A systematic review of the prevalence of Null Hypothesis Significance Testing,

sample sizes, and implied statistical power in research using

the Implicit Relational Assessment Procedure (2006-2022)

Ian Hussey

Here I provide a systematic review of published research using the Implicit Relational Assessment Procedure (IRAP). This manuscript does not perform any synthesis of the research, it merely describes the search and exclusion process of this systematic review. The publications returned by this systematic review can be used in other work that seeks to do such syntheses (e.g., reviews of methods or findings).

In 2011, two remarkable articles were published, and the field of psychology is still grappling with their implications. 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 across science has been heterogenous. Various other fields have

## Sample size, statistical power and replicability

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 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 higher 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).[[1]](#footnote-1) 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).

In order 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, the 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 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 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 attenutation 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 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), are almost ubiquitous in IRAP research. Research has demonstrated that the modal use of multiway ANOVA inflates false positive rates much higher than the 5% rate implied by the standard alpha level of 0.05 (Cramer et al., 2016). 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 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 high statistical power is a desirable property regardless of whether a researcher is employing Null Hypothesis Significance Testing in the in 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 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). Regardless of a researchers’ 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. This effort is in line with the Association for Contextual Behavioral Science’s recently announced support for an active focus on the replicability (Task Force on the Strategies and Tactics of Contextual Behavioral Science Research, 2021). I performed a systematic review of published research using the IRAP and applied Fraley et al.’s (2022) approach to estimating median sample size over time and its implied statistical power. Comparisons were then made between the IRAP literature and 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. In order 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 based on the example provided by Fraley et al. (2022).

## Systematic review of research designs in the IRAP research (2006-2022)

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. 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.

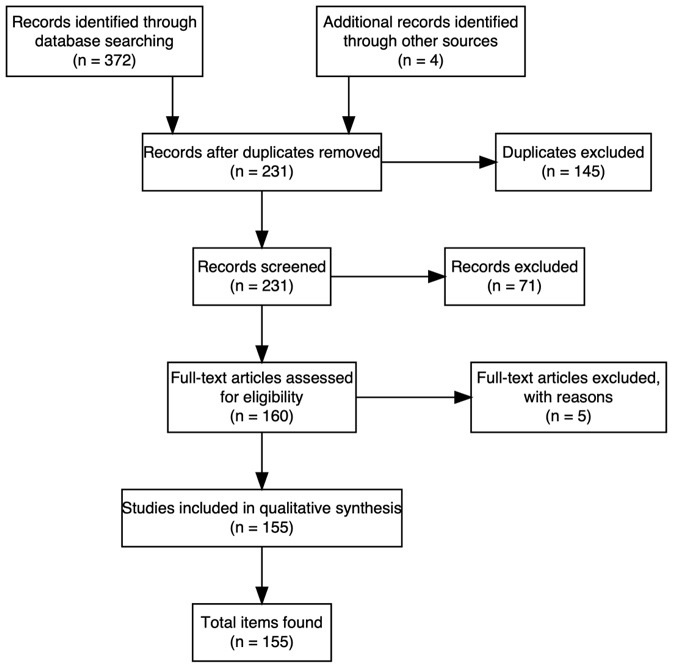
A PRISMA flow chart detailing all exclusions can be found in Figure 1 (Moher et al., 2009). After duplicates were removed, 231 records remained. These were screened based on their title and abstract. Inclusion criterion was the use of the Implicit Relational Assessment Procedure (IRAP) within the study. 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. 160 records remained after title and abstract exclusions. The full texts of these articles was then screened using the same inclusion criterion. Five articles were excluded based on this full text search. In four of these cases, we sought additional details of the procedure used from the authors of the articles. Based on the information provided, four articles were excluded for employing a task other than a standard IRAP (Baker et al., 2015, 2017; Smith et al., 2022; Szarko et al., 2022). The fifth article was excluded because it did not employ an IRAP, although this was not initially apparent from its title and abstract (Perez et al., 2020). After all exclusions, 155 published articles and book chapters using the IRAP were returned by the review. All materials needed to computationally reproduce each stage of this systematic review are freely available in the supplementary materials (XXX) and can be updated by others or used for other evidence synthesis or meta-science purposes.

The full text for each record was inspected in order 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-subjects conditions; and whether the study reported employing Null Hypothesis Significance Testing (NHST). Note that comparisons among multiple IRAP trial-types was excluded from consideration when labelling 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 labelled “mixed”. As such, “mixed” refers not only to mixed within-between research designs but also articles that report both within and between designs. This was suitable for the current analytic purposes, which required excluding the purely within-subject studies from the analyses in order to estimate statistical power correctly (i.e., using those studies employing at least one between-subjects analysis).

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

Fraley et al. (2022) recently updated their previous review of the research designs employed in nine Social and Personality Psychology journals (Fraley & Vazire, 2014): 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, the date range corresponded to the beginning of the Replication Crisis in psychology, specifically the publication of impactful papers by Bem (2011) and Simmons et al. (2011).

**Figure 1.** PRISMA flow chart for systematic review



# Results

## Analytic strategy

The analyses reported here broadly follow those reported by Fraley et al. (2022). In the first part, I assess the distribution of sample sizes in the IRAP literature as a whole, and how median sample sizes have changed over time. The median is more suitable than the mean due to the strong skew exhibited by count data.

In the second part, I quantify the median statistical power that these median sample sizes imply. Importantly, statistical power is not a property of a sample size and the two 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) limit my consideration to specific analyses (i.e., independent *t*-tests); (2) employ the standard alpha level of .05; (3) employ the modal two-tailed comparison; (4) estimate 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-groups comparison); and (5) estimate the ability to detect an effect size of Cohen’s *d* = .408, as used in as used in Fraley et al. (2022). This effect size is equivalent to a Pearson’s *r* = .20, 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). Other effect sizes are of course possible, but the specific choice of effect size is relatively less important. Although estimates of power will differ between different effect sizes of interest, any reasonable choice of effect size allows us to study (a) changes in power over time and (b) differences in average power between research literatures.

## Prevalence of Null Hypothesis Significance Testing in the IRAP literature

XXX

## Sample size in the IRAP literature

XXX

## Statistical power in the IRAP literature

XXX

## Comparing sample size and statistical power in the IRAP literature versus Social and Personality Psychology literature

XXX

I quantify median sample size over time and its implied statistical power in two different ways. First, I do this in a more precise way that makes fewer assumptions… [using sample size per cell AND D = .408 AND N PER GROUP to estimate median N and power for the IRAP literature over time]

Second, I compare the IRAP literature to the social and personality psychology literature by reanalyzing Fraley et al.’s (2022) data. This requires relaxing an assumption, from N per cell to sample size as a whole USING R = .20 AND THE WHOLE SAMPLE SIZE. This necessarily produces different estimates of power between the two analyses, and the previous one should be considered the more precise one. The latter one facilitates comparisons between fields, and over time. It demonstrates that the IRAP literature has lower sample sizes and power than social and personality psychology did at the start of the replicability crisis.

[include ranked regressions on individual Ns?? Someone is going to ask for change over time]

# Discussion

XXX

# Author note

Ian Hussey, Ruhr University Bochum, Germany. [ian.hussey@rub.de](mailto:ian.hussey@rub.de).

# Statements and Declarations

## Conflict of Interest

The author declares that he has no relevant financial or non-financial interests to disclose.

## Funding

This research was supported by the META-REP Priority Program of the German Research Foundation (#464488178).

## Availability of data, code and materials

All data, code and materials are available at XXX.

# 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, 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

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*, 0956797616647519. 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

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*, 527–542.

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

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

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

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

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

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

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

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

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

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

Hussey, I., & Drake, C. E. (2020). 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

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

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. APA PsycInfo.

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. psyh. 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

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

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. psyh. https://doi.org/10.1037/h0100911

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

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

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

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

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

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

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

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

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

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

1. For a beginner introduction to statistical power using interactive visualizations, see Magnusson (2023). For a seminal book-length treatment see Cohen (1977). For accessible implementations of power analyses in R see the pwr package (Champely, 2016). [↑](#footnote-ref-1)