Journal of Contextual Behavioral Science

The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess

--Manuscript Draft--

Manuscript Number:	
Article Type:	Empirical Research
Keywords:	implicit relational assessment procedure; confound; Validity
Corresponding Author:	Ian Hussey Ruhr University Bochum GERMANY
First Author:	Ian Hussey
Order of Authors:	Ian Hussey
	Chad E Drake, PhD
Abstract:	Several recent articles have reached the same conclusion that effects on the Implicit Relational Assessment Procedure (IRAP) are biased in some way or demonstrate generic patterns of effect regardless of what domain is being assessed. Multiple accounts have been advanced to explain why this might be the case. However, no work has sought to either (a) precisely estimate this generic effect or (b) consider its implications for the validity of conclusions in published and future research. This study used a large open dataset (N = 753) of IRAPs capturing implicit evaluations in multiple domains. Results demonstrated a specific generic pattern among IRAP effects that was common across domains. The majority of variance in IRAP effects is attributable to the generic pattern rather than the domain being assessed. The IRAP is therefore relatively insensitive to the attitudes or learning histories that it is intended to assess, and effects on the task are heavily confounded. The existence of the generic pattern may also undermine the validity of many conclusions made in the published IRAP literature.
Suggested Reviewers:	Brian O'Shea, PhD University of Amsterdam boshea@fas.harvard.edu Brian is an author of the first publication to describe the generic pattern on the IRAP. The current manuscript includes a replication of his study that a nonsense word IRAP demonstrating the pattern.
	Chad E Shenk, PhD Associate Professor, The Pennsylvania State University ces140@psu.edu Chad has conducted large scale IRAP studies, has experience with advanced statistical methods, and knowledge of R. Prior to submission, he contacted me to point out an error in the publicly available R code. He seems to already be aware of the project's data and code.
	Solomon Kurz, PhD Postdoc, Central Texas Veterans Healthcare System a.solomon.kurz@gmail.com Solomon has worked with IRAP data and has knowledge of advanced statistical methods
Opposed Reviewers:	Dermot Barnes-Holmes, PhD d.barnes-holmes@ulster.ac.uk Several of the studies included in this manuscript were conducted as part of my PhD under Dermot's supervision. Dermot therefore would not be independent/has a COI.
	Yvonne Barnes-Holmes, PhD yvonne.barnesholmes@ugent.be Several of the studies included in this manuscript were conducted as part of my PhD under Yvonne's co-supervision. Dermot therefore would not be independent/has a COI.

RUHR UNIVERSITY BOCHUM | 44780 Bochum | Germany

FACULTY OF PSYCHOLOGY

Psychology of
Human Technology Interaction

Dr. Ian Hussey
Phone +32 (0)470 396842
Email Ian.hussey@rub.de

July 29, 2022

Manuscript Submission: " The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess"

Dear Prof Levin,

Please find attached our manuscript "The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess".

The IRAP is of clear interest to CBS. The majority of publications that employ it examine draw conclusions from (differences in) mean *D* scores. At the same time, there is already agreement that these scores are biased in some way. E.g. O'Shea et al (2016) observed this and called it a 'positive framing bias'. Finn et al. (2016) also observed it and referred to it by different terms. While there is general agreement that this issue exists, our manuscript is the first to (1) correctly label this issue as a confound, (2) to quantify it precisely using a large sample and multiple domains, and (3) consider the implications of this confound for the valid interpretation of past and future IRAP research.

After demonstrating the effect in a large sample, we consider which statistical tests, comparisons and inferences are confounded vs. unaffected by the generic pattern. This provides guidance to how readers should and should not (re)interpret results from IRAP data.

We hope that this manuscript will be of great interest to your readers given CBS's close ties to the IRAP.

Kind regards,

Ian Hussey
&
Chad Drake

Highlights

- Effects on the Implicit Relational Assessment Procedure are seriously confounded
- A similar generic pattern is found regardless of the evaluative domain assessed
- Previously published research may have reached invalid conclusions

The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess

Ian Hussey & Chad E. Drake

Word count: 4883

Author note: Ian Hussey, Ruhr University Bochum, Germany. ian.hussey@rub.de. Chad Drake, SIU Carbondale, USA. chad.e.drake@gmail.com. IH was supported by Ghent University grant 01P05517 and the META-REP Priority Program of the German Research Foundation (#464488178).

Abstract

Several recent articles have reached the same conclusion that effects on the Implicit Relational Assessment Procedure (IRAP) are biased in some way or demonstrate generic patterns of effect regardless of what domain is being assessed. Multiple accounts have been advanced to explain why this might be the case. However, no work has sought to either (a) precisely estimate this generic effect or (b) consider its implications for the validity of conclusions in published and future research. This study used a large open dataset (N = 753) of IRAPs capturing implicit evaluations in multiple domains. Results demonstrated a specific generic pattern among IRAP effects that was common across domains. The majority of variance in IRAP effects is attributable to the generic pattern rather than the domain being assessed. The IRAP is therefore relatively insensitive to the attitudes or learning histories that it is intended to assess, and effects on the task are heavily confounded. The existence of the generic pattern may also undermine the validity of many conclusions made in the published IRAP literature.

The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess

Implicit measures have seen widespread use across many clinical and social domains over the past two decades and are now a mainstay of psychological measurement (Greenwald & Lai, 2020; Nosek et al., 2011). A wide variety of implicit measures have been created, with each procedure having unique features and benefits. In particular, the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2010) is acknowledged to be one of few implicit measures that allows researchers to assess the relations between stimuli of interest (Gawronski & De Houwer, 2011). That is, the IRAP can assess not only how automatically concepts and attributes are associated (e.g., self and negative) but also the manner in which they are related. For example, the distinction between "I am bad" versus "I want to be bad" (Remue et al., 2013, 2014). IRAP researchers have offered this distinction as a potential benefit of the procedure compared to associative measures (Gawronski & De Houwer, 2011; Hughes et al., 2012).

Generic patterns in IRAP data

However, the IRAP has also been subject to an important criticism: effects on the task appear to suffer from a 'positive framing bias'. Stimulus categories are often evaluated positively on the IRAP even when the participant would be expected to hold neutral or negative attitudes towards that category (O'Shea et al., 2016). For example, normative participants apparently demonstrate positive evaluations of both death (Hussey, Daly, et al., 2015) and Hitler (previously unpublished data reported in this article).

O'Shea et al. (2016) argued that this effect occurs due to the valence of the IRAP response options: 'True' is more positively valenced than 'False', and valence congruence between the response option and the valence of the attribute stimuli generate positive IRAP

effects. To put this another way, whereas the IRAP is intended to provide a measure of the automatic relating of the category and attribute stimuli (e.g., 'Black people' and 'Pleasant'), the effect may instead be driven by the congruence between the valence of the attribute stimuli and response option (e.g., 'Pleasant' and 'True' vs. 'False'). O'Shea et al. (2016) therefore advance two key ideas: (1) they argue that IRAP effects are driven in large part by some factor that is unrelated to the phenomenon that is of direct interest to researchers using the task, and (2) they advance a specific explanation of this, which will we refer to here as the 'valence congruence account'.

Subsequent research has agreed with the idea that IRAP effects are influenced by factors other than category-attribute relations but has provided alternative explanations of why this phenomenon occurs. Finn et al. (2016) employed an IRAP which involved relating non-evaluative stimuli (i.e., colors and shapes). Despite including no evaluative stimuli, a comparable bias was demonstrated, whereby effects on some trial types were larger than others. This would seem to suggest that O'Shea et al.'s (2016) valence congruence account is insufficient. Finn et al. (2016) advanced an alternative account of this effect, which they continued to develop in a subsequent publication (Finn et al., 2018). The key point to be appreciated here is that while O'Shea et al. (2016) and Finn et al. (2016, 2018) disagree as to the *cause* of this bias in IRAP effects, the presence, replicability, and generalizability of these biases in IRAP effects is apparently uncontroversial.

Goals of the current research

In contrast with previous work that has focused on explanations of these biases in IRAP effects, the current research seeks to (1) quantify this bias more precisely, and (2) consider its implications for the validity of the conclusions made in the published literature. We will hereafter refer to these biases as the 'generic pattern' observed in IRAP effects.

Previous debate about the nature of any generic pattern may have been driven by the fact that this pattern has not yet been well estimated, due to a combination of small sample sizes (typically around 40) and a limited range of domains. In order to overcome this, this article used an unprecedently large sample (N = 753) addressing multiple attitude domains (k = 10). This was achieved by collating data from published and unpublished IRAP studies conducted across two labs that undertook multi-year IRAP research programs. This work aimed to (1) assess the evidence that IRAP effects tend to follow a generic pattern by estimating the generic pattern more precisely; (2) understand the severity of the generic pattern by quantifying the proportion of variance in IRAP effects that comes from undesirable sources (i.e., the generic pattern) versus desirable sources (i.e., sensitivity to the domain being assessed); and (3) make recommendations about which common analytic strategies give rise to valid versus invalid inferences as a result of this generic pattern.

Method

All data and code for data processing and analysis code is available on the Open Science Framework (https://osf.io/vhzsn/?view_only=60e9e24c080e410db9c929914cf7eec4). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons et al., 2012). All analyses were done in R (R Core Team, 2022) using the packages ez (Lawrence, 2016), schoRsch (Pfister & Janczyk, 2019), and lme4 (Bates et al., 2015).

Data

Data was taken from an existing, publicly available dataset of IRAP data [reference removed for peer review blinding on editor's instruction]. The current study therefore employs secondary analysis of existing data, with sample size being determined by data availability. Inclusion criteria were as follows: (1) The study must have included at least one standard IRAP (i.e., not variants such as the MT-IRAP or Training IRAP); (2) The IRAP must employ single-

word valenced attribute category stimuli (e.g., positive vs. negative). This did not include other more specific categorizations (e.g., masculine/feminine) or more elaborate propositions (e.g., 'I can approach' vs. 'I cannot tolerate it'). This served to limit the differences between IRAPs to the domain being evaluated while keeping other aspects of the procedure relatively consistent; (3) The IRAP must have used 'True' and 'False' as response options within the procedure; (4) When a study employed multiple IRAPs within participants, only the first IRAP that each participant completed was included. Data from 12 IRAPs across 10 domains using 12 stimuli sets and a total of 753 participants met inclusion criteria. See Figure 2 for a list of all domains.

Performance exclusions

Participants whose percentage accuracy or mean reaction time on the IRAP test blocks were more than 2.5 standard deviations from the mean were excluded as outliers. These exclusions were calculated separately for each domain to allow for differences in the distributions of mean reaction times. This method was adaptive to differential mean response latencies between domains, removed the need for an arbitrary cutoff, and is consistent with recommendations for the treatment of outliers in the wider reaction-time literature (Ratcliff, 1993; Whelan, 2008). A total of 44 participants (5.8%) were excluded on this basis.

Participants

Ethical approval for each original study was granted by the local institutional review board, and informed consent was obtained from all individuals prior to participation. The final analytic sample after performance exclusions contained 709 participants. Where demographics data was present, participants were typically female (193 women [62.5%], 159 men [37.2%], 1 identified as nonbinary [0.3%]) and young adults (M_{age} = 20.1, SD = 4.7). Sample size by attitude domain ranged from 19 to 131 (M = 70.9, SD = 38.8).

Measures

The IRAP is a computer-based reaction time task. Its procedural parameters have been discussed in great detail in many other papers (Barnes-Holmes et al., 2010; Hussey, Thompson, et al., 2015), and so only a brief overview will be provided here (see Hussey, 2020). On each block of trials, participants are presented with images or words at the top of the screen and in the middle of the screen. Response options are presented on the bottom left and bottom right sides of the screen and are mapped to the left and right response keys. In order to progress to the next trial, the correct response must be given. Incorrect responses result in a red X being presented on screen. The correct responses alternated between blocks. For example, an IRAP study examining racial attitudes might include "White people" and "Black people" as category stimuli and positive and negative words as attribute stimuli. In this example, a participant must respond to "White people" and "Dangerous" with "True" on one block and "False" on the other block. Participants initially complete pairs of these blocks during a practice phase of the task; once responding meet predetermined criteria for a both of a pair of practice blocks (typically a median reaction time < 2000 ms and a percentage accuracy > 80%), participants proceed to the test block (typically three pairs). Only data from the test blocks is used in the analyses (Hussey, Thompson, et al., 2015). Details of each IRAP's stimuli, task parameters, and responding rules can be found in the Supplementary Materials of the original dataset (https://osf.io/vhzsn/?view_only=60e9e24c080e410db9c929914cf7eec4).

Data processing

IRAP studies typically use the *D* scoring method to convert each participant's reaction times into analyzable values. The *D* score has some similarities to Cohen's *d*, insofar as it is a standardized difference in mean reaction time between the two block types. The specifics of the *D* score have been discussed in precise detail in other publications (Barnes-Holmes et al., 2010; Hussey, Thompson, et al., 2015) and therefore will only be summarized here. Its key

points are that reaction times > 10,000 ms are trimmed, a mean reaction time is calculated for the trials in each block type, and a standard deviation is calculated for the pooled trials in both blocks. The difference between the means is then divided by the standard deviation, resulting in a *D* score.

Four *D* scores were calculated for each IRAP, one for each of the four trial-types (e.g., 'life – positive', 'life – negative', 'death – positive', 'death – negative'). Data for each study was scored so that positive *D* scores referred to faster responding on the blocks assumed to be consistent with participants' learning histories, as in the original studies. For example, positive *D* scores on trial-type 1 on the body-shape IRAP referred to faster responding to the stimulus pair 'thin – positive' with 'True' relative to 'False'; comparably, positive *D* scores on trial-type 1 on the Christian-Muslim IRAP referred to faster responding to 'Christians – Safe' with 'True' relative to 'False'.

Results

Evidence for the generic pattern

We hypothesized that if the IRAP is relatively sensitive to the domain being assessed then a greater proportion of variance will be attributable to the main effect for IRAP stimuli domain and/or the interaction between domain and trial type than for the main effect for trial type. However, if IRAP effects are mostly driven by the generic pattern then the main effect for trial type effect would be larger than the main effect for domain or their interaction effect. This latter scenario would imply that the IRAP is relatively insensitive to the stimulus domain being assessed.

A mixed within-between ANOVA was run using type III sum of squares method with IRAP *D* scores as the dependent variable, IRAP trial-type as the within subjects independent variable (i.e., Category 1 – Positive, Category 1 – Negative, Category 2 – Positive, Category 2 – Negative), and domain as the between subjects independent variable. Only data from the

domains featuring known-words was used (i.e., all domains other than the non-words IRAP). Because our hypothesis involved comparing the proportion of variance attributable to the effects effect rather than assessing differences in means, only the ANOVA effect sizes are reported. Three different metrics of effect size are reported: η^2, η_p^2, and η_g^2, where η^2 refers to the percentage of variance explained (akin to r^2); η_p^2 ('partial') refers to the percentage of variance explained after controlling for all other main and interaction effects; and η_g^2 ('generalized') also refers to percentage of variance after controlling for other effects, but is corrected to be interpretable across differential factorial designs (e.g., number of levels and within vs. between factors; see Lakens, 2013). All three metrics are reported here to maximize informativeness. Confidence intervals were bootstrapped using the case removal and percentile method with 2000 replications. Point estimates for effect sizes are computed as median bootstrapped estimates for robustness. Following standard practice, 90% confidence intervals are reported rather than 95% confidence intervals on the basis that squared effect sizes can only be positive numbers.

Hypothesis tests were conducted via the comparison of confidence intervals. Given the mix of within and between-subject effects, η_g^2 was the most appropriate effect size to compare in order to make inferences. Indeed the generalized form of this effect size was created for exactly such purposes. Results demonstrated that the majority of variance in IRAP effects was explained by the main effect for trial type (η^2 = 0.70, 90% CI [0.66, 0.75], η_p^2 = 0.24, 90% CI [0.22, 0.27], η_g^2 = 0.17, 90% CI [0.15, 0.19]) with only a small fraction explained by either the main effect for domain (η^2 = 0.10, 90% CI [0.07, 0.14], η_p^2 = 0.08, 90% CI [0.06, 0.10], η_g^2 = 0.03, 90% CI [0.02, 0.04]) or the interaction between domain and trial type (η^2 = 0.09, 90% CI [0.07, 0.12], η_p^2 = 0.04, 90% CI [0.03, 0.05], η_g^2 = 0.03, 90% CI 90% CI [0.02, 0.04]). Results are illustrated in Figure 1. Results therefore supported the conclusion that variation in the IRAP

effects are mostly attributable to a generic pattern among the IRAP trial types. IRAP effects are therefore relatively insensitive to the attitude domain being assessed.

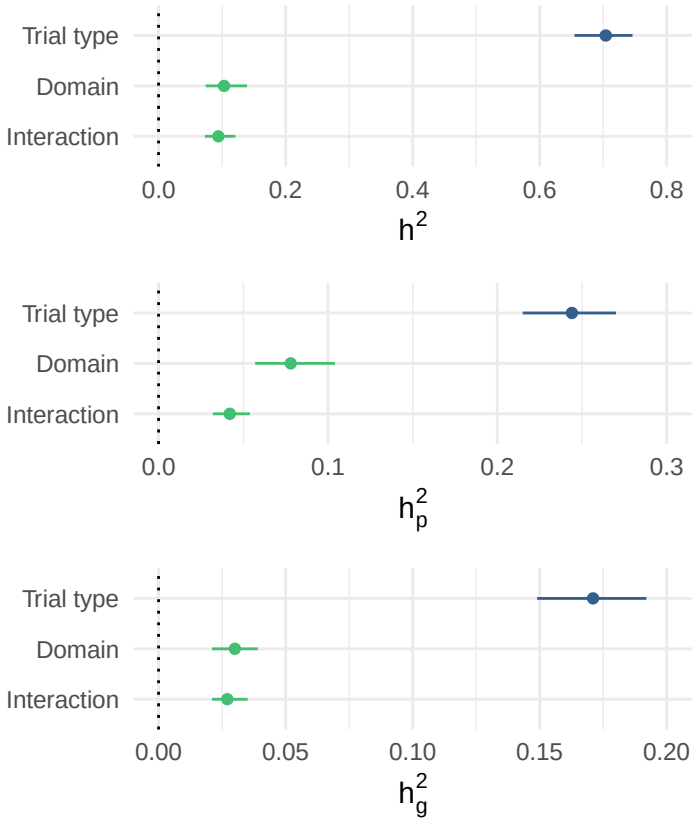

Figure 1. Effect size estimates for the ANOVAs.

Estimating the generic pattern

A meta-analytic model was then used to estimate the generic pattern. Specifically, a crossed random effects model with *D* scores as the dependent variable, trial type as independent variable, and both domain and participant as random effects (i.e., random intercepts). The Wilkinson notation of this model was as follows:

```
D ~ 1 + trialtype + (1 | domain) + (1 | participant)
```

The generic pattern was therefore estimated via the estimated means for each trial type. Estimated means were Category 1 – Positive: M = 0.34, 95% CI = [0.30, 0.38], p < .001; Category 1 – Negative: M = 0.14, 95% CI = [0.06, 0.21], p < .001; Category 2 – Positive: M = -0.20, 95% CI = [-0.28, -0.12], p < .001; Category 2 – Negative: M = 0.02, 95% CI = [-0.06, 0.09], p < .001. This pattern is illustrated in Figure 2 (upper panel), along with the data from each attitude domain.

Figure 2 (lower panel) also suggests that the generic pattern is present not only in IRAPs assessing known attitude domains, but possibly also in an IRAP assessing evaluations of the non-words CUG and VEC (which should intuitively be neutral). Due to the very small sample size for this IRAP (N = 19), no meaningful quantitative analyses could be conducted to compare known words (i.e., all data from the attitude domains analyzed previously) and non-word stimuli. Nonetheless, visual inspection of the plot reveals a strikingly similar pattern between the trial types, despite one set of IRAPs supposedly measuring attitudes to a set of domains, and the other employing nonsense words.

In order to facilitate the understanding of this generic pattern, the Category 2 trial types were inverted following standard guidelines for the interpretation of IRAP effects (Hussey, Thompson, et al., 2015). This provided a common interpretation across trial types: positive D scores represent more positive evaluations and negative D scores represent more negative evaluations (i.e., quicker responding to positive attribute stimuli with ‘True’, or negative attribute stimuli with ‘False’). These inverted D scores are illustrated in Figure 3. The generic pattern therefore implies that, regardless of what attitude domains served as Category 1 and Category 2, participants evaluative Category 1 more positively than Category 2, and affirm positively more than they reject negativity. As can be seen from the estimated means, the ordinal ranking among the trial types is Category 1 – Positive > Category 1 – Negative > Category 2 – Positive > Category 2 – Negative.

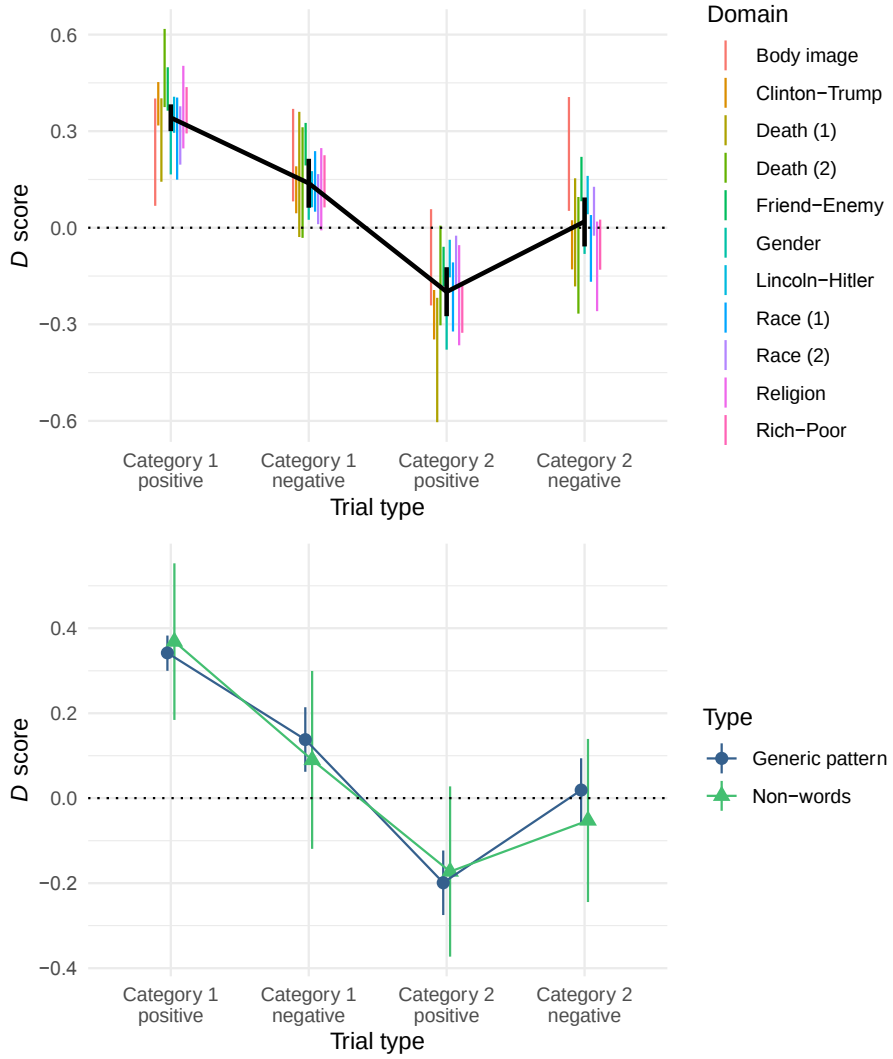

Figure 2. The generic pattern in IRAP effects. Upper panel illustrates mean IRAP effects for each attitude domain, with the meta-analyzed generic pattern in black. Lower panel compares effects on a non-words IRAP with the generic pattern. Error bars represent 95% CIs.

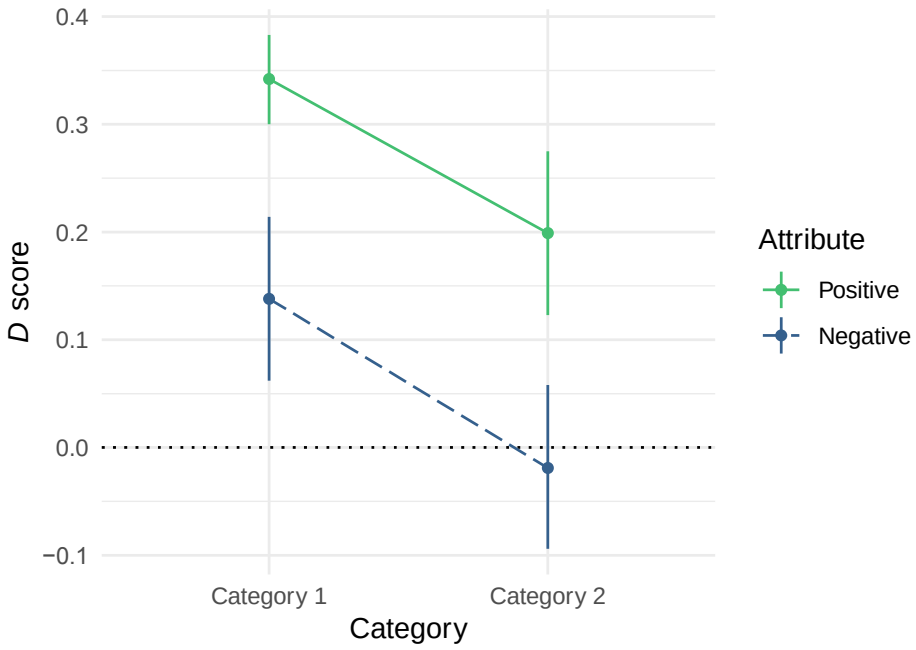

Figure 3. Meta-analyzed estimates of the generic pattern in IRAP effects. Scores for Category 2 have been inverted for interpretability. Error bars represent 95% CIs.

Discussion

Results demonstrate that IRAPs assessing implicit evaluations demonstrate a generic pattern among trial types that is unrelated to the domain supposedly being assessed. Variation in IRAP effects is attributable in large part to this generic pattern, much more so than the category stimuli employed in the procedure. This suggests that the IRAP is relatively insensitive to the attitudes and learning histories that it was designed to measure.

Results from our meta-analytic model provided insight into the nature of the generic pattern (see Figures 2 and 3). This appears to take the form of a specific ordinal ranking in mean scores between the four trial types. Our results, based on a far larger sample size and range of domains, suggest that both existing accounts of the generic pattern are incorrect: the pattern is not either a ‘positivity bias’ (O’Shea et al., 2016) or a ‘single trial type dominance effect’ (Finn et al., 2018), but instead represents both differences between positive and negative

attributes and also differences between the two categories. This pattern is difficult to attribute to genuine properties of the attitude domains themselves, and is more easily attributed to as-yet-known features of the IRAP task itself that elicit such behaviour within the task. The generic pattern among IRAP effects seems to be a replicable effect, but undermines the IRAP's utility in assessing the implicit attitudes or learning histories that most researchers are interested in when using the task. Put another way, the analysis of mean scores on the IRAP trial types is therefore severely confounded by the generic pattern.

In behavioural terms, the category stimuli appear to exert relatively weak stimulus control over reaction times relative to other, likely less interesting, sources of control. However, it should be noted that these negative implications for the IRAP are agnostic to the level of analysis used by a researcher, whether represenationalist (e.g., that IRAP effects can be used to measures implicit attitudes or associations in memory, etc.) or functional analytic-abstractive (e.g., in terms of relational responding or other concepts developed within Relational Frame Theory; see Barnes-Holmes & Hussey, 2016; Hughes et al., 2011, 2012).

Implications for the validity of conclusions in the published literature

The presence of this generic pattern is problematic for most research using the IRAP. Generally speaking, when researchers use the IRAP in their research, they wish to use the task to help explain another phenomenon of interest (i.e., behaviour within the IRAP functions as the thing that explains: the *explanans*) rather than in order to investigate behaviour within the IRAP itself (i.e., where behaviour within the IRAP represents the thing to be explained: the *explanandum*), although exceptions do exist (Finn et al., 2016, 2018; O'Shea et al., 2016). In the IRAP's modal use-case, the presence of a generic pattern is likely to represent a strong barrier to the task being useful to their goals. And, more worryingly, the generic pattern may cause researchers to make invalid inferences, by misattributing the presence of IRAP effects to attitudes or learning histories (i.e., driven by the category stimuli) when they are instead merely

instances of the generic pattern. For example, Hussey, Daly, et al. (2015) concluded that normative participants demonstrated counter-intuitive positive evaluations of death. In light of the generic pattern among IRAP effects, it would be more accurate – and less interesting – to characterize these results merely as 'the generic pattern among IRAP effects was observed', with no reference to what this might say about the original domain of interest.

The existence of a generic pattern has significant implications for how the results of past and future IRAP studies should be interpreted. Indeed, many of the conclusions made in the published literature may be undermined or invalidated due to the confound that the generic pattern represents. To understand why this is the case, consider that, by definition, the generic pattern means that non-zero IRAP effects are likely to be observed regardless of whether participants possess attitudes or learning histories that would previously be expected to be the source of such IRAP effects. As such, the presence of IRAP effects – that is, *D* scores that are significantly different from zero – cannot reasonably be equated with evidence for implicit attitudes (i.e., at the cognitive level of analysis) or learning histories involving the category stimuli (i.e., at the behavioural analytic-abstractive level). Analyses that treat *D* = zero as a reference point, even tacitly, will therefore produce false or invalid domain level inferences. Somewhat unfortunately, this concern about the validity of conclusions when a *D* score at or near zero is treated as meaningful have been around as long as the IRAP itself (Blanton & Jaccard, 2006). However, these have previously been conceptual arguments, where the current work is empirical.

Table 1. A description of commonly-used methods of analysis for IRAP data, as well as the validity of the inferences which are typically made from them.

Comparison	Example research question	Analytic method	Common inference	Conclusions
Mean *D* scores from a single trial-type compared against 0	Is the "White people – positive" *D* score significantly different from zero?	One-sample *t*-test	A "White people – positive" bias was observed.	Confounded
Mean *D* scores within-subject compared between-trial types	Does participant 1's "White people – positive" *D* score differ from their "White people – negative" *D* score?	Within-subjects *t*-test/ANOVA	"White people – positive" biases were larger than "White people – negative" biases.	Confounded
Mean *D* scores from a given trial-type and participant compared between time points	Do *D* scores on the "white people – positive" differ between timepoints 1 and 2?	Within-subjects *t*-test/ANOVA	"White people – positive" bias changed between timepoints/after the intervention.	Unaffected
Mean *D* scores from a given trial type compared between-subjects	Do effects on the "White people – positive" *D* score differ between Black and White participants?	Between-subjects *t*-test/ANOVA	White people demonstrated a larger "White people – positive" bias than Black people.	Unaffected
D scores from a given trial type correlated with other trial-types	Are "White people – positive" *D* scores negatively associated with "White people – negative" *D* scores?	Correlation/regression	Positive evaluations of White people are negatively associated with negative evaluations of White people.	Unaffected
D scores from a given trial type correlated with external variables	Are "Black people – negative" *D* scores positively associated with self-reported racism?	Correlation/regression	Negative evaluations of Black people on the IRAP and in a self-reported racism scale are positively associated.	Unaffected

Notes: Conclusions refers to the validity of substantive domain-level conclusions in light of the existence of the generic pattern among IRAP effects.

In order to explicate which specific types of analyses and conclusions are impacted by the presence of the generic pattern among IRAP effects, we discuss each of the most common forms of analysis of IRAP data in turn. Table 1 provides a summary of the most common comparisons, research questions, analytic methods, and inferences from IRAP data, and the validity of such conclusions in light of the generic pattern. Figure 4 uses results from a hypothetical between-groups IRAP study to illustrates some of these common comparisons and the validity of their domain level conclusions.

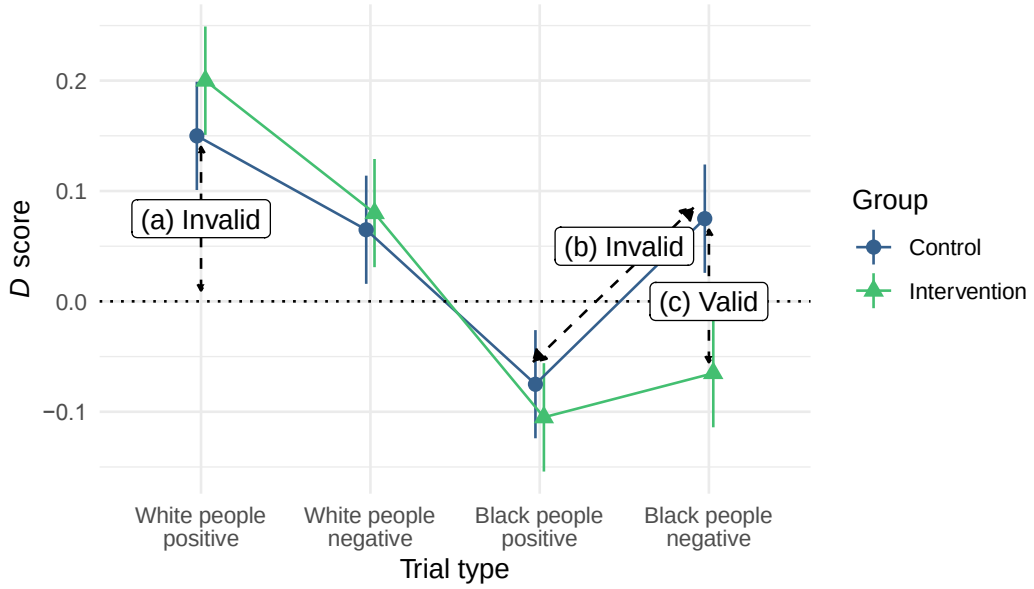

Figure 4. Statistical comparisons whose substantive conclusions are invalidated by the existence of the generic pattern among IRAP effects.

It is useful to unpack the example comparisons made in Figure 4 in detail in order to understand the validity of their domain level conclusions. In the case of the comparison labelled (a), a researcher might observe that the 'White people – positive' trial type showed an effect that was significantly greater than zero (e.g., using a one-sample *t* test). While it is indeed

correct to describe the group as having demonstrated a non-zero IRAP effect here, it would be invalid to interpret this as evidence of a substantive domain-specific effect regarding evaluations of 'White people' more generally. For example, conclusions such as "the sample evaluated White people positively" would be invalid because this particular IRAP effect would likely be generated regardless of what category stimuli were used (i.e., it is confounded). That is, our understanding of the generic pattern implies that the effect in our hypothetical study likely has little to do with the stimulus category 'White people', and therefore no conclusions regarding participant evaluations of 'White people' should be made. In general, we therefore recommend that comparisons of IRAP *D* scores against the zero point (e.g., via one sample *t* tests) should be avoided when attempting to make substantive conclusions about the domain being assessed in an IRAP.

For the comparison labelled (b), a researcher might observe that the 'Black people – negative' trial type showed an effect that was significantly larger than the 'Black people – positive' trial type (e.g., using a paired samples *t* test). They might then make the domain level conclusion that 'implicit negativity towards Black people is stronger than implicit positivity towards them'. Although initially less obvious, this inference still relies on a common interpretation of the zero point between the two trial types (i.e., that $D = 0$ has some shared domain-level meaning between trial types). However, our results demonstrated that this is not the case, as the generic pattern takes the form of IRAP effects of different magnitudes between trial-types (see Figure 2 and 3). As such, this substantive conclusion would also be invalid. In general, we therefore recommend that comparisons of IRAP *D* scores between trial types within a single IRAP (e.g., via paired-samples *t* tests) should be avoided when attempting to make domain level conclusions.

Finally, for the comparison labelled (c), a researcher might observe that mean effects on the 'Black people – negative' trial type were significantly different between control and

intervention conditions. They might then conclude that their intervention 'served to reduce implicit negativity towards Black people'. Because this comparison involves scores on only a single trial-type, with no direct or tacit reliance on interpretation of the zero point, this domain level conclusion would not be invalidated by the existence of the generic pattern among IRAP effects. Similarly, a comparison made within-subjects on the same trial-type (e.g., pre-post intervention) would also remain valid. Although not illustrated in Figure 4, domain-level conclusions of the results of correlations among trial types and between trial types and external variables (e.g., self-report or behavioural tasks) would also not be invalidated by the existence of the generic trial type effect (see Table 1 and Figure 4).

Given that we have argued that many common analyses of IRAP data give rise to invalid results, it would seem important to assess the prevalence of such invalid inferences and conclusions in the published literature. While this is beyond the scope of the current article, a systematic review of the IRAP literature is being conducted to address this question. We readily admit that many articles we ourselves have written are likely to contain inferences and conclusions that we are now recognizing as invalid.

Conclusions

Evidence from a large dataset of published and unpublished IRAP studies show that IRAPs examining very different domains – even those using non-words – demonstrate startlingly similar patterns of effects. This finding is in agreement with general conclusions of several recent articles that there is a generic pattern among IRAP effects. However, due to its large sample size relative to the existing IRAP literature, this study is the first to quantify the generic pattern more precisely and to consider its implications for the valid interpretation of published and future IRAP studies. Results demonstrated that majority of variance in effects on evaluative IRAPs is attributable to this generic pattern rather than the domain it is intended to measure. The IRAP is therefore relatively insensitive to the attitudes and learning histories

it is intended to assess. This represents a serious confound and has negative implications for the published literature: multiple common analyses of IRAP data are likely to produce invalid domain level conclusions. There is therefore a strong need to systematically evaluate the prevalence of these types of analyses and invalid inferences in the published IRAP literature.

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*, 527–542.

Barnes-Holmes, D., & Hussey, I. (2016). The functional-cognitive meta-theoretical framework: Reflections, possible clarifications and how to move forward. *International Journal of Psychology*, *51*(1), 50–57. https://doi.org/10.1002/ijop.12166

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

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

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

Gawronski, B., & De Houwer, J. (2011). Implicit measures in social and personality psychology. In C. M. Judd (Ed.), *Handbook of research methods in social and personality psychology* (Vol. 2). Cambridge University Press. 10.1017/CBO9780511996481.016

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

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, *61*(3), 465–498.

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. (2020). The IRAP is not suitable for individual use due to very wide confidence intervals around D scores. *Preprint*. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I., Daly, T., & Barnes-Holmes, D. (2015). Life is Good, But Death Ain't Bad Either: Counter-Intuitive Implicit Biases to Death in a Normative Population. *The Psychological Record*, *65*(4), 731–742. https://doi.org/10.1007/s40732-015-0142-3

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. https://CRAN.R-project.org/package=ez

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, *15*(4), 152–159. https://doi.org/10.1016/j.tics.2011.01.005

O'Shea, B., Watson, D. G., & Brown, G. D. A. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment*, *28*(2), 158–170. https://doi.org/10.1037/pas0000172

Pfister, R., & Janczyk, M. (2019). *schoRsch: Tools for Analyzing Factorial Experiments* (1.7). https://CRAN.R-project.org/package=schoRsch

R Core Team. (2022). *R: A language and environment for statistical computing* (4.2). R Foundation for Statistical Computing. https://www.R-project.org/

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(4), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M. A., & De Raedt, R. (2013). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self-versus ideal self-related cognitions in dysphoria. *Cognition & Emotion*, *27*(8), 1441–1449. https://doi.org/10.1080/02699931.2013.786681

Remue, J., Hughes, S., De Houwer, J., & De Raedt, R. (2014). To Be or Want to Be: Disentangling the Role of Actual versus Ideal Self in Implicit Self-Esteem. *PLoS ONE*, *9*(9), e108837. https://doi.org/10.1371/journal.pone.0108837

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. Social Science Research Network. http://papers.ssrn.com/abstract=2160588

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482.

Declaration of conflicts of interest: none