The Implicit Relational Assessment Procedure's trial-types are not independent

Ian Hussey

The Implicit Relational Assessment Procedure (IRAP) can be scored in different ways, as a single overall score or as one score for each its four trial-types. Over time, studies have tended to employ the latter method on the basis of a pervasive belief that the four trial types are independent of one another. Inspection of the literature shows no direct evidence to support this claim and one experiment contradicting it. I conducted a more general investigation of this question using a large existing dataset of 1688 participants who completed one of 41 distinct IRAPs in 17 domains. Scores for each trial-type within each IRAP were correlated with one another in order to assess their independence. 27% of correlations were significantly different from zero. 74% of IRAPs demonstrated at least one detectable correlation among its trial-types. A multi-level meta-analysis demonstrated that IRAP trial-types are on average significantly correlated ($r = .21$, $p < .00000000000001$), albeit with significant heterogeneity between domains. In addition to the existing experimental evidence, this provides large scale correlational evidence that against the claim that IRAP's trial-types are in general independent. Based on the available evidence, researchers should start from the default assumption that the trial types are non-independent, and therefore that the IRAP should be analyzed as a single score. Recommendations for how scoring decisions should be made are provided, along with more general recommendations for ensuring the replicability of findings.

The Implicit Relational Assessment Procedure (IRAP) has been used various as an measure of implicit attitudes and of the strength of relational responding (Barnes-Holmes et al., 2010; Vahey et al., 2015). Whereas other common implicit measures, such as the Implicit Association Test (Greenwald et al., 1998), are typically scored and analyzed as a single score (e.g., representing Black people – positive / White people – negative), the IRAP is sometimes scored as a single overall score, other time one score is calculated for each trial-type (e.g., separately for Black people – positive, Black people – negative, White people – positive, and White people – negative), and indeed other combinations (Barnes-Holmes et al., 2010; Hussey et al., 2015).

Choices on whether to score IRAP data as one overall score or four trial-type scores are frequently guided by the claim that the four IRAP trial-types are independent of one another and therefore should not be comingled. Finn et al. (2016a) stated this most clearly: "the IRAP is seen as providing a measure of the strength or probability of four functionally independent [relational responses]" (p.310). This claim is not always stated so explicitly elsewhere, but nonetheless it appears to be pervasive in guiding data scoring and analysis choices in the IRAP literature. Anecdotally, this claim is often encountered in peer review. For example, when authors present data from IRAPs scored as a single overall score, reviewers often respond that it should be scored and analysed by trial-type instead. It should be noted that although standardization of procedures is very important to replicability (Elson, 2019), that this suggestion is not merely related to standardization: it is typically attributed to the specific claim that the IRAP trial-types are independent and therefore averaging them is inappropriate. Indeed, this claim has theoretical implications: the separation of the IRAP into distinct trial-types has served as the basis for recent experimentation and theorising within Relational Frame Theory (e.g., Finn et al., 2016b, 2017, 2018). However, inspection of the literature demonstrates very little direct support for this claim, and one study directly contradicting it.

With regard to supportive evidence, studies that are sometimes cited to support it typically some form of the argument that one IRAP trial-types often demonstrates criterion associations and others do not (e.g., Hussey, Barnes-Holmes, et al., 2016; Nicholson & Barnes-Holmes, 2012). However, such a conclusion represents an instance of making claims about comparisons between two effects without directly comparing them. This represents a unfortunately common statistical fallacy that (i.e., the difference between "significant" and "non-significant" is not itself significant: Gelman & Stern, 2006; Nieuwenhuis et al., 2011).

Even if they did not rely on a statistical fallacy, these studies would represent indirect evidence at best, because they appeal to a likely consequence of the independent of the trial-types (i.e., differential criterion associations) rather than a direct assessment of their independence. Direct and statistically valid assessments are possible, and indeed already exist. Hussey, Ní Mhaoileoin, et al. (2016) demonstrated stimulus control over the stimulus function governing responses on a given IRAP trial-type (e.g., the 'women – human' trial-type in a dehumanization of women IRAP) by manipulating the content of the trial-type it was being contrasted with between IRAPs (i.e., by contrasting it with 'men' vs. 'inanimate objects'). Despite the content of the women trial-types being identical between IRAPs, women were evaluated to be more human when the other trial-types referred to inanimate objects and less human when they referred to men. This provided experimental evidence that the IRAP trial-types are not independent of one another. However, it is also just one study. It is prudent to directly evaluate the independence of the trial-types in other, larger studies, across multiple domains. That is the purpose of the current study.

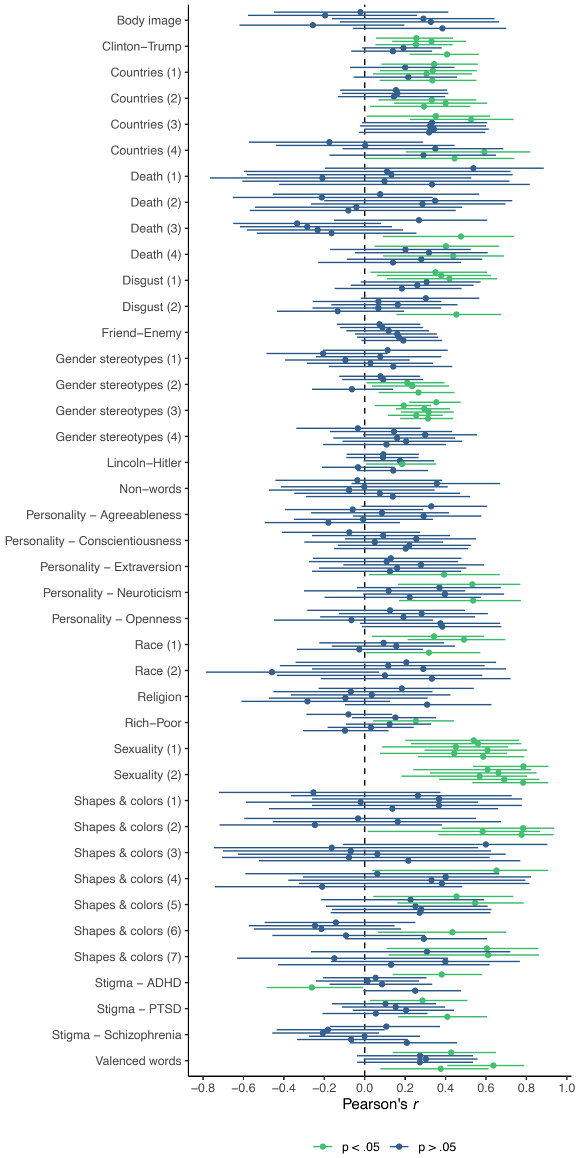
The claim that the IRAP trial-types are independent provides a precise and testable statistical claim: statistical independence requires the trial-types not be correlated. This claim was investigated by examining average correlations across a wide range of IRAP domains using meta-analytic models applied to a large existing dataset.

Method

Data

Data was taken from a publicly available dataset of published and unpublished IRAP data. All details of the component studies and task parameters are available in Hussey and Drake (2020). After excluding outliers based on mastery criteria in the task, the analytic sample contained 1688 participants who completed one of 41 distinct IRAPs in 17 different attitude domains. Data from each IRAP was converted to IRAP D scores (see Hussey et al., 2015). Only participants who passed the test blocks and maintained performance criteria on the test blocks were included (i.e., $\geqslant$ 78% accuracy and median reaction time $\leqslant$ 2000ms on both consistent and inconsistent blocks). For full details of the dataset, see Hussey and Drake (2020). All data and code is available (osf.io/tgajb).

Figure 1. Correlations among IRAP trial-types



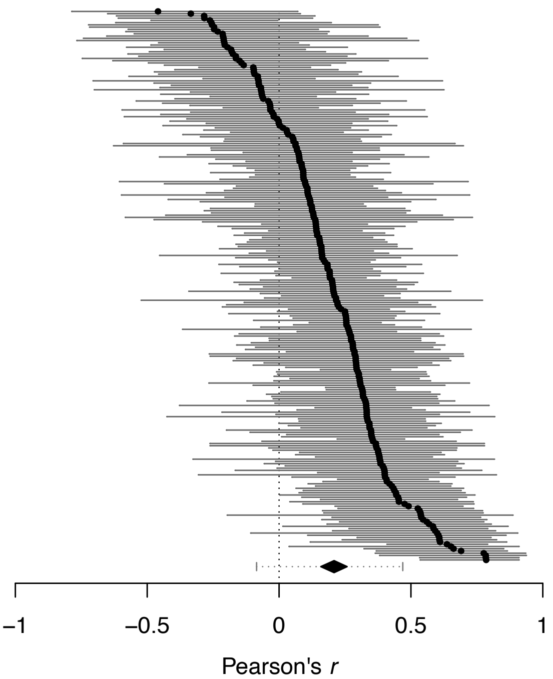
Results

Correlations among IRAP trial-types

Pairwise associations between scores on each IRAP trial-type within each domain were quantified using Pearson's r correlations, one for each permutation of the four trial-types, resulting in six correlations for each domain (i.e., between trial-types 1 & 2, 1 & 3, 1 & 4, 2 & 3, 2 & 4, and 3 & 4). All correlations were transformed prior to analysis using Fischer's r-to-z transformation. This normalization transformations produces less biased estimates of variance and allows confidence intervals to conform to the bounded range of correlations (i.e., 0 to 1). Sample sizes were then used to calculate variances and confidence intervals around each estimate using the R package metafor (Viechtbauer, 2010). Estimates and confidence intervals were back-transformed for reporting and plotting. Correlations are illustrated in Figure 1 (intervals represent 95% Confidence Intervals).

Inspection of the correlates and their 95% Confidence Intervals demonstrated that 27.2% of correlations were significantly different from zero. All detectably non-zero correlations were positive (see Figure 1). 63.4% of IRAPs demonstrated at least one pair of detectable correlations among its trial-types. This represents initial evidence that the IRAP trial-types are not in general independent: data suggest they often correlate. However, given large differences in sample sizes between studies and therefore differences in implied power (see Hussey, 2023), rather than merely quantifying the proportion of detectable correlations, this question should more appropriately be answered using a meta-analytic model. This is done in the section below.

Figure 2. Caterpillar plot. Diamond represents the 95% Confidence Interval on the meta-estimate. The dashed interval represents the 95% Prediction Interval.

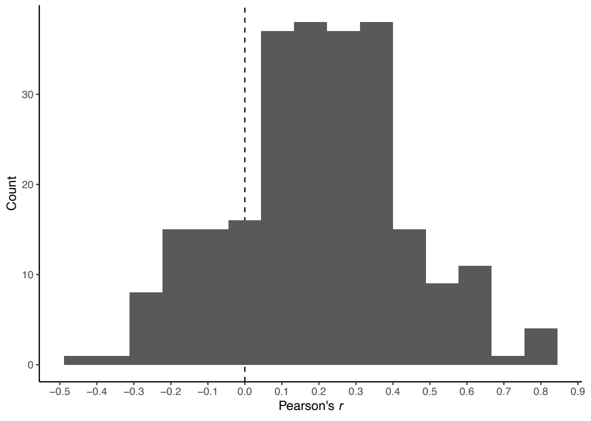


Meta-analysis

A three-level meta-analytic model was fit to these correlations using the metafor package. Domain was included as a random effect in order to acknowledge the non-independence of the six correlations for each domain, and to acknowledge the non-exhaustive nature of the domain variable (e.g., in order to increase generalizability of the results to other unobserved domains). Results demonstrated that correlations among IRAP trial-types are on average positively correlated, r = .21, 95% CI [.16, .25], 95% PI [-.09, .47], p < .00000000000001 with observed heterogeneity, Q(df = 245) = 399.03, $p < .0001$, $\tau^2 = 0$. This provides direct evidence against the claim that IRAP trial-types are in general independent. A caterpillar plot of results is presented in Figure 2.

95% Prediction Intervals (95% PIs) were calculated as well as Confidence Intervals. These take the observed heterogeneity of effect sizes into account in order to estimate the probably range of effect sizes that are likely to be observed (whereas 95% CIs are an estimate of the true population value). The width of the 95% Prediction Intervals (-.09, .47) suggested that IRAP studies are likely to observe correlations between trial-types that are anywhere from very small negative correlations (possibly indicating non-independence) to zero (possibly indicating independence at times) to moderate positive correlations (again possibly indicating non-independence). In light of this heterogeneity, in order to test the idea that some IRAP trial-types are independent and others are not (e.g., between domains), the distribution of correlations was inspected. No evidence of multimodality was observed (see histogram in Figure 3), suggesting that there are not two subsets of dependent versus independent trial-types. Unimodality of effect sizes combined with a statistically significant positive meta-effect size suggests that the IRAP trial-types are non-independent, and that observed variation is due to heterogeneity between domains from unmodelled sources rather than due to moderation. Simply put, results suggest that different implementations of the IRAP differ and this causes differences in the degree of correlation among the trial-types, but that the underlying relationship between trial-types is one of non-independence.

Figure 3. Distribution of correlations among IRAP trial-types



Discussion

Previous research has claimed that the IRAP's four trial-types are independent, and that IRAP data should therefore be scored and analyzed at the level of the four trial-types rather than one overall score (Finn et al., 2016a). However, the only direct experimental evidence test of this claim demonstrated otherwise, that performance on the IRAP trial-types are dependent on one another (Hussey, Ní Mhaoileoin, et al., 2016). In order to provide an equally direct yet more general test of this claim, I meta-analysed correlations among the four IRAP trial-types across 41 IRAPs in 17 domains. 27% of correlations were significantly different from zero. 63% of IRAPs demonstrated at least one detectable correlation among its trial-types. The meta-analysis demonstrated that IRAP trial-types are correlated (*r* = .21). The observed heterogeneity between domains combined with their unimodal distribution implies that the IRAP trial-types are not independent, although variation

Interpreting the magnitude of correlations

The magnitude of the meta-analytic correlation is worth considering. By Cohen's (1988) commonly employed interpretation guidelines this represents a "small" correlation. However, Cohen (1988) was explicit that researchers should also evaluate effect sizes in normative terms. Multiple large scale meta-scientific studies have estimated the average correlation observed in psychological studies to be around *r* = .20 (e.g., Gignac & Szodorai, 2016; Hemphill, 2003; Richard et al., 2003). The correlation between IRAP trial-types is therefore of similar magnitude to correlations observed in psychological research generally, and should therefore cannot be dismissed based on their magnitude. Comparisons can also be drawn within the field of implicit social cognition specifically, by comparing the magnitude of the correlations between the IRAP trial-types to that of correlations employed to argue that implicit measures have convergent validity. A meta-analysis of the association between the Implicit Association Test (IAT) and socially relevant criterion variables demonstrated an average effect of *r* = .23 (Kurdi et al., 2019). Elsewhere, the average correlation between implicit measures (IATs) and explicit measures (self-report scales) to the same attitude domain was found to be *r* = .36 (Nosek, 2005). If correlations with implicit measures of this magnitude are interpreted as evidence of convergent validity, then correlations of similar magnitude must also be accepted as evidence that the IRAPs trial-types are not independent.

Recommendations for the analysis of IRAP data

The evidence does not support the claim that IRAP's trial-types are in general independent. Blanket statements about whether IRAP data should be analyzed as one overall score or four trial-types scores should be therefore be avoided. Where researchers encounter recommendations to analyze IRAP data as four scores due to the independence of the trial-types, these recommendations should be countered with reference to empirical data showing this is not the case (e.g., the current article and Hussey, Ní Mhaoileoin, et al., 2016).

Decisions about scoring methods should also be made with reference to other measurement properties including reliability. Reliability should be expected *a priori* to be lower when scoring four the trial-types separately than one overall score given that reliability is directly related to the number of data points used to calculate it (van Zyl et al., 2000). This has been shown to be the case in the same dataset employed in the current article (Hussey & Drake, 2020).

Equally, the current results are not a panacea, nor do they imply that IRAP data should always be analyzed as one overall score and never as separate scores – only that blanket recommendations to do so are at odds with the data. There was observable variation between domains which may represents cases where one approach is more appropriate than the other. However, determinations about how to analyze data from specific IRAPs (i.e., data generated from specific stimuli and task parameters) should be made with reference to larger studies focusing on addressing such measurement questions.

While researchers are encouraged to base their analytic decisions on data, they should be acutely aware of the risks of overfitting. For example, based on the current research, a researcher might decide whether to analyze their IRAP data as one score or four based on an assessment of the correlations among the trial-types (e.g., one score if correlated, four scores if not). However, this risks over fitting, or conditioning analyses on results obtained within the same sample. This represents a second reason why larger studies focused specifically on measurement are warranted prior to using data from a given IRAP to make substantive claims. That is, the movement from measure development and refinement to substantive use should follow the well-established practices and norms of the development of any measure (e.g., Nunnally, 1978; Nunnally & Bernstein, 1994).

Equally, the current results should not be interpreted as carte blanche to score IRAP data however researchers please, and especially not as an invitation to score it multiple ways and report only a subset of those results. Researchers already face a large number of choices when scoring IRAP data (e.g., by trial-type or overall score, or collapsing the sample trial-types, or collapsing the target trial-types, the choice to invert trial-types or not, choices between multiple scoring algorithms, and between multiple performance criteria: Hussey et al., 2015). Each decision point represents an Experimenter Degree of Freedom, which are a key contributor to poor replicability (Simmons et al., 2011). The published IRAP literature has many examples of the same researchers making different decisions between studies, increasing the risk that that these represent analytic choices that are overfitted on the data at hand. Two strategies that would improve the replicability of findings in light of a larger number of Experimenter Degrees of Freedom are preregistration of those choices and their rationales (Nosek et al., 2018; although see: Akker et al., 2022; Bakker et al., 2020) and/or results-blind analyses (MacCoun & Perlmutter, 2015; Sarafoglou et al., 2023).

Author note

Ian Hussey, University of Bern, Switzerland. ian.hussey@unibe.ch.

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 osf.io/tgajb.

References

Akker, O. van den, Assen, M. A. L. M. van, Enting, M., Jonge, M. de, Ong, H. H., Rüffer, F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J., & Bakker, M. (2022). *Selective Hypothesis Reporting in Psychology: Comparing Preregistrations and Corresponding Publications*. MetaArXiv. https://doi.org/10.31222/osf.io/nf6mq

Bakker, M., Veldkamp, C. L. S., Assen, M. A. L. M. van, Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

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

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

Elson, M. (2019). Examining Psychological Science Through Systematic Meta-Method Analysis: A Call for Research. *Advances in Methods and Practices in Psychological Science*, *2*(4), 350–363. https://doi.org/10.1177/2515245919863296

Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016a). 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

Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016b). Exploring the behavioral dynamics of the Implicit Relational Assessment Procedure: The impact of three types of introductory rules. *The Psychological Record*, *2*, 309–321.

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2017). Exploring the Single-Trial-Type-Dominance-Effect in the IRAP: Developing a Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) Model. *The Psychological Record*, 1–15. https://doi.org/10.1007/s40732-017-0262-z

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect in the IRAP: Developing a differential arbitrarily applicable relational responding effects (DAARRE) model. *The Psychological Record*, *68*(1), 11–25. https://doi.org/10.1007/s40732-017-0262-z

Gelman, A., & Stern, H. (2006). The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

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., 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

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

Hussey, I. (2023). *A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using the Implicit Relational Assessment Procedure*. PsyArXiv. https://doi.org/10.31234/osf.io/g2x9p

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. (2020). The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/ge3k7

Hussey, I., Ní Mhaoileoin, D., Barnes-Holmes, D., Ohtsuki, T., Kishita, N., Hughes, S., & Murphy, C. (2016). The IRAP Is Nonrelative but not Acontextual: Changes to the Contrast Category Influence Men’s Dehumanization of Women. *The Psychological Record*, *66*(2), 291–299. https://doi.org/10.1007/s40732-016-0171-6

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

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature News*, *526*(7572), 187. https://doi.org/10.1038/526187a

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

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. https://doi.org/10.1038/nn.2886

Nosek, B. A. (2005). Moderators of the Relationship Between Implicit and Explicit Evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565

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

Nunnally, J. (1978). An Overview of Psychological Measurement. In B. B. Wolman (Ed.), *Clinical Diagnosis of Mental Disorders: A Handbook* (pp. 97–146). Springer US. https://doi.org/10.1007/978-1-4684-2490-4_4

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

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

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2023). Comparing Analysis Blinding With Preregistration in the Many-Analysts Religion Project. *Advances in Methods and Practices in Psychological Science*, *6*(1), 25152459221128319. https://doi.org/10.1177/25152459221128319

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

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

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*(3), 271–280. https://doi.org/10.1007/BF02296146

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03