The IRAP is not very sensitive to the attitudes and learning histories it seeks to assess

Ian Hussey & Chad Drake

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 This study pooled published and unpublished (file-drawer) studies in multiple domains to obtain a large sample size (N = 501). 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. The existence of the generic pattern may also undermine the validity of many conclusion made in the published IRAP literature.

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).

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). This effect may explain some of the counter-intuitive findings in the IRAP literature, such as the finding that normative participants demonstrate positive evaluations of death on the IRAP (Hussey, Daly, et al., 2015).

More specifically, 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 have 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 subsequent publications (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, they presence, replicability, and generalizability of these biases in IRAP effects is apparently uncontroversial.

Goals of the current research

In contrast with previous work, which 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 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 40 to 60 participants) and a limited range of domains. In order to overcome this, this article therefore used an unprecedently large sample ($N = 501$) across multiple attitude domains ($k = 7$). 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; (2) estimate the generic pattern more precisely; (3) 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 (4) make recommendations about which common analytic strategies give rise to valid versus invalid inferences as a result of this generic pattern.

Method

All analysis code is available on the Open Science Framework (osf.io/vhzsn). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons et al., 2012).

Data

Data was pooled from among the authors' published and unpublished (file-drawer) IRAP studies. The current study therefore employs secondary analysis of existing data, with sample size being determined by data availability. Inclusion criteria were as follows: (1) Studies must include at least one standard IRAP (i.e., not variants such as the MT-IRAP or Training IRAP). (2) The study must have been completed in an experimental setting from university student populations. (3) 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. (4) The IRAP must have used 'True' and 'False' as response options within the procedure. (5) When a study employed multiple IRAPs within participants, only the first IRAP that each participant completed was included. Data from tasks other than the IRAP and demographics items were not considered.

Data from 11 studies across 7 domains and a total of 501 participants met inclusion criteria. Some of this data has been used in other publications for different purposes (Drake et al., 2016; Hussey, Daly, et al., 2015). It is worth noting that this sample size is roughly than 12 times the size of the typical IRAP study, and includes more participants that the all studies combined in a previous meta-analysis of clinically-relevant IRAP effects (Vahey et al., 2015).

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 21 participants (4%) were excluded on this basis. In the remaining participants, accuracy and latency performances were typical of previous IRAP studies ($M_{\text{RT}} = 1482$, $SD = 268$; $M_{\text{accuracy}} = 92\%$, $SD = 4.7\%$).

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 480 participants (307 women [64%], 159 men [33%], 1 identified as nonbinary; $M_{\text{age}} = 21.4$, $SD = 6.5$). 13 participants had missing demographics data. *N*s per domain were: 29 Body Image, 122 Life-Death, 97 Lincoln-Hitler, 62 Personalized Friend-Enemy, 112 Race, 35 Religion, and 23 Non-Words.

Table 1. Stimuli used in each IRAP.

	Death			Race		Body shape	Christian-Muslim	Lincoln-Hitler	Friend-enemy	Non-words
	Death 1	Death 2	Death 3	Race 1	Race 2					
Category 1	My Life	My Life	Living	*Images of white faces*	White people	Skinny	Christians	Abraham Lincoln	[friend's name]	CUG
Category 2	My Death	My Death	Dying	*Images of black faces*	Black people	Fat	Muslims	Adolf Hitler	[enemy's name]	VEC
Attributes 1	Exciting	Escape	Exciting	Safe	Good	Active	Compassionate	Caring	Caring	Exciting
	Great	Calm	Great	Friendly	Worthy	Attractive	Correct	Friend	Friend	Great
	Lovely	Soothing	Lovely	Polite	Deserving	Desirable	Good	Good	Good	Lovely
	Pleasant	Relief	Pleasant	Kind	Superior	Disciplined	Loving	Nice	Nice	Pleasant
	Satisfying	Peaceful	Satisfying		Motivated	Good	Safe	Safe	Safe	Satisfying
	Enjoyable	Comfort	Enjoyable		Smart	Healthy	Truthful	Trustworthy	Trustworthy	Enjoyable
Attributes 2	Distressing	Distressing	Distressing	Dangerous	Bad	Bad	Cruel	Bad	Bad	Distressing
	Awful	Awful	Awful	Aggressive	Deficient	Ill	Flawed	Cruel	Cruel	Awful
	Hurtful	Hurtful	Hurtful	Rude	Inadequate	Lazy	Bad	Dangerous	Dangerous	Hurtful
	Horrible	Horrible	Horrible	Violent	Inferior	Sloppy	Hateful	Enemy	Enemy	Horrible
	Painful	Painful	Painful		Lazy	Disgusting	Dangerous	Hateful	Hateful	Painful
	Upsetting	Upsetting	Upsetting		Stupid	Ugly	Dishonest	Selfish	Selfish	Upsetting

Notes: The friend-enemy IRAP employed personalized category 1 and 2 stimuli that were provided by the participant.

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 presents 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 hand 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. Between blocks of trials, this correct response changes so that, for example, participants must respond to "white people" and "dangerous" with "True" on one block and "False" on the other block. Participants complete pairs of these blocks in two phases: practice and testing. In order to progress from practice to testing, the participant must respond quickly and accurately on both blocks within the pair (typically with median reaction time $<$ 2000 ms and percentage accuracy $>$ 80%). Should they fail to meet this criteria, the participant completes another pair of practice blocks. Should they meet the criteria, they progress to the testing phase where they complete three pairs of blocks in a row. Following standard practice, only reaction time data from the test blocks is used in the analyses (Hussey, Thompson, et al., 2015).

Data processing

IRAP studies typically using 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 trimmed and 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'. Details of each IRAP's stimuli, task parameters, and responding rules can be found in Table 1.

Results

Comparisons of different domains

We report both frequentist (*p*-value based) and Bayes Factors ANOVAs. The former has the benefits of providing multiple metrics of effect size. The latter have the benefit of being able to quantify the evidence in favor of the null hypothesis, which was relevant to our research question here. Note that it is a common misconception that a non-significant *p* value can be interpreted as evidence for absence of differences between condition, but this is in fact not the case (Greenland et al., 2016). All analyses were done in R (R Core Team, 2020) using the packages ez (Lawrence, 2016), schoRsch (Pfister & Janczyk, 2019), and BayesFactor (Morey et al., 2018).

First, we assessed the evidence for a generic pattern among IRAP effects. Next, we assessed the relatively magnitudes of the generic pattern to the IRAP's sensitive to the stimulus domain being assessed. We hypothesized that if the IRAP is relatively sensitive to the domain being assessed, a greater proportion of variance should be attributable to the main effect for IRAP stimuli domain and/or the interaction between domain and trial type. However, if IRAP effects are mostly driven by the generic pattern, then the main effect for trial type effect would be larger. This would imply that the IRAP is relatively insensitive to the stimulus domain being assessed.

A mixed within-between frequentist ANOVA was run using type III sum of squares method with IRAP *D* scores as the DV, IRAP trial-type as the within subjects IV (category A-positive, category A-negative, category B-positive, category B-negative), and domain as the between subjects IV. Only data from the domains using known-words was used (i.e., all domains other than the non-words IRAP). This revealed a main effect for both domain, $F(5, 451) = 3.87$, $p = .002$, and trial-type, $F(3, 1353) = 123.65$, $p < .001$, but not their interaction, $F(15, 1353) = 0.69$, $p = .789$. A comparable Bayes Factor ANOVA was run using a default prior (JZS prior with Cauchy distribution with $r = 0.5$ placed on the effect size). In the case of main effects, Bayes Factors compared the hierarchical models of (H1) all main effects, relative to (H0) all main effects other than the effect of interest (i.e., the equivalent of frequentist type II sum of

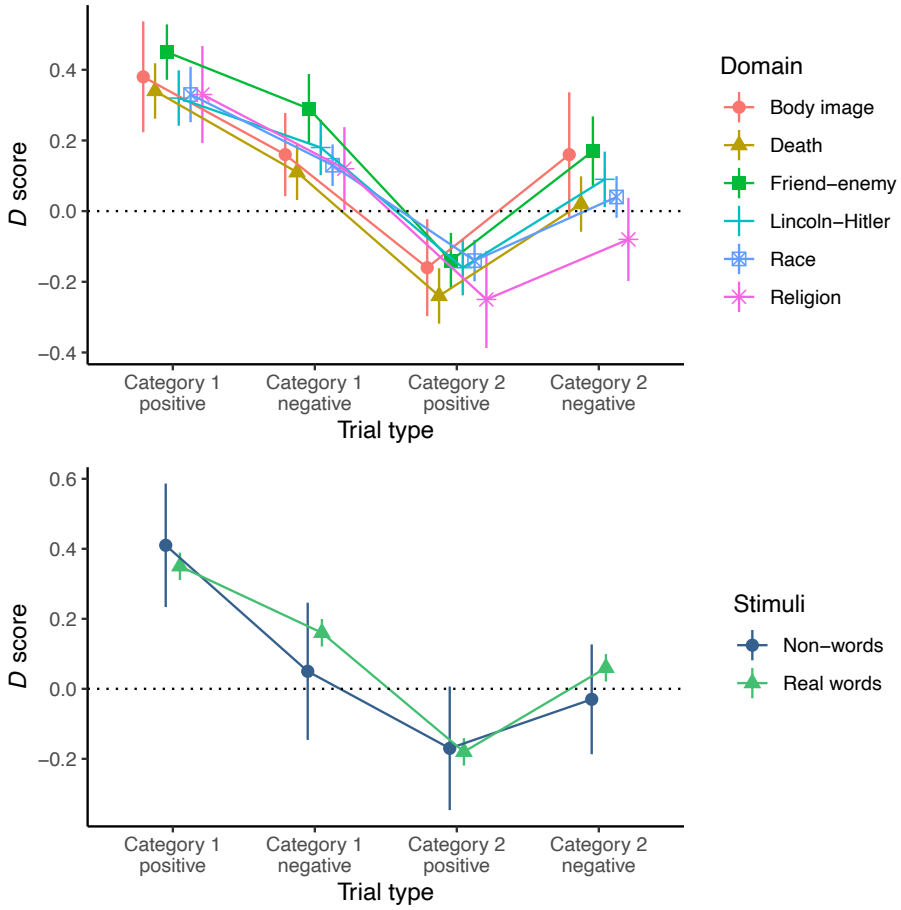

Figure 1. IRAP effects between domains (upper panel) and between words vs. non-words IRAPs (lower panel). Error bars represent 95% CIs.

squares method). In the case of interaction effects, Bayes Factors compared the hierarchical models of (H1) all main effect and their interaction effect, relative to (H0) all main effects without the interaction effect of interest (i.e., the equivalent of frequentist type III sum of squares method). Results were interpreted following the guidelines defined by Lee and Wagenmakers (2013). For readers less familiar with Bayes factors, $BF_{10} > 3$ corresponds very roughly with $p < .01$ in the frequentist framework (although it should be noted that the two do not have a one-to-one interpretation and can provide divergent results; for a thorough yet accessible introduction to analyses using Bayes Factors see Rouder et al., 2009). Results provided no strong evidence for or against the main effect for domain ($BF_{10} = 1.63$), extreme evidence for trial type ($BF_{10} > 1000$), and extreme evidence against the interaction effect ($BF_{10} < 0.001$). Results from both frequentist and Bayes Factors hypothesis testing methods therefore supported the conclusion that IRAP effects are mostly driven by the generic pattern rather than the domain being assessed. Figure 1 (upper panel illustrates the comparable generic pattern in between-trial type effects that is observed in all domains.

Frequentist standardized effect sizes were also calculated, where η^2 refers to the percentage of variance explained (akin to r^2); ('partial') η^2_p refers to the percentage of variance explained after controlling for all other effects; and ('generalized') η^2_g 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 1000 replications. Median bootstrapped estimates are reported for robustness. Following standard practice, 90% confidence intervals are reported on the basis that squared effect sizes can only be positive numbers. Comparisons can be made between the effects on the basis of the non-overlap of these intervals. Given the mix of within and between-subject effects, η^2_g is the most appropriate metric to compare in order to make such inferences. 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 [0.63, 0.77], η^2_p = 0.22, 90% CI [0.19, 0.25], η^2_g = 0.16, 90% CI [0.13, 0.18]) with only a small fraction explained by either the main effect for domain (η^2 = 0.06, 90% CI [0.03, 0.10], η^2_p = 0.05, 90% CI [0.02, 0.07], η^2_g = 0.02, 90% CI [0.01, 0.03]) or the interaction between domain and trial type (η^2 = 0.04, 90% CI [0.02, 0.06], η^2_p = 0.01, 90% CI [0.01, 0.02], η^2_g = 0.01, 90% CI [0.00, 0.02]). Results from the effect sizes therefore also supported the conclusion that IRAP effects are mostly driven by the generic pattern are relatively insensitive to the domain being assessed.

Comparisons of known domains and non-word stimuli

A second set of ANOVAs compared IRAP effects between trial type and stimuli type – that is, between IRAPs that employed real words from known domains words as concept category domains (e.g., race, religion, etc.) versus unknown non-words (i.e., evaluations of the non-words CUG and VEC, as used in O'Shea et al., 2016). This revealed a main effect for trial type, $F(3, 1434) = 34.30$, $p < .001$, $\eta^2 = 0.82$, 90% CI [0.68, 0.89], $\eta^2_p = 0.07$, 90% CI [0.04, 0.09], $\eta^2_g = 0.05$, 90% CI [0.03, 0.06], but no evidence for a main effect for stimuli type, $F(1, 478) = 0.41$, $p = .520$, $\eta^2 = 0.01$, 90% CI [0.00, 0.06], $\eta^2_p < 0.01$, 90% CI [0.00, 0.01], $\eta^2_g < 0.01$, 90% CI [0.00, < 0.01], or their interaction effect, $F(3, 1434) = 0.99$, $p = .394$, $\eta^2 = 0.04$, 90% CI [0.00, 0.11], $\eta^2_p < 0.01$, 90% CI [0.00, 0.01], $\eta^2_g < 0.01$, 90% CI [0.00, 0.01]. Bayes Factors provided extreme evidence for trial type ($BF_{10} > 1000$), moderate evidence against stimuli type ($BF_{10} = 0.18$), and strong evidence against the interaction effect ($BF_{10} = 0.09$). Results suggested that, given available data, there is reason to believe that a similar generic IRAP pattern emerges regardless whether the concept words employed in the task are from known domains (e.g., race, religion) or unknown non-words. The generic pattern among IRAP effects, even among non-word stimuli, is illustrated in Figure 1 (lower panel).

Discussion

Both hypothesis testing methods and effect sizes support the conclusion that IRAP effects are predominantly driven by a generic pattern among the trial-types rather than the domains being assessed. The IRAP is therefore relatively insensitive the attitudes or learning histories that it is intended to assess. Due to our far larger sample size and variety of domains, results also shed light on the nature of the effect: the generic pattern appears to take the form of a specific pattern among the trial-types (see Figure 1). In behavioural terms, the category stimuli exert relatively weak stimulus control over the relative speed of responses 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).

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., 2018, 2018; O'Shea et al., 2016). In the IRAP's modal use-case, the presence of a generic pattern is likely to either represent a strong barrier to the task being useful to their goals. Or, more worryingly, the generic pattern may cause researchers to make invalid inferences, by misattributing the presence of IRAP effects to attitudes or learning histories 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.

Implications for the validity of conclusions in the published literature

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. 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 be equated with evidence of implicit attitudes (i.e., cognitive abstractions) or learning histories involving the category stimuli (i.e., behavioural abstractions). Analyses that treat D = zero point as a reference

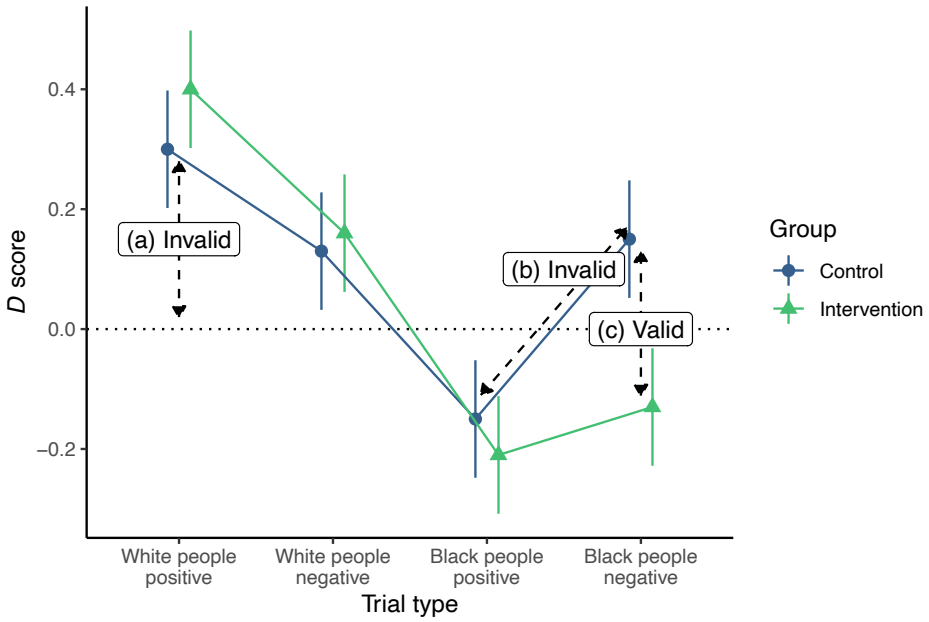

Figure 2. Comparisons that would produce valid versus invalid domain-level conclusions on a hypothetical IRAP.

point, even tacitly, will therefore produce false or invalid domain level inferences. Somewhat unfortunately, this concerns about the validity of conclusions when the zero point 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.

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 the most common forms of analysis of IRAP data in turn. Table 2 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 2 uses results from a hypothetical between-groups IRAP study to illustrates some of these common comparisons and the validity of their domain level conclusions.

It is useful to unpack the example comparisons made in Figure 2 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 domain-specific effect. For example, the conclusion that "the sample evaluated White people positively", or other similar conclusions, conclusions would be invalid because this particular IRAP effect would likely be generated regardless of what category stimuli were used. 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 participants 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 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 positively 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, the empirical data we presented earlier demonstrates that this is not the case, as the generic pattern takes the form of IRAP effects of different magnitudes between trial-types (see Figure 1). As such, this conclusions would also be invalid. In general, we therefore recommend that comparisons of IRAP D scores between trial

Table 2. A description of the commonly-used methods of analysis for IRAP data, as well as the inferences which tend to be made on their basis.

Comparison	Example research question	Analytic method	Common inference	Validity
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.	Invalid
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.	Invalid
Mean *D* scores from a given trial-type and participant compared between time points	Do *D* score on the "white people – positive" differ between timepoints 1 and 2?	Within-subjects *t*-test/ANOVA	"White people – positive" bias did not change between timepoints/after the intervention.	Valid
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.	Valid
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.	Valid
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.	Valid

Notes: Validity refers to the validity of domain-level conclusions in light of the presence of the generic pattern among IRAP effects.

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 2, 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) are also unaffected by the existence of the generic trial type effect (see Table 2).

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 invalid inferences.

Conclusions

Evidence from a large dataset of published and unpublished IRAP studies demonstrated that IRAP from very different domains – even those using non-words – demonstrate startlingly similar patterns of effect. This finding is in agreement with several recent papers in the literature that argue that there is a generic pattern among IRAP effects. However, due to its relatively large sample size, this study is the first to quantify the generic pattern more precisely, and to consider its implications for the validity of published and future IRAP studies. 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.

Notes

Funding: This research was conducted with the support of Ghent University grant 01P05517 to IH.
Conflicts of interest: None.

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

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

Drake, C. E., Seymour, K. H., & Habib, R. (2016). Testing the IRAP: Exploring the Reliability and Fakability of an Idiographic Approach to Interpersonal Attitudes. *The Psychological Record*, *66*(1), 153–163. https://doi.org/10.1007/s40732-015-0160-1

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

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

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*. 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

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs* (Version 0.9.12-4.2) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

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* (Version 1.7) [Computer software]. https://CRAN.R-project.org/package=schoRsch

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0) [Computer software]. 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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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

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

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