Interpreting and inverting with less cursing: A guide to interpreting IRAP data

This Professional Interest Brief seeks to provide a clear guide to interpreting data generated by Implicit Relational Assessment Procedure (IRAP). The interpretation of IRAP data is not immediately intuitive and yet has received little explicit attention in the published literature. As such, it is hoped that this guide will help clarify this matter, particularly for those new to using the IRAP or intending to use the measure in the future. In doing so, we hope to make the measure more accessible and facilitate continued use of the methodology and its contribution to the contemporary Relational Frame Theory (RFT) literature.

Keywords: IRAP; Implicit Relational Assessment Procedure

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. http://doi.org/10.1016/j.jcbs.2015.05.001

Article DOI: http://doi.org/10.1016/j.jcbs.2015.05.001

Interpreting and inverting with less cursing: A guide to interpreting IRAP data

Ian Hussey[1], Miles Thompson[2], Ciara McEnteggart[1], Dermot Barnes-Holmes[1] & Yvonne Barnes-Holmes[1]

[1]Maynooth University, [2]Goldsmiths, University of London

Accepted: Journal of Contextual Behavioral Science

Author note

Correspondence should be addressed to the first author (Ian.Hussey@nuim.ie). IH was assisted by a Government of Ireland Scholarship by the Irish Research Council.

Abstract

This Professional Interest Brief seeks to provide a clear guide to interpreting data generated by Implicit Relational Assessment Procedure (IRAP). The interpretation of IRAP data is not immediately intuitive and yet has received little explicit attention in the published literature. As such, it is hoped that this guide will help clarify this matter, particularly for those new to using the IRAP or intending to use the measure in the future. In doing so, we hope to make the measure more accessible and facilitate continued use of the methodology and its contribution to the contemporary literature on Relational Frame Theory (RFT).

Interpreting and inverting with less cursing: A guide to interpreting IRAP data

One of the cornerstones of Contextual Behavioral Science (CBS: Hayes, Barnes-Holmes, & Wilson, 2012) is its appeal to a basic account of human language and cognition through Relational Frame Theory (RFT). Relational Frame Theory argues that the fundamental building block of human cognitive abilities, such as abstract reasoning and generative language is "arbitrarily applicable relational responding" (AARR: see Hayes, Barnes-Holmes, & Roche, 2001). Much early RFT research revolved around demonstrating its proposed analytic units, relational frames, that were established in the laboratory (see Hughes & Barnes-Holmes, in press, for review). However, in recent years, RFT researchers have attempted to extend RFT's conceptual account by also assessing histories of relational responding that were established outside of the laboratory (see Barnes-Holmes, Barnes-Holmes, & Hussey, in press), such as by posing questions about the probability or "strength" of individuals' relational responding in applied domains such as obsessive compulsive tendencies, depression, or professional burnout (see Nicholson & Barnes-Holmes, 2012a; Hussey & Barnes-Holmes, 2012; Kelly & Barnes-Holmes, 2013, respectively). In order to do this, RFT researchers have built on methodologies frequently used in cognitive and social psychology to assess what are referred to as "implicit attitudes" (see De Houwer & Moors, 2010; see also Hughes, Barnes-Holmes, & Vahey, 2012). This has produced a procedure that has shown utility in assessing the relative strength of relational responding: the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). IRAP research now represents one of the forefronts of RFT research (Barnes-Holmes et al., in press).

Task structure

A brief description of the procedure will now be provided, as the interpretation of IRAP data is best understood through an understanding of the structure of the task itself. The

IRAP involves presenting pairs of stimuli to participants on a computer screen. Participants respond to blocks of these stimulus pairings, and are required to respond as accurately and quickly as possible according to what we will describe as two responding rules. In some IRAP studies, specific instructions regarding these rules are provided before each block (e.g., "Respond as if I am positive and others are negative"). However, in other studies, specific instructions to respond according to a particular rule for each block are not provided (e.g., "Try to get as many correct as possible – go fast, making a few errors is ok"). For the purposes of communication, however, we will describe the task in terms of utilizing two specific types of rule. In short, the IRAP compares, under accuracy and latency pressure, the relative ease (i.e., speed) with which participants respond according to one rule relative to the other. In other words, the IRAP is a procedure that is used to assess subtle reaction time biases that are often referred to as reflecting "implicit attitudes" (De Houwer & Moors, 2010).

Table 1. *Example rules and stimuli for an IRAP investigating self-esteem*

Rule A Respond as if I am good and others are bad	
Rule B Respond as if I am bad and others are good	
Label 1: "Self"	**Label 2: "Others"**
I am	Other people are
I'm	Others are often
I think I am	Other people can be
Target 1: "Positive"	**Target 2: "Negative"**
Loyal	Manipulative
Trustworthy	Dishonest
Kind	Cruel
Moral	Horrible
Generous	Selfish
Friendly	Heartless
Response Option 1	**Response Option 2**
True	False

For illustrative purposes, consider the stimulus set for the hypothetical self-esteem IRAP outlined in Table 1 (see Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2010, for an alternative published version of a self-esteem IRAP). One of the two responding rules, Rule A (e.g., "Respond as if I am good and others are bad") or Rule B (e.g., "Respond as if I am bad and others are good"), is presented to participants before each block of trials .We will refer to these as Rule A blocks and Rule B blocks. Table 1 also lists what are arbitrarily referred to as label stimuli (presented at the top of the screen), target stimuli (presented in the middle of the screen) and response options (presented at the bottom of the screen). Label stimuli frequently contain what can loosely be referred to as categories (e.g., self vs. others), whereas target stimuli frequently contain attributes (e.g., positive vs. negative). These four classes of stimuli each contain one or more exemplars of the relevant category and attribute (e.g., loyal, trustworthy, kind, etc.). However, when describing the data researchers typically refer only to the overarching functional class (e.g., positive or self). Finally, participants respond using one of the two response options, which are typically mapped to the "D" and "K" keys (e.g., similar and different, or true and false).

Each IRAP trial presents one label stimulus and one target stimulus and both response options. The combination of two label categories (e.g., self and others) and two target categories (e.g., positive and negative) produce four possible "trial-types" (e.g., trial-type 1 = self-positive, trial-type 2 = self-negative, trial-type 3 = others-positive, and trial-type 4 = others-negative). It is important to note that the trial-types are procedurally separated, insofar as label 1 stimuli are never presented within the same trial as label 2 stimuli, and target 1 stimuli are never presented within the same trial as target 2 stimuli.

The required correct and incorrect response options for each trial-type in each of the two rule blocks are pre-determined by the task structure itself (Table 2; see Barnes-Holmes, Barnes-Holmes, et al., 2010). To illustrate, let us return to the hypothetical self-esteem IRAP.

Rule A block employs contingencies that require participants to respond as if "I am good and others are bad". For example, a self-positive trial (i.e., trial-type 1) might present the participant with the stimuli "I am" and "loyal" and the response options "True" and "False". In this case, True would be the correct response, by definition, while selecting False would present the participant with a red X. However, if these same stimuli appeared on a Rule B block trial, the correct response would now be False. The IRAP is arranged in this way in order to assess the difference in reaction times between Rule A and Rule B blocks for each trial-type (e.g., the difference in speed between responding True on Rule A blocks vs. False on Rule B blocks). Furthermore, participants are presented with pairs of Rule A and Rule B blocks, each of which contains a large number of IRAP trials in order to capture a sufficient number of reaction times to conduct a meaningful analysis (e.g., 48). Typically, participants complete pairs of practice blocks until they meet both accuracy and latency mastery criteria, followed by three test block pairs (see Barnes-Holmes, Barnes-Holmes, et al., 2010). Given that the IRAP effect is produced via accuracy and latency pressure, these criteria should be set as high as is feasible. Recent studies have frequently employed accuracy $\geq$ 80% and median time to first correct response $\leq$ 2000 ms, but future work may of course tighten these criteria further. It should be noted that both mastery criteria must be met within both blocks in a block pair for the criteria to have been met. On balance, variations on these criteria have not been systematically explored, and future efforts might revise these practices.

Table 2. *Required responses on each trial-type within a hypothetical self-esteem IRAP*

	Trial-type 1: **Self-positive**	**Trial-type 2: Self-negative**	**Trial-type 3: Others-positive**	**Trial-type 4: Others-negative**
Rule A block ('Respond as if I am good and others are bad')	True	False	False	True
Rule B block ('Respond as if I am bad and others are good')	False	True	True	False

Interpretation of IRAP effects

Methods of quantifying effects on the IRAP

To reiterate, the IRAP presents stimuli to participants in pairs of blocks. The same categories of stimuli are presented in both blocks. However, the critical difference between the two blocks is that the required response option for each trial-type alternates between them. For example, on one block, participants must respond to a given stimulus pair (e.g., "I am" and "Loyal") with one response option (e.g., "True"), whereas on the other block, participants must respond with the other response option (e.g., "False"). The IRAP researcher then seeks to quantify the difference in responding speed between the two blocks in any pair. Loosely, this difference indicates which responding direction makes more intuitive sense or is more "automatic" for an individual (De Houwer & Moors, 2012).

While this difference can be quantified in numerous ways, specific common practices have emerged from the broader literature on the analysis of reaction-time data (see Ratcliff, 1993; Whelan, 2008). In particular, due to the distribution of reaction times, some form of normalization technique is recommended when quantifying the differences between block pairs. The most common way to quantify the difference between responding on two paired blocks is to treat the difference as an effect, and thus to estimate the size of this effect using an adaptation of Cohen's *d* (Cohen, 1977) known as a *D* score (see Barnes-Holmes, Barnes-

Holmes, et al., 2010; Greenwald, Nosek, & Banaji, 2003). In short, the *D* score involves removing all latencies above 10,000 ms, and then calculating the difference between the mean reaction time to first correct response on each of the two blocks divided by the standard deviation of all the reaction times in both blocks (see Table 3). *D* scores have a theoretical range of -2 to +2. The generic concept of the *D* score has been applied to the IRAP's specific structure (i.e., by calculating scores for each trial-type rather than for the overall block). These are often referred to as "D_{IRAP} scores" when applied to the IRAP, but will be referred to here as *D* scores for simplicity (Barnes-Holmes, Barnes-Holmes, et al., 2010).

It should be noted that the *D* scores produced by several implementations of the IRAP (e.g., the IRAP 2008, 2010, 2012, and 2014 programs: Barnes-Holmes, 2014) are calculated based on time from stimulus onset to first correct response. Although other versions of the *D* score do exist (see Greenwald et al., 2003), no published work has systematically explored the relative utility of different scoring algorithms within the context of the IRAP. Nonetheless, this version of the *D* score appears to be performing well, given that it was employed in the overwhelming majority of studies included in a recent meta analysis of clinically relevant IRAPs (Vahey, Nicholson, & Barnes-Holmes, 2015).

It is critical to note, however, that although the *D* score is included in several implementations of the IRAP, it is an analytic method and not an essential part of the procedure itself. Indeed, not all studies have employed it (e.g., Kishita, Takashi, Ohtsuki, & Barnes-Holmes, 2014), and many other strategies could be used to quantify the difference in reaction times depending on a researcher's goals. For example, *G* scores are a non-parametric alternative that convert reaction times into fractional ranks (Nosek, Bar-Anan, Sriram, Axt, & Greenwald, 2014).

Table 3. *Scoring and processing IRAP data using D scores or G scores.*

Scoring	*D* score	*G* score
1. Define measurement unit	Reaction times are defined as the time from stimulus presentaiton to first *correct* response.	Reaction times are defined as the time from stimulus presentaiton to first *correct* response.
2. Exclude 'fast' responders	For each participant, exclude all IRAP data if the reaction times of more than 10% of test block trials are < 300 ms.	For each participant, exclude all IRAP data if the reaction times of more than 10% of test block trials are < 300 ms.
3. Remove outliers	For each trial-type in each block pair, remove latencies > 10,000 ms.	For each trial-type in each block pair, remove latencies > 10,000 ms.
4. Calculate scores	For each trial-type in each block pair: i. $D = (M_B - M_A)/SD_{AB}$, where M_A = mean latencies block A, M_B = mean latencies block B, SD_{AB} = standard deviation of latencies in both blocks A and B[1]	For each trial-type in each block pair: i. Fractionally rank the *N* latencies across both block A and block B[1]: $\frac{1}{1}, \frac{1}{2}, \frac{1}{3} \ldots \frac{1}{N}$ In case of ties, ranks are averaged across tied values ii. Subtract $\frac{1}{2N}$ from each fractional rank iii. For each of the fractional ranks, computer the standard normal deviate (mean = 0 and SD = 1) iv. $G = M_B - M_A$, where M_A = mean of normal deviates in block A, M_B = mean of normal deviates in block B
Processing	**Both scoring algorithms**	
5. Exclude data that does not maintain accuracy and latency mastery criteria	For each participant, exclude data for participants who fail to maintain the mastery criteria in the test blocks. These mastery criteria may applied at the block level or at the level of the individual trial-type. Additionally, they have usually been applied in one of two ways: a) If participant fails to meet accuracy and latency criteria on either block within any test block pair, exclude this participant's IRAP data. This is more conservative (e.g., Barnes-Holmes, Murtagh, et al., 2010); or b) If participant fails to meet accuracy and latency criteria on either block in a test block pair, exclude data from this block pair. If more than one block pair is excluded, exclude this participant's data. This causes less attrition (e.g., Nicholson & Barnes-Holmes, 2012).	
6. Average scores across block pairs	For each trial-type, average *D* (or *G*) scores across test block pairs, e.g., $D_{\text{final}} = (D_{\text{test1}} + D_{\text{test2}} + D_{\text{test3}})/3$	
7. Perform trial-type inversions	Depending on the analysis and the contents of the IRAP's trial-types, invert two trial-types so as to create a common axis across the trial-types. See elsewhere in this brief.	

Note: It is important to note that several versions of the IRAP program (e.g. IRAP 2008, 2010, 2012, and 2014: see IRAPresearch.org) calculate D scores automatically using steps 2, 3, and 5. However, steps 1, 4, and 6 must be performed manually. Depending on the mastery criterion exclusion method (e.g., option b), step 5 may also need to be recalculated manually. [1]Blocks A and B are defined by the responding contingencies within them (e.g., in our example IRAP Rule A block = I am positive and others are negative vs. Rule B block = I am negative and others are positive) and do not refer to the order of presentation of the blocks.

For the purposes of clarity, and to encourage the consideration of a wider variety of scoring procedures, the specific steps involved in calculating both *D* scores and *G* scores are included in Table 3. This table also includes the details of two ways in which data has been excluded when participants fail to maintain criteria within the test blocks (i.e., removing data from particular blocks vs. excluding all of it, see Nicholson & Barnes-Holmes, 2012b; Barnes-Holmes, Murtagh, Barnes-Holmes, & Stewart, 2010). We have included this information here because, on reflection, previous papers have not always been clear about which steps are performed automatically by several implementations of the procedure and which must be performed manually by the researcher.

Due to the *D* score's popularity, the remainder of this article will discuss the interpretation of IRAP effects using this metric alone. Nonetheless, the following points are likely also to be applicable to other metrics. Whichever strategy is employed, researchers should ensure that their analysis takes into account the distribution of reaction times (Ratcliff, 1993; Whelan, 2008). For example, contrary to very early recommendations, simply calculating difference scores and using an arbitrary cut off value to create groups is not appropriate (e.g., Milne, Barnes-Holmes, Barnes-Holmes, & Stewart, 2005). It is important to note, however, that normalization techniques, such as the *D* score, normalize data within an individual's data and not across individuals. As such, they remain compatible with single-subject design approaches.

Interpreting trial-types

Let us assume that we have collected data from a number of participants using the self-esteem IRAP using a typical setup (e.g., completing practice blocks until mastery criteria are met followed by three pairs of test blocks; see Barnes-Holmes, Barnes-Holmes, et al., 2010). After averaging each participant's *D* scores across the three test block pairs, we are left with four *D* scores, one for each trial-type. Let us assume that the average of these *D*

scores across participants is as follows: self-positive, $D = 0.41$; self-negative, $D = 0.29$; others-positive, $D = -0.33$; and others-negative, $D = -0.14$. These notional D scores and their standard errors are graphed in Figure 1, upper panel ("Self-esteem IRAP"). We will now discuss how these effects should be interpreted. As discussed above, the absolute magnitude of the D score provides a metric of the size of the difference in reaction times between Rule A and Rule B blocks. However, due to the IRAP's multiple trial-types, interpreting the results is less intuitive than within other procedures such as the Implicit Association Test (IAT: Greenwald et al., 2003). The following paragraphs therefore aim to clarify and simplify the interpretation of D scores.

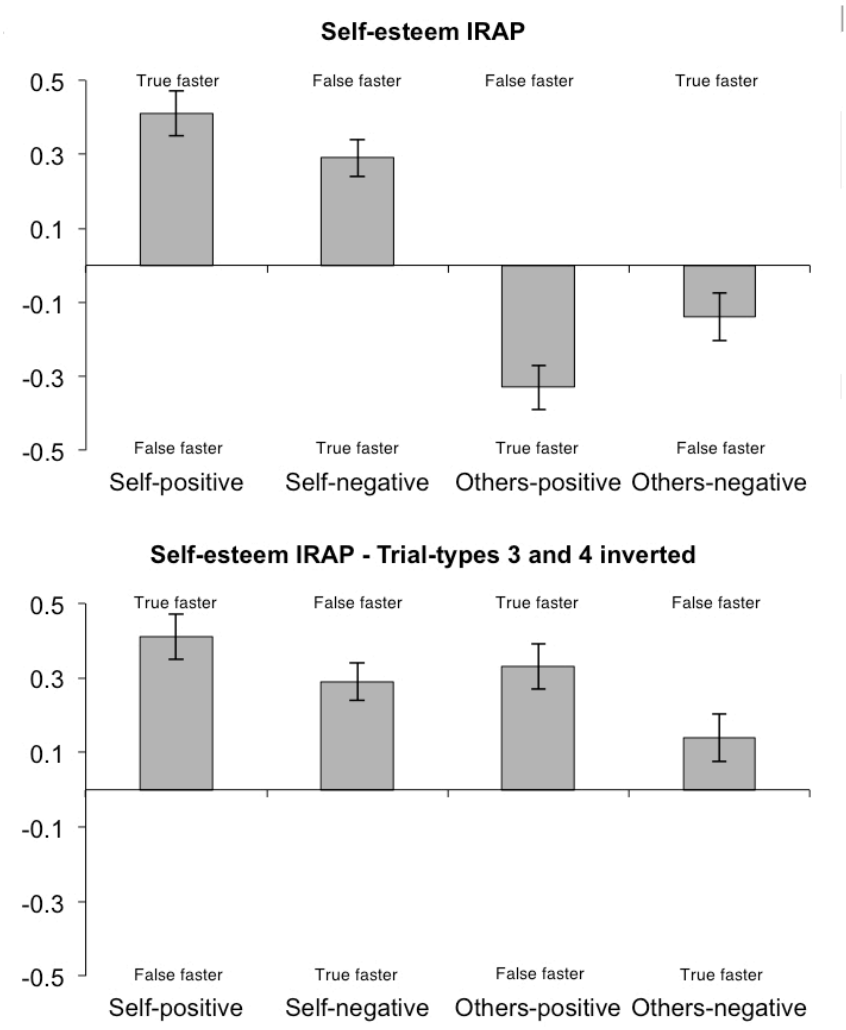

Figure 1. Self-esteem IRAP D scores by trial-type. Upper panel: D scores as output by the IRAP program. Lower panel: IRAP D scores with trial-types 3 and 4 inverted for clarity of interpretation.

On any given trial-type, a *D* score greater than zero indicates that participants responded more quickly during Rule A blocks than Rule B blocks. Conversely, a *D* score less than zero suggests the reverse (responding more quickly on Rule B blocks than Rule A blocks). Table 3 provides interpretations of potential *D* scores on the hypothetical self-esteem IRAP. As a memory aid, we often speak about the IRAP's output as following a "true-false-false-true" format. That is, assuming the IRAP uses "True" and "False" as response options, *D* scores greater than zero on each trial-type represent faster responding for True, False, False, and True, respectively. This rule of thumb may help to remind the researcher whether *D* scores greater than zero on each trial-type represent the *assertion* or *rejection* of the proposition contained within the trial-type – for example, whether responding on trial-type 1 (self-positive) was biased towards responding with "True" (which could be interpreted as an "I am positive" effect) or "False" (which could be interpreted as an "I am *not* positive" effect).

Table 4. *Interpretation of D scores greater than or less than zero for each trial-type within the self-esteem IRAP.*

	Trial-type 1: **Self-positive**	**Trial-type 2: Self-negative**	**Trial-type 3: Others-positive**	**Trial-type 4: Others-negative**
***D* scores > 0** (faster responding on Rule A block)	"True' faster than "False"	"False" faster than "True"	"False" faster than "True'	"True" faster than "False'
	I am positive	I am not negative	Others are not positive	Others are negative
***D* scores < 0** (faster responding on Rule B block)	"False" faster than "True"	"True" faster than "False"	"True" pressed faster than "False"	"False" faster than "True"
	I am not positive	I am negative	Others are positive	Others are not negative

Some published research has also calculated "overall" *D* scores by averaging the four trial-types (e.g., Hussey & Barnes-Holmes, 2012; Remue, De Houwer, Barnes-Holmes, Vanderhasselt, & De Raedt, 2013). Interpretations of such overall *D* scores must therefore be based on all four categories contained within the IRAP. For example, within our self-esteem IRAP, an overall *D* score greater than zero would represent a "self-positive/others-negative" effect (i.e., on the whole, participants evaluated self as being positive or *not*-negative to the degree that they evaluated others as negative or *not*-positive). In contrast, an overall *D* score less than zero would represent a "self-negative/others-positive" effect (i.e., on the whole, participants evaluated self as being negative or *not*-positive to the degree that they evaluated others as positive or *not*-negative). Overall *D* scores are therefore comparable to the interpretation of IAT *D* scores, in that they provide a single overall bias score (Greenwald, McGhee, & Schwartz, 1998). It should be noted, however, a primary reason for employing an IRAP over other procedures, such as the IAT, is its ability to parse out an overall effect into four individual bias scores (i.e., one for each trial-type). We therefore encourage researchers to make use of trial-type-level analyses wherever feasible.

When interpreting individual trial-types, it is important to ensure that the interpretation only ever refers to the classes of label, target, and response options that were presented in that trial-type, and not other classes. For example, the *D* scores greater than zero on trial-type 2 in the self-esteem IRAP should be interpreted as "I am *not* negative". While it may be tempting to collapse this double negative into a more commonsense "I am positive" interpretation, this must not be done because trial-type 2 only presents participants with negative attributes. Effects on this trial-type are therefore a specific *rejection* of "self-negative". This does not necessarily equate to an *assertion* of "self-positive". While this may initially seem pedantic, the ability to unpack subtle effects such as this is a key rationale for

using the IRAP in the first place. With this in mind, researchers must be careful that their interpretations map onto what participants actually respond to on-screen.

Closely related to this point is the issue of using negations in a stimulus set (e.g., not me, not good, etc.). To be specific, we strongly suggest that researchers avoid the use of negations in their stimulus sets. Such IRAPs can, of course, generate useful results (e.g., Remue et al., 2013), but participants often report finding such IRAPs difficult to complete. Imagine, for example, our hypothetical self-esteem IRAP used "I am not" as Label 2 rather than "Others are"; and that a *D* score less than zero was obtained on trial-type 3 (i.e., "I am not-positive"). In effect, participants have shown a bias towards responding with a double negative (i.e., "I am not-positive-False"), which of course is quite "cognitively" demanding when responding at speed. Of course, the issue of how stimuli should be selected, more generally, is a broader one that requires separate discussion in its own right. For the purposes of the current brief, it is suffice to say that researchers should be cognizant of the demands particular trial-types (and stimuli more generally) place on participants when designing stimulus sets.

Trial-type inversions

Many types of analyses require careful understanding not just of the interpretation of individual trial-types, but also of how they are interpreted in relation to one another. For example, we might ask whether participants implicitly evaluate self and others as equally positive. If a paired *t*-test were conducted on our hypothetical *D* scores produced by the IRAP, it would reveal a significant difference between trial-types 1 (self-positive) and 3 (others-positive; see Figure 1, upper panel). One might, therefore, erroneously conclude that there is a significant difference between the "I am positive" and "others are positive" effects in our hypothetical results. However, this is not the case, because we have not compared like with like: scores greater than zero on the self-positive trial-type represent a bias toward

confirming positive attributes, whereas scores greater than zero on the others-positive trial-type represent a bias towards *refuting* positive attributes. In other words, scores in the same direction (greater than zero) reflect positive bias for self, but a negative bias for others. As such, *D* scores greater than zero on the vertical axis in Figure 1 (upper panel) are not united by a common interpretation, such as representing a bias towards confirming (rather than rejecting) positive attributes.

In order to appropriately conduct an analysis that involves the comparison of, or interaction between, IRAP trial-types, one must therefore first invert half of the IRAP trial-types (i.e., multiply by -1) so that *D* scores greater than zero on the trial-types of interest share a common interpretation – for example, representing a bias towards confirming (rather than rejecting) positive attributes. Doing so will also allow for intuitive interpretation of IRAP data, due to a common vertical axis. Which trial-types ought to be inverted depends on the IRAP stimulus set and analytic question, as will be discussed below. Although this manipulation is commonly performed within published articles, its details are often limited to a single line in the results section (e.g., Hussey & Barnes-Holmes, 2012). This has proven to be somewhat opaque to researchers who are not familiar with the procedure.

In order to explain this manipulation, let us take a step back to consider the relationship between stimulus categories in the IRAP. Specifically, research with the IRAP has typically employed label and target categories that are opposite from one another along a meaningful dimension. For example, “self” and “others” (labels) as opposed to “self” and “bananas”, or “positive” and “negative” (targets) as opposed to “positive” and “tasty”. Which trial-types are to be inverted depends, however, on the specific nature of the relationship between label and target stimuli (e.g., self and positive). In general, the reversals are conducted such that *D* scores above zero represent biases towards positivity for both labels 1 and 2. Thus, in the self-esteem IRAP example, inverting trial-types 3 and 4 yields a

meaningful vertical axis across all four trial-types (see Table 4). That is, *D* scores above zero for self and for others represent positive or not-negative effects.

Table 5. *Interpretation of D scores greater than and less than zero on the Self-Esteem IRAP when the 3rd and 4th trial-types are inverted.*

	Trial-type 1: **Self-positive**	**Trial-type 2: Self-negative**	**Trial-type 3: Others-positive**	**Trial-type 4: Others-negative**
***D* score > 0** ('positive' or 'not-negative' effects)	I am positive	I am not negative	Others are positive	Others are not negative
***D* scores < 0** ('negative' or 'not-positive' effects)	I am not positive	I am negative	Others are not positive	Others are negative

If we apply this trial-type inversion to our hypothetical data, trial-types 3 and 4 change from -0.33 to 0.33, and -0.14 to 0.14, respectively (see Figure 1, lower panel). Now, scores greater than zero represent positive or not-negative effects (see Table 4). A paired *t*-test would now, correctly, show no differences between effects on the self-positive and others-positive trial-types, and we would conclude that participants implicitly evaluate self and others as being equally positive. Furthermore, the results of the IRAP across all four trial-types are now readily interpretable from the plotted data, and can be summarized as effects that indicate that "I am positive" ($D = 0.41$); "I am *not* negative" ($D = 0.29$); "others are positive" ($D = 0.33$); and "others are *not* negative" ($D = 0.14$). We should also note at this point that we are in the habit of speaking about IRAP data that has been inverted in this manner as following a "true-false-true-false" format, because *D* scores greater than zero now represent faster responding for True, False, True, and False on each trial-type, respectively.

Of course, not all IRAPs follow this format. For example, Nicholson and Barnes-Holmes (2012b) employed "Scares me" and "I can approach" as label stimuli categories 1

and 2, respectively, and pictures of spiders and pleasant nature scenes as target stimuli categories 1 and 2, respectively. In cases such as this, we would invert trial-types 2 and 4 to organize the vertical axis around scaring versus approaching. *D* scores greater than zero would then represent "*Does* scare me" or "I *cannot* approach", and *D* scores less than zero would represent "*Does not* scare me" or "I *can* approach".

Finally, not all IRAPs follow the simple format of categories and attributes, such as in our self-esteem IRAP example. Indeed, an additional primary reason for employing an IRAP over another procedures, such as the IAT, is its ability to assess responses to relatively complex relational networks (e.g., involving statements and conditionals: Hussey & Barnes-Holmes, 2012), rather than the relative strengths of category pairings (Gawronski & De Houwer, 2011). As such, the heuristic strategy of conceptualizing labels and target stimuli as categories and attributes cannot always be employed. An understanding of how to interpret trial-types in relation to one another, and when trial-types should be inverted, is therefore key to appropriate analyses of IRAP data. Whichever strategy is employed, researchers should clearly state 1) if trial-types have been inverted, 2) the rationale for this strategy, and 3) the resulting interpretation of a plot's vertical axis. In closing, it is worth emphasizing that researchers should work through the stages of interpretation laid out in this article, with hypothetical data if necessary, before they start collecting data. Doing so can help refine stimulus sets and highlight potential problems early on, thus minimizing time spent hand wringing and cursing, having invested considerable time and resources on data collection only to find that the resultant IRAP effects are difficult if not impossible to interpret or understand.

References

Barnes-Holmes, D. (2014). Implicit Relational Assessment Procedure (Version 2008, 2010, 2012, 2014). Maynooth, Ireland. Retrieved from IRAPresearch.org

Barnes-Holmes, D., Barnes-Holmes, Y., & Hussey, I. (in press). Relational Frame Theory: Finding its historical and intellectual roots and reflecting upon its future development. In S. C. Hayes, D. Barnes-Holmes, R. D. Zettle, & A. Biglan (Eds.), *Handbook of contextual behavioral science*. New York, NY: Wiley-Blackwell.

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., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes toward meat and vegetables in vegetarians and meat-eaters. *The Psychological Record*, *60*, 287–306.

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

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–193). New York, NY: Guildford Press.

De Houwer, J., & Moors, A. (2012). How to define and examine implicit processes. In *Psychology of Science: Implicit and Explicit Processes* (pp. 183–198). Oxford: Oxford University Press.

Gawronski, B., & De Houwer, J. (2011). Implicit measures in social and personality psychology. In *Handbook of research methods in social and personality psychology* (Vol. 2). New York, NY: Cambridge University Press.

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.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. doi:10.1037/0022-3514.85.2.197

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. New York: Plenum Press.

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual Behavioral Science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*, *1*(1-2), 1–16. doi:10.1016/j.jcbs.2012.09.004

Hughes, S., & Barnes-Holmes, D. (in press). Relational Frame Theory: The Basic Account. In S. C. Hayes, D. Barnes-Holmes, R. D. Zettle, & A. Biglan (Eds.), *Handbook of contextual behavioral science*. New York, NY: Wiley-Blackwell.

Hughes, S., Barnes-Holmes, D., & Vahey, N. (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. doi:10.1016/j.jcbs.2012.09.003

Hussey, I., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice*, *19*(4), 573–582. doi:10.1016/j.cbpra.2012.03.002

Kelly, A., & Barnes-Holmes, D. (2013). Implicit attitudes towards children with autism versus normally developing children as predictors of professional burnout and psychopathology. *Research in Developmental Disabilities*, *34*(1), 17–28.

Kishita, N., Takashi, M., Ohtsuki, T., & Barnes-Holmes, D. (2014). Measuring the Effect of Cognitive Defusion using the Implicit Relational Assessment Procedure: An Experimental Analysis with a Highly Socially Anxious Sample. *Journal of Contextual Behavioral Science*, *3*, 8–15. doi:10.1016/j.jcbs.2013.12.001

Milne, R., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2005). *Measures of attitudes to autism using the IAT and IRAP*. Presented at the 31st Annual Convention of the Association for Behavior Analysis, Chicago, IL, USA.

Nicholson, E., & Barnes-Holmes, D. (2012a). 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. doi:10.1016/j.jbtep.2012.02.001

Nicholson, E., & Barnes-Holmes, D. (2012b). The Implicit Relational Assessment Procedure (IRAP) as a Measure of Spider Fear. *The Psychological Record*, *62*(2), 263–278.

Nosek, B. A., Bar-Anan, Y., Sriram, N., Axt, J., & Greenwald, A. G. (2014). Understanding and Using the Brief Implicit Association Test: Recommended Scoring Procedures. *PLoS ONE*, *9*(12), e110938. doi:10.1371/journal.pone.0110938

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(4), 510–532.

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.

Vahey, N. A., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2010). A first test of the Implicit Relational Assessment Procedure as a measure of self-esteem: Irish prisoner groups and university students. *The Psychological Record*, *59*(3), 4.

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. doi:10.1016/j.jbtep.2015.01.004

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