The Implicit Relational Assessment Procedure’s trial-types

are not independent

Ian Hussey

I have frequently encountered the claim, whether in published literature or in peer review, that data from the Implicit Relational Assessment Procedure must be analyzed at the trial type level because the IRAP trial types are functionally independent”. Inspection of the literature shows no direct evidence to support this claim. In order to investigate this directly, I employed a large existing dataset of 1464 participants who completed one of 35 IRAPs in 16 domains. Scores for each trial type within each IRAP were correlated with one another. 27% of correlations were significantly different from zero. 74% of IRAPs demonstrated at least one detectable correlation among its trial types. A meta-analysis demonstrated that IRAP trial types are typically correlated, albeit with significant heterogeneity between domains, *r* = .21, 95% CI [.16, .26], 95% PI [-.10, .48]. 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. Equally, this should not be interpreted as carte blanche to score IRAP data however we please. Scoring choices should be made with reference to evidence for their relative reliability and validity. The replicability of claims in IRAP studies could be increased via preregistration of researchers’ chosen scoring methods.

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. (2016b) stated this 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 often always not 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., 2016a, 2017, 2018).

However, inspection of the literature demonstrates very little direct support for this claim. 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 et al., 2016; Nicholson & Barnes-Holmes, 2012). However, this represents a common statistical fallacy (i.e., the difference between “significant” and “non-significant” is not itself significant: Gelman & Stern, 2006). To the best of my knowledge, no published IRAP study among those listed in the 151 publications found by a recent systematic review (Hussey, 2023) has conducted a direct assessment of this claim by assessing the average correlations among IRAP trial types.

The claim that the IRAP trial types are independent provides a precise one precise and testable statistical claim: statistical independence requires the trial types not be detectably correlated. This claim was investigated by examining average correlations across a wide range of IRAP domains.

# Method

## Data

Data was taken from a publicly available dataset of published and unpublished IRAP data (Hussey & Drake, 2020). After excluding outliers based on mastery criteria in the task, the analytic sample contained 1464 participants who completed one of 35 distinct IRAPs in 16 different attitude domains. Data from each IRAP was converted to IRAP *D* scores (see Hussey et al., 2015). For full details of the dataset, see Hussey and Drake (2020). All data and code is available (osf.io/XXX).

# 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 (i.e., one for each permutation of the four trial types). Variances and confidence intervals were calculated around each estimate using the R package metafor (Viechtbauer, 2010). All correlations are illustrated in Figure 1 (intervals represent 95% Confidence Intervals).

Inspection of the 95% Confidence Intervals demonstrated that 27.1% of correlations were significantly different from zero. All detectably non-zero correlations were positive (see Figure 1). 74.3% of IRAPs demonstrated at least one pair of detectable correlations among its trial types. Even descriptively, this observation represents initial evidence that the IRAP trial types are not independent: data suggest they often correlate.

## Meta-analysis

A three-level meta-analytic model was fit to these correlations, again 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, .26], 95% PI [-.10, .48], *p* < .000000000001. 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.

**Figure 1.** Correlations among IRAP trial types



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 (-.10, .48) 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), I inspected the distribution of correlations. 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 2.** Caterpillar plot. Diamond represents the 95% Confidence Interval on the meta-estimate. The dashed interval represents the 95% Prediction Interval.



**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., 2016b).

The claim that the IRAP trial types is frequently repeated in peer review to argue that IRAP data should be analyzed at the trial type level

Of course, its possible that some might argue that this statistical claim of near zero correlation among trial types is not what they meant by “functional independence”. If so, I encourage such authors to translate their broad verbal claims into testable predictions, or remain unsupported. Scheel (2022) highlights that many such verbal claims are “not even wrong” in the sense that they are so vague as to be unsupportable, untestable, and incapable of being correct or wrong. Claims not supported by evidence should not rigidly dictate our research practices.

[Something about how this doesn’t mean it’s a free for all either. Pre-registration is even more needed when there are experimenter degrees of freedom like this. Larger studies on measurement properties needed.]

# 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 osf.io/XXX.

# References

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

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*, *2*, 309–321.

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*, *66*(2), 309–321. https://doi.org/10.1007/s40732-016-0173-4

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

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

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. *Preprint*. https://doi.org/10.31234/osf.io/ge3k7

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

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

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

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