Gloster et al., 2020). This effect size is no more or less meaningful than any other for the sake of illustration but is chosen to leverage intuitions the reader may have to define its Smallest Effect Size of Interest. That is, such a study would be powered to detect mean differences on the IRAP between groups that are no smaller than the general efficacy of ACT. Loosely speaking, I am inviting the reader to think about other effect sizes they have an intuition for in order to provide a benchmark. For example, if you are interested in between groups differences that are smaller than the average efficacy of ACT, you would need more participants than this again. With an effect size of Cohen's *d* = .46 (i.e., ignoring the very small difference between Hedges' *g* and Cohen's *d* at this sample size), an α-level of 0.05, a two-sided test, and two equally sized groups; 152 participants would be needed to obtain the minimum recommended power (80%), or 248 participants for high power (95%). Comparing these sample sizes to the sample sizes observed in published IRAP research employing between-groups or mixed within-between designs, 2% of such published IRAP studies employed a sample size sufficient to detect an effect size the same size as the average transdiagnostic efficacy of Acceptance and Commitment Therapy (Hedges' *g* = 0.46: Gloster et al., 2020) with minimum recommended power (80%), and 0% could detect this with high power (95%).

Please note again that the above power analyses are not specific recommendations for future IRAP research, and this article should not be cited as a source of such recommendations, but rather are worked examples of (a) how authors should begin to engage with conducting their own a priori power analyses, and (b) illustrative of how underpowered the published IRAP literature is to detect effect sizes of these magnitudes. I reiterate this caution due to The Law of Lakens' Guidelines, which states that whenever you try to make the point that researchers should not follow certain guidelines, you will nonetheless sometimes be cited as a source of said guidelines (Rohrer, 2023).

Limitations

Extraction of sample sizes

Sample sizes were extracted from the original articles by a single coder and no estimates of inter-rater reliability were therefore produced. All data and code for the current manuscript are public, therefore making the accuracy of these extractions testable in principle. However, given this is a non-trivial task, and inaccurate data extractions could lead to bias, the lack of a second scorer must be acknowledged as a limitation.

Choice of effect size

One limitation of the current study is the choice of the effect size for which power was calculated (i.e., Pearson's r = .20, equivalent to Cohen's d = .408). This value was selected to be able to draw direct comparisons with Fraley et al. (2022) and therefore the literature on Social and Personality psychology. Additionally, this estimate was derived from multiple large-scale meta-science studies estimating the average effect size reported in psychology research across thousands of papers (Gignac & Szodorai, 2016; Hemphill, 2003; Richard et al., 2003). No such estimates exist for the IRAP literature.

One meta-analysis of the IRAP's criterion validity does exist (Vahey et al., 2015), but that study involved a different estimand to the current work (i.e., it provided an answer to a different question). Whereas the current analyses require an estimate of all average observed effect sizes, Vahey et al. (2015) included just 15 of 46 publications they found, and 56 effect sizes of the at least 308 effect sizes reported in those articles. These publications and effect sizes were not sampled randomly but by applying inclusion criteria related to clinical relevance, making their estimate unsuitable for the current purposes.

Potential for systematic differences in effect sizes between literatures

It is important to consider the possibility that the effect sizes observed within the IRAP literature are simply much larger than those observed in other areas of psychology. This would undermine the comparisons between the implied power in the IRAP literature versus the Social and Personality Psychology literature which are based on an assumption of similar average effect sizes. However, this would require that the IRAP itself is unique or distinct in some way that allows it to capture larger-than-average effect sizes. Given that the IRAP is used to study a diverse range of topics covering social psychology (e.g., race and gender: De Schryver et al., 2018; Hughes et al., 2016), clinical psychology (e.g., depression, suicidality, eating disorders, and OCD: Hussey et al., 2016; Hussey & Barnes-Holmes, 2012; Nicholson & Barnes-Holmes, 2012; Parling et al., 2012), and behaviorism (e.g., within Relational Frame Theory: Finn et al., 2016; Pidgeon et al., 2021), the common link across the IRAP literature is therefore not the domain in which the task is employed, but the use of the task itself. While it is quite feasible that effect sizes differ in one domain relative to another (e.g., effect sizes in persuasion research might simply be larger than in perception research, or vice versa), this is not the case when a range of domains is considered.

It is relatively easier to estimate the probability that larger effect sizes will be observed using one task relative to others: average observable effect sizes are limited by the reliability of the task being used to capture them (i.e., via attenuation: Revelle, 2009). The IRAP's reliability has been shown to be less than acceptable according to common recommendations (e.g., $\alpha = .60$ and test-retest $r = .43$ according to one meta-analysis: Greenwald & Lai, 2020; reliability was found to be substantially lower in a different analysis: Hussey & Drake, 2020a; recommendation for $\alpha > .70$: Nunnally & Bernstein, 1994). Measures in Social and Personality psychology have generally been shown to possess higher reliability than this in reliability-generalization meta-analyses (Greco et al., 2018). It is mathematically implausible that a less reliable than average task can consistently capture larger than average effect sizes. As such, it is unlikely that average effect sizes in the IRAP literature are larger than those in other literatures (although the reverse is plausibly the case).

Generalizability

The limitations of the dataset generated by Fraley et al. (2022) covering the Social and Personality psychology literature are discussed in their original article and will be reiterated here. Those authors elected to code articles from a subset of the more prestigious Social and Personality psychology journals, and there is potential that their estimates are not perfectly representative of articles published in other journals in that field. Only increasingly larger-scale assessments of literature can resolve this. There are already signs of progress here: Fraley et al. (2022) have already updated their dataset once (see Fraley & Vazire, 2014) and state that they will continue to do so again in the future.

The current research represents an effort to extend this form of assessment beyond Social and Personality psychology. It is unclear whether the present results for the IRAP literature would generalize to the broader behavioral literature. Given recent calls for scrutiny of the replicability of behavioral research (e.g., Hantula, 2019; Task Force on the Strategies and Tactics of Contextual Behavioral Science Research, 2021), future research should examine median sample sizes in studies employing NHST in the behavioral literature (e.g., journals such as Journal of Contextual Behavioral Science, The Psychological Record, and Perspectives of Behavioral Science).

Conclusion

Given that statistical power across a literature is a key determinant of the replicability of the findings in that literature, these results paint a worrying picture for the replicability of

IRAP research. These concerns add to concerns vocalized elsewhere about the IRAP's reliability (Hussey, 2020; Hussey & Drake, 2020a), a method factor that confounds several common analyses of IRAP data (Hussey & Drake, 2020b), and the fact that most IRAP studies come from a very narrow range of individuals and labs, potentially impacting the replicability and generalizability of claims (Hussey, 2022). Researchers should therefore interpret the results and conclusions of published IRAP research with some caution and be cautious in choosing to employ the IRAP in their work. In line with a recent statement by the Association for Contextual Behavioral Science explicitly embracing the need for replication studies (Task Force on the Strategies and Tactics of Contextual Behavioral Science Research, 2021), direct assessments of the reproducibility and replicability of the published IRAP literature is likely warranted.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology: Recommendations for increasing replicability. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Baker, M., & Dolgin, E. (2017). Cancer reproducibility project releases first results. *Nature*, *541*(7637). https://doi.org/10.1038/541269a

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & Maas, H. L. J. van der. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*. https://doi.org/10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*(7), 169–177.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and

Coherence (REC) model. *The Psychological Record*, *60*(3), 527–542. https://doi.org/10.1007/BF03395726

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Boulesteix, A.-L., Hoffmann, S., Charlton, A., & Seibold, H. (2020). A replication crisis in methodological research? *Significance*, *17*(5), 18–21. https://doi.org/10.1111/1740-9713.01444

Buckwalter, W. (2022). The replication crisis and philosophy. *Philosophy and the Mind Sciences*, *3*. https://doi.org/10.33735/phimisci.2022.9193

Burgos, J. E., & Killeen, P. R. (2019). Suing for Peace in the War Against Mentalism. *Perspectives on Behavior Science*, *42*(2), 241–266. https://doi.org/10.1007/s40614-018-0169-2

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Caldwell, A., Lakens, D., DeBruine, L., Love, J., & Aust, F. (2022). *Superpower: Simulation-Based Power Analysis for Factorial Designs* (0.2.0). https://CRAN.R-project.org/package=Superpower

Champely, S. (2016). *pwr: Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr

Clayton, K., Horrillo, J., & Sniderman, P. M. (2023). The BIAT and the AMP as measures of racial prejudice in political science: A methodological assessment. *Political Science Research and Methods*, *11*(2), 363–373. https://doi.org/10.1017/psrm.2022.56

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12). https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. https://doi.org/10.1037/0033-2909.112.1.155

Corneille, O., & Hütter, M. (2020). Implicit? What Do You Mean? A Comprehensive Review of the Delusive Implicitness Construct in Attitude Research. *Personality and Social Psychology Review*, *24*(3), 212–232. https://doi.org/10.1177/1088868320911325

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L. J., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

Cullen, C., & Barnes-Holmes, D. (2008). Implicit pride and prejudice: A heterosexual phenomenon? In M. A. Morrison & T. G. Morrison (Eds.), *The psychology of modern prejudice* (pp. 195–223). Nova Science Publishers.

De Schryver, M., Hussey, I., De Neve, J., Cartwright, A., & Barnes-Holmes, D. (2018). The PIIRAP: An alternative scoring algorithm for the IRAP using a probabilistic semiparametric effect size measure. *Journal of Contextual Behavioral Science*, *7*, 97–103. https://doi.org/10.1016/j.jcbs.2018.01.001

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in*

Methods and Practices in Psychological Science, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Farrell, L., Cochrane, A., & McHugh, L. (2015). Exploring attitudes towards gender and science: The advantages of an IRAP approach versus the IAT. *Journal of Contextual Behavioral Science*, *4*(2), 121–128. https://doi.org/10.1016/j.jcbs.2015.04.002

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016). Exploring the behavioral dynamics of the implicit relational assessment procedure: The impact of three types of introductory rules. *The Psychological Record*, *66*(2), 309–321. https://doi.org/10.1007/s40732-016-0173-4

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal N-Pact Factors From 2011 to 2019: Evaluating the Quality of Social/Personality Journals With Respect to Sample Size and Statistical Power. *Advances in Methods and Practices in Psychological Science*, *5*(4), 251524592211202. https://doi.org/10.1177/25152459221120217

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLOS ONE*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Gelman, A. (2016, September 21). What has happened down here is the winds have changed. *Statistical Modeling, Causal Inference, and Social Science*. http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gloster, A. T., Walder, N., Levin, M. E., Twohig, M. P., & Karekla, M. (2020). The empirical status of acceptance and commitment therapy: A review of meta-analyses. *Journal of Contextual Behavioral Science*, *18*, 181–192. https://doi.org/10.1016/j.jcbs.2020.09.009

Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Twardy, C., Wang, J., & Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, *7*(7). https://doi.org/10.1098/rsos.200566

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2018). Meta-Analysis of Coefficient Alpha: A Reliability Generalization Study. *Journal of Management Studies*, *55*(4), 583–618. https://doi.org/10.1111/joms.12328

Greenwald, A. G., & Lai, C. K. (2020). Implicit Social Cognition. *Annual Review of Psychology*, *71*(1), 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hantula, D. A. (2019). Editorial: Replication and Reliability in Behavior Science and Behavior Analysis: A Call for a Conversation. *Perspectives on Behavior Science*, *42*(1), 1–11. https://doi.org/10.1007/s40614-019-00194-2

Hayes, J., Stewart, I., & McElwee, J. (2016). Assessing and Training Young Children in Same and Different Relations Using the Relational Evaluation Procedure (REP). *The Psychological Record*, *66*(4), 547–561. https://doi.org/10.1007/s40732-016-0191-2

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual Behavioral Science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*, *1*(1–2), 1–16. https://doi.org/10.1016/j.jcbs.2012.09.004

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. https://doi.org/10.1037/0003-066X.58.1.78

Hughes, S., Barnes-Holmes, D., & Vahey, N. A. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science*, *1*(1–2), 17–38. https://doi.org/10.1016/j.jcbs.2012.09.003

Hughes, S., Cummins, J., & Hussey, I. (2022). Effects on the Affect Misattribution Procedure are strongly moderated by influence awareness. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01879-4

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes□Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology*, *46*(5), 632–648. https://doi.org/10.1002/ejsp.2207

Hussey, I. (2020). The IRAP is not suitable for individual use. *Preprint*. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I. (2022). *Reply to Barnes-Holmes & Harte (2022) "The IRAP as a Measure of Implicit Cognition: A Case of Frankenstein's Monster."* PsyArXiv. https://doi.org/10.31234/osf.io/qmg6s

Hussey, I., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice*, *19*(4), 573–582. https://doi.org/10.1016/j.cbpra.2012.03.002

Hussey, I., Barnes-Holmes, D., & Booth, R. (2016). Individuals with current suicidal ideation demonstrate implicit "fearlessness of death." *Journal of Behavior Therapy and Experimental Psychiatry*, *51*, 1–9. https://doi.org/10.1016/j.jbtep.2015.11.003

Hussey, I., & Drake, C. E. (2020a). The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis. *Preprint*. https://doi.org/10.31234/osf.io/ge3k7

Hussey, I., & Drake, C. E. (2020b). *The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess*. PsyArXiv. https://doi.org/10.31234/osf.io/sp6jx

Hussey, I., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J., & Nosek, B. A. (2019). *The Attitudes, Identities, and Individual Differences (AIID) Study and Dataset*. https://doi.org/10.17605/OSF.IO/PCJWF

Hussey, I., Thompson, M., McEnteggart, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). Interpreting and inverting with less cursing: A guide to interpreting IRAP data. *Journal of Contextual Behavioral Science*, *4*(3), 157–162. https://doi.org/10.1016/j.jcbs.2015.05.001

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jackson, M. L., Larry Williams, W., Hayes, S. C., Humphreys, T., Gauthier, B., & Westwood, R. (2016). Whatever gets your heart pumping: The impact of implicitly selected reinforcer-focused statements on exercise intensity. *Journal of Contextual Behavioral Science*, *5*(1), 48–57. https://doi.org/10.1016/j.jcbs.2015.11.002

Kavanagh, D., Hussey, I., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2016). Using the IRAP to explore natural language statements. *Journal of Contextual Behavioral Science*, *5*(4), 247–251. https://doi.org/10.1016/j.jcbs.2016.10.001

Kavanagh, D., Matthyssen, N., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., & Vastano, R. (2019). Exploring the use of pictures of self and other in the IRAP: Reflecting upon the emergence of differential trial type effects. *International Journal of Psychology & Psychological Therapy*, *19*(3), 323–336.

Kavanagh, D., Roelandt, A., Van Raemdonck, L., Barnes-Holmes, Y., Barnes-Holmes, D., & McEnteggart, C. (2019). The On-Going Search for Perspective-Taking IRAPs: Exploring the Potential of the Natural Language-IRAP. *The Psychological Record*, *69*(2), 291–314. https://doi.org/10.1007/s40732-019-00333-w

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kilroe, H., Murphy, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2014). Using the T-IRAP interactive computer program and applied behavior analysis to teach relational responding in children with autism. *Behavioral Development Bulletin*, *19*(2), 60–80. https://doi.org/10.1037/h0100578

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R.,

… Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Lakens, D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*(2), 230–243. https://doi.org/10.1037/pspi0000049

Leech, A., & Barnes-Holmes, D. (2020). Training and testing for a transformation of fear and avoidance functions via combinatorial entailment using the Implicit Relational Assessment Procedure (IRAP): Further exploratory analyses. *Behavioural Processes*, *172*. https://doi.org/10.1016/j.beproc.2019.104027

Levin, M. E., Hayes, S. C., & Waltz, T. (2010). Creating an implicit measure of cognition more suited to applied research: A test of the Mixed Trial—Implicit Relational

Assessment Procedure (MT-IRAP). *International Journal of Behavioral Consultation and Therapy*, *6*(3), 245–262. https://doi.org/10.1037/h0100911

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology / Psychologie Canadienne*, *61*, 281–288. https://doi.org/10.1037/cap0000236

Magnusson, K. (2023). *Understanding Statistical Power and Significance Testing—An Interactive Visualization*. https://rpsychologist.com/d3/nhst/

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Maloney, E., & Barnes-Holmes, D. (2016). Exploring the Behavioral Dynamics of the Implicit Relational Assessment Procedure: The Role of Relational Contextual Cues Versus Relational Coherence Indicators as Response Options. *The Psychological Record*. https://doi.org/10.1007/s40732-016-0180-5

McShane, B. B., & Böckenholt, U. (2014). You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*, b2535. https://doi.org/10.1136/bmj.b2535

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nicholson, E., & Barnes-Holmes, D. (2012). Developing an implicit measure of disgust propensity and disgust sensitivity: Examining the role of implicit disgust propensity and sensitivity in obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(3), 922–930. https://doi.org/10.1016/j.jbtep.2012.02.001

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In JA. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). Psychology Press.

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd edition). McGraw-Hill.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Page, L., Noussair, C. N., & Slonim, R. (2021). The replication crisis, the rise of new research practices and what it means for experimental economics. *Journal of the Economic Science Association*, *7*(2), 210–225. https://doi.org/10.1007/s40881-021-00107-7

Parling, T., Cernvall, M., Stewart, I., Barnes-Holmes, D., & Ghaderi, A. (2012). Using the Implicit Relational Assessment Procedure to Compare Implicit Pro-Thin/Anti-Fat Attitudes of Patients With Anorexia Nervosa and Non-Clinical Controls. *Eating Disorders*, *20*(2), 127–143. https://doi.org/10.1080/10640266.2012.654056

Payne, K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, *89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277

Pidgeon, A., McEnteggart, C., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2021). Four self-related IRAPs: Analyzing and interpreting effects in light of the DAARRE model. *The Psychological Record*, *71*(3), 397–409. https://doi.org/10.1007/s40732-020-00428-9

Rafacz, S. D., Houmanfar, R. A., Smith, G. S., & Levin, M. E. (2019). Assessing the effects of motivative augmentals, pay-for-performance, and implicit verbal responding on cooperation. *The Psychological Record*, *69*(1), 49–66. https://doi.org/10.1007/s40732-018-0324-x

Revelle, W. (2009). Chapter 7: Classical Test Theory and the Measurement of Reliability. In *An introduction to psychometric theory with applications in R*. https://personality-project.org/r/book/Chapter7.pdf

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rohrer, J. M. (2023, March 7). Non-representative samples! What could possibly go wrong? *The 100% CI*. https://www.the100.ci/2023/03/07/non-representative-samples-what-could-possibly-go-wrong/

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the Implicit Association Test: The Recoding-Free Implicit Association Test (IAT-RF). *Quarterly Journal of Experimental Psychology (2006)*, *62*(1), 84–98. https://doi.org/10.1080/17470210701822975

Schimmack, U. (2021). The Implicit Association Test: A Method in Search of a Construct. *Perspectives on Psychological Science*, *16*(2), 396–414. https://doi.org/10.1177/1745691619863798

Sidman, M. (1960). *Tactics of scientific research*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-Positive Citations. *Perspectives on Psychological Science*, *13*(2), 255–259. https://doi.org/10.1177/1745691617698146

Sommet, N., Weissman, D., Cheutin, N., & Elliot, A. J. (2022). *How many participants do I need to test an interaction? Conducting an appropriate power analysis and achieving sufficient power to detect an interaction*. OSF Preprints. https://doi.org/10.31219/osf.io/xhe3u

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's Replication Crisis and Clinical Psychological Science. *Annual Review of Clinical Psychology*, *15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Task Force on the Strategies and Tactics of Contextual Behavioral Science Research. (2021). *Adoption of Open Science Recommendations | Association for Contextual Behavioral Science*. https://contextualscience.org/news/adoption_of_open_science_recommendations

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 59–65. https://doi.org/10.1016/j.jbtep.2015.01.004

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Xu, F. K., Nosek, B. A., & Greenwald, A. G. (2014). Psychology data from the Race Implicit Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data*, *2*(1). https://doi.org/10.5334/jopd.ac