The PI_{IRAP}:

An alternative scoring algorithm for the IRAP using a probabilistic semiparametric effect size measure

Maarten De Schryver[1], Ian Hussey[1], Jan De Neve[1], Aoife Cartwright[2],

& Dermot Barnes-Holmes[1]

[1]Ghent University, Belgium

[2]Maynooth University, Ireland.

Author note

This research was supported by grant BOF09/01M00209 of Ghent University to Jan De Houwer and an Odysseus Type 1 award (2015-2010) from the FWO to Dermot Barnes-Holmes. Correspondence concerning this paper should be addressed to Maarten De Schryver, Department of Experimental-Clinical and Health Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent (Belgium), Email: Maarten.DeSchryver@ugent.be, Phone: +32-(0)92649107.

Abstract

The Implicit Relational Assessment Procedure (IRAP) has been used to assess the probability of arbitrarily applicable relational responding or as an indirect measure of implicit attitudes. To date, IRAP effects have commonly been quantified using the D_{IRAP} scoring algorithm, which was derived from Greenwald, Nosek and Banaji's (2003) *D* effect size measure. In the article, we highlight the difference between an effect size measure and a scoring algorithm, discuss the drawbacks associated with *D*, and propose an alternative: a probabilistic, semiparametric measure referred to as the Probabilistic Index (Thas, De Neve, Clement, & Ottoy, 2012). Using a relatively large IRAP dataset, we demonstrate how the PI is more robust to the influence of outliers and skew (which are typical of reaction time data). Finally, we conclude that PI models, in addition to producing point estimate scores, can also provide confidence intervals, significance tests, and afford the possibility to include covariates, all of which may aid single subject design studies.

Keywords: implicit measure; IRAP; effect size; D score; probabilistic index model

The first purpose of the current paper is to consider the relative benefits of effect size measures when scoring data from the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010), a reaction time task that is frequently employed within research related to Relational Frame Theory (RFT: Hayes, Barnes-Holmes, & Roche, 2001). The second purpose is to propose a probabilistic, semiparametric measure referred to as the Probabilistic Index (Thas, De Neve, Clement, & Ottoy 2012) for use with the IRAP, which appears to provide some advantages over the currently most widely used measure (i.e., the *D*-IRAP score). On balance, it is not our intention to argue that the PI is a replacement for the *D* score, but provides an alternative measure that may be particularly useful when the response-time distributions are skewed and the potential for outliers and extreme scores are present. Before proceeding, however, it seems important to provide a brief overview of the IRAP, as an RFT-based methodology, to contextualize the current work.

Historical and conceptual background to the IRAP

Much of the early research in RFT consisted of demonstration studies to test the theory’s basic assumptions and core ideas. One of the defining features of this research was a dichotomous approach to arbitrarily applicable relational responding (AARRing), which is a central idea within the account of human language and cognition provided by RFT (see Hughes & Barnes-Holmes, 2016, for an accessible overview). That is, laboratory studies in RFT often focused on showing that particular patterns of AARRing were either present or absent. Within a few years of the publication of the 2001 RFT book (Hayes et al., 2001), however, the need to develop procedures that could, in principle, provide a measure of AARRing that was non-dichotomous became increasingly apparent. The initial response to this need was the development of what came to be known as the IRAP. Specifically, the IRAP was a response to the question, “How can we capture relational frames in flight”, which essentially is a question about the relative strength of AARRing in the natural environment.

In developing the IRAP, two separate methodologies were combined. The first of these was an RFT-based procedure for training and testing multiple stimulus relations, the Relational Evaluation Procedure (REP: Cullinan, Barnes, & Smeets, 1998) and the second was the Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998). The REP presents participants with two stimuli and requires them to provide a relational response (e.g., “same” or “different”). The IAT was developed by social-cognition researchers as a method for measuring what are frequently conceptualized as associative strengths in memory by comparing the relative speed of categorization of stimuli. The IRAP combined features from these two tasks by requiring participants to provide one relational response in some blocks (e,g., “same”) and another in other blocks (e.g., “different”), and comparing the relative speed of relational responding between block types. The IRAP was therefore conceptualized as a procedure for measuring the relative strength of AARRing in a non-dichotomous manner (see Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008).

It has been argued that due to its close connection to the IAT research with the IRAP quickly became dominated by studies focused on so-called implicit attitudes and implicit cognition more generally (Barnes-Holmes, Barnes-Holmes, Hussey, & Luciano, 2016). On the one hand, this strategy was very useful because it provided a means by which to assess the validity of the IRAP as a measure of natural verbal relations (see Vahey, Nicholson, & Barnes-Holmes, 2015). On the other hand, it also served as a distraction from a focus on RFT and AARRing *per se* (Barnes-Holmes, Barnes-Holmes, Barnes-Holmes, Luciano, & McEnteggart, 2017). Furthermore, the historical connection between the IRAP and IAT was instrumental in developing a version of the IAT D_1 score, which is used to analyze the response latency data from the IAT. The IRAP version, the D_{IRAP} algorithm, is described later in the current article, but as was pointed out by Barnes-Holmes et al. (2010) the D_{IRAP} algorithm should not be seen as prescriptive or necessarily the "best way" to analyze IRAP data." (p.533). Consistent with this view, and the ongoing development of the IRAP as an RFT-based method for analyzing human language and cognition, the current article presents another algorithm for analyzing IRAP data that appears to offer a number of advantages over the D_{IRAP} algorithm.

A brief description of the IRAP

The IRAP is a computer-based task on which an individual responds to a series of trials, each of which usually presents pairs of stimuli on screen (although see Kavanagh, Hussey, McEnteggart, Barnes-Holmes, & Barnes-Holmes, 2016, for an alternative format using natural language statements). To illustrate, we use an IRAP that was designed to assess gender stereotypes. (see Cartwright, Hussey, Roche, Dunne, & Murphy, 2017 for similar IRAPs and discussion of the topic). Subsequently, data collected using this IRAP will then be presented. On each trial a label stimulus appears at the top of the screen, such as either "Men are" or "Women are". Target stimuli appear in the middle of the screen, such as stereotypically masculine traits (witty, competitive, decisive, and charismatic) or feminine traits (nurturing, gentle, affectionate, and sensitive). Two response options are also provided on each trial, such as "true" and "false". The IRAP operates by requiring opposite patterns of responding across successive blocks of trials. For example, "men are-masculine" trials would require participants to respond with "true" on one block and "false" on the next block. If the correct response is emitted the task simply continues to the next trial, but if the incorrect response is emitted a red X appears on screen and the next trial is not presented until the correct response option is provided. The IRAP thus involves presenting four trial types within each block and participants are required to emit opposing patterns of responding across successive blocks of trials. The four trial-types for the example of the IRAP described above may be summarized as: men-masculine, men-feminine, women-masculine, and women-feminine. For half of the blocks of trials, participants would be required to respond as if men are masculine and women are feminine (consistent trials[1]; i.e., men-masculine/true; men-feminine/false; women-masculine/false; women-masculine/true), and for the remaining blocks to respond as if men are feminine and women are masculine (inconsistent trials; i.e., men-masculine/false; men-feminine/true; women-masculine/true; women-masculine/false). Finally, it is worth noting that the IRAP typically involves allowing the participant to complete a number of pairs of consistent and inconsistent blocks until they reach mastery criteria (e.g., for each block in a pair, median latency < 2000 ms and accuracy > 80%), followed by a static number of test blocks pairs (usually 3) from which data are analyzed. This was the case for the gender IRAP dataset used in the current paper (for paper length discussions of the task see Barnes-Holmes et al., 2010; Hussey, Thompson, McEnteggart, Barnes-Holmes & Barnes-Holmes, 2015).

Broadly speaking, the IRAP is usually scored by subtracting the mean response latency for one pattern of responding from the mean response latency of the opposite pattern of responding; the difference score is typically normalized (i.e., the D_{IRAP} algorithm). The difference score thus reflects a response bias in one direction or the other, such as responding “True” more quickly than “False” across blocks of trials when presented with the men-masculine trial-type. Specific response biases are usually predicted based on the behavioral

[1] We employ the terms “consistent” and “inconsistent” here based on their usage in the literature. However, we recognize that these terms are potentially confusing. Given that Barnes-Holmes et al. (2010) stated that the faster/more probable response is by definition consistent with an individual’s learning history, one could argue that a block of trials should only be designated as “consistent” for that individual after the fact (i.e., as an outcome) based on which block was faster, rather than a priori. As such, consistent and inconsistent are typically used in two ways: to note the (in)congruence between an individual’s learning history and which block produces faster RTs, and as a label to differentiate the two types of blocks within the IRAP as the researcher sees them (likely influenced by social-normative expectations). These meanings will not always overlap, leading to occasional confusion. Here, we employ the latter sense of the words. For this reason, some researchers have referred to the blocks using arbitrary designations such as “A” and “B” (e.g., Hussey, Barnes-Holmes, & Booth, 2016) or as “pro” versus “anti” the domain of interest targeted by the IRAP (as in pro- and anti-spider; e.g., Nicholson & Barnes-Holmes, 2012)

histories of the participants. In the case of the current example, participants who report strongly gender-stereotypical biases would, for example, be expected to produce a larger men-masculine response bias. To put it simply, the original basic hypothesis behind the IRAP is that, all things being equal, mean response latencies should be shorter across blocks of trials that are consistent with a participant's behavioral history relative to those blocks of trials that require responses that are inconsistent with that history (see Barnes-Holmes, Finn, McEnteggart, & Barnes-Holmes, in press, for a recent and more sophisticated approach to explaining IRAP effects). In what follows, we will explain why the focus on the mean or average latency, which are used to calculate D_{IRAP} scores, may be problematic and outline an another analytic method for the IRAP. We will start by considering the general issue of scoring reaction time measures, and making an important distinction between the concept of a "scoring algorithm" and an "effect size measure".

Scoring reaction time measures

Since the introduction of the IRAP, IRAP scores have most frequently been calculated using the D_{IRAP} scoring algorithm. A scoring algorithm typically contains a set of consecutive steps that a researcher follows in order to obtain a final score, or scores, for each participant. For instance, a scoring algorithm might specify which trials should be taken into account, how to treat errors, how to treat response latencies that are deemed to be excessively short or long, and how to calculate the final score(s). Calculating the final score(s) usually involves adopting a particular effect size measure, which may be defined as the mathematical formula used to calculate the quantity reflecting the magnitude of the difference in performance between conditions (e.g., between blocks of consistent and inconsistent trials). A specific scoring algorithm thus includes the type of effect size measure that should be used to obtain the score, or scores, in addition to other steps such as data exclusions. For instance, the D_{IRAP}

scoring algorithm requires that all RTs > 10000 ms are discarded and then, for each trial-type for each pair of consistent and inconsistent blocks, a *D* effect size measure should be calculated. The means of the obtained *D* scores (calculated across the block pairs) serves as the final D_{IRAP} score for each trial type. Other D_{IRAP} scoring algorithms could propose similar steps, but include more stringent exclusion criteria than simply removing RTs > 10000ms, such as removing entire data sets for participants who failed to maintain a mean response latency < 2000ms on one or more of the four trial-types. Of course, other scoring algorithms could employ a different effect size measure. In the current article we propose one such measure: a semi-parametric probabilistic index.

Before proceeding, it should be noted that the decision to employ a particular effect size measure may be based in part on the stringency of the exclusion criteria. For example, if particularly stringent exclusion criteria are adopted for removing relatively long latencies, using an effect size measure that aims to reduce the impact of such latencies may be of little benefit. The basis for deciding how stringent the exclusion criteria should be is a highly complex issue, a discussion of which is beyond the scope of the current article. For present purposes, we will focus on a situation in which the exclusion criterion for response latencies is relatively relaxed (i.e., > 10000ms). As we shall see, when such a relaxed criterion is adopted, relatively long response latencies (e.g., lying somewhere between the mean response latency plus 2.5 standard deviations and 10000ms) may introduce unwanted "noise" into the dataset and thus it may be wise to adopt an effect size measure that will reduce the effects of such "noise"[2]. In what follows, we will begin by providing a concise overview of different types of effect size measures and discuss how appropriate these are to answer the main question: "is an individual faster (or slower) to respond on consistent trials compared to inconsistent trials?"

[2] We intentionally use quotes to indicate that extreme values in the context of some research questions may not be considered "noise" but constitute data points that have important theoretical significance, and should not, therefore, be removed from the dataset.

Non-standardized effect size measures

Non-standardized effect size measures summarize differences between distributions in the same unit as the unit of measurement. For instance, we can estimate the mean response time (RT, in milliseconds, ms) across consistent trials and the mean RT (also in ms) across inconsistent trials. The difference between these two means expresses the differences in ms between the “typical” RT for each block of consistent and inconsistent trials. Because RT distributions are typically right-skewed, other measures of central tendency are sometimes used, such as the median. Another way to deal with skewness is by transforming RTs using a log-, square root- and/or reciprocal transformation (e.g., the *C*-measures originally used for the IAT: Greenwald et al., 1998).

Non-standardized effect size measures have the advantage that they are easy to interpret, particularly when the unit of measurement (e.g., such as the RT rather than a difference score) is important. A serious limitation, however, is that effects tend to correlate with general responding speed (GRS); that is, participants responding slower during a task show typically larger effects compared to participants with faster responses (Fazio, 1990; Faust, Balota, Spieler, & Ferraro, 1999; Greenwald et al., 2003). Relatedly, O’Toole and Barnes-Holmes (2009) reported that raw latency and difference scores from the IRAP correlated with intelligence scores. This makes it difficult, or even impossible, to make meaningful comparisons of non-standardized effect sizes among participants, and even among different experiments.

Standardized effect size measures

Standardized effect size measures may be seen as “canceling out” the unit of measurement. Perhaps the most well-known of these is the standardized mean difference, or Cohen’s *d*. Because the difference between the means is divided by the pooled standard deviation, Cohen's *d* can be interpreted as the difference relative to the variability on RTs between conditions[3].

In their original article, Greenwald and colleagues. (2003, p. 201) explicitly draw a link between the *D* effect size measure and Cohen's *d*, both in terms of their calculation and interpretation (Cohen, 1988). However, it is important to note that *D* is, in actuality, mathematically more comparable to a different standardized effect size measure: the point-biserial correlation (r_{pb}) coefficient (see Ruscio, 2008). As will be discussed below, this categorization may be important when considering the disadvantages associated with different classes of effect size measures. Point-biserial correlations are expressed as the correlation between the RTs of both condition and a dummy variable indicting to which condition the RT belongs (0 for consistent responses; 1 for inconsistent responses). If the number of trials are equal in both conditions ($n_1 = n_2$), the *D* can be considered as a scaled point-biserial correlation coefficient*:* $r_{pb} = D \times \sqrt{\frac{n_1 n_2}{N^2 - N}}$, with $N = n_1 + n_2$. In contrast to Cohen's *d*, the *D* effect size measure is obtained by dividing the difference of the means by the standard deviation of the pooled sample of RTs (i.e., the standard deviation of RTs independent of condition; in the case of the IRAP, this means calculating single standard deviations across pairs of blocks of consistent and inconsistent trials). Broadly speaking, as measures of effect size, both *d* and r_{pb} can both be interpreted as a signal to noise ratio (i.e., in the differences in mean RTs between consistent and inconsistent blocks proportionate to the variance in those RTs).

Both *d* and r_{pb} effect size measures are popular, and both appear to reduce the unwanted correlation between effect sizes and GRS. However, they also have disadvantages

[3] Note, that some researchers divide the difference between the means by the standard deviation of a reference group. This standardized effect size measure is known as Glass d (d_{G}) and should not be confused by Cohen's *d*.

in common (see Ruscio, 2008). First, r_{pb} seems to be sensitive to base rates. That is, when the number of trials substantially differs among conditions, r_{pb} will decrease in value if the difference in number of trials increases. In those cases where heterogeneity exists, the same observation is made for Cohen's *d*. Second, both measures are sensitive to violations of parametric assumptions such as normality and heterogeneity of variances. For example, Cliff (1993) argued that it is not guaranteed that if the mean of a first distribution is larger than the mean of a second distribution, the majority of scores of the first distribution will be larger compared to the second one. For instance, this could be the case when the mode of a right heavy-tailed distribution is smaller compared to the mode of another (see also McGraw & Wong, 1992). Nonlinear data transformations, such as log-, square root-, and/or reciprocal transformations, have often been suggested as a way to correct for the non-normal distribution of RT data. Third, both Cohen's *d* and r_{pb} are sensitive to nonlinear transformations of the data. As such, different values might be obtained when one of the aforementioned transformations was used prior to the calculation of effect sizes. Fourth, both effect size measures are very sensitive to the presence of outliers. Finally, r_{pb} has the disadvantage that it is difficult to interpret, especially for non-experts.

At this point, let us focus on the potential impact of outliers and how we might deal with them in a set of IRAP data (as noted earlier, RT distributions are typically skewed to the right and this makes it difficult to decide which observations may be considered outliers). The data were taken from an unpublished IRAP study that was designed to examine gender stereotyping. In our sample ($N = 188$), we observed that 85% (81%) of the distributions of the consistent (inconsistent) trials show a skew to the right (here, we defined skewness when the Pearson's moment coefficient of skewness > 1.00). To test if the variances of the consistent block significantly (with alpha set to .05) differ from the variance of the inconsistent block, we performed F-ratio tests. Our results suggest that in 45% of the cases, unequal variances were observed (i.e., heterogeneity of variances). Next, we counted the number of outliers for each participant and for each type of block (i.e., consistent versus inconsistent). Here, any observation that differs at least 2.5 standard deviations from the mean (within participant, within block) is considered an outlier[4]. Our results show that at least one trial out of 24 trials can be identified as an outlier in 85% (79%) of the participants for the consistent (inconsistent) trials. In total, 3.8% of the trials were considered as outliers[5].

From these results, and in light of the aforementioned limitation of Cohen's *d* and r_{pb}/D effect size measures, it should be clear that there are potential problems in using either as effect size measures to reflect the difference between conditions. To illustrate, for each participant we calculated a *D* score based on the full data set (D_{full}) and a second *D* score based on the set of trials after excluding outliers (D_{excl}). That is, we implemented the *D* effect size measure within two different scoring algorithms that differed in how they deal with outliers. For illustrative purposes, we will only consider the data from one of the IRAP's four trial-types, the men-masculine trial type (and not men-feminine, women-masculine, or women-feminine trial types). Although D_{full} and D_{excl} correlate highly ($r = .93$), the mean absolute difference between these scores was found to equal $M = .12$, with $SD = .11$. As illustrated in Figure 1a, deviations between D_{full} and D_{excl} can be as extreme as .53. From this figure, we observe that for 5% of participants in our sample, the deviation between the two versions of *D* was larger than the standard deviation of the D_{full} scores in the sample (SD = .34). That is, the data points for 10 participants fell outside the dotted lines on the graph. Outlier data points therefore may have an unwanted influence on the scored data. This is especially problematic given that extremity is defined by arbitrary rules that may differ among researchers.

[4] We fully recognize that 2.5 SD is an arbitrary choice to define outliers (as is excluding RTs > 10000 ms in the D_{IRAP} algorithm). However, this is a common practice to detect outliers in psychological research (Leys, Ley, Klein, Bernard, & Likata, 2013). Importantly, our goal here is just to illustrate the presence of "extreme observations".

[5] Under normality and using the 2.5 SD criterion, 1.24% of the trials would be identified as outliers.

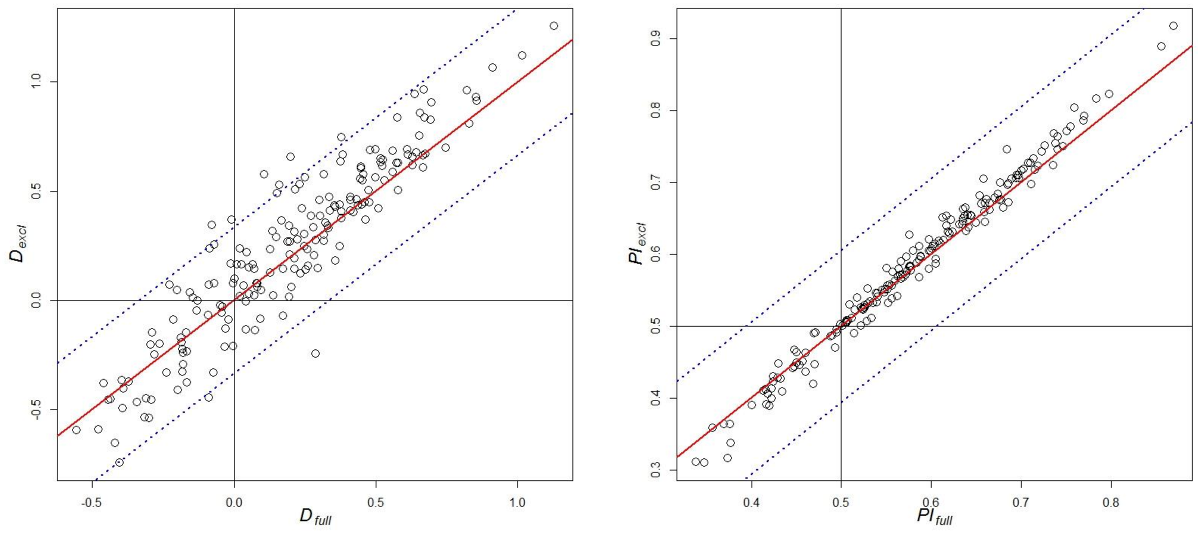

Figure 1. Left Panel: Figure 1a; Scatterplot of the obtained *D* scores using the full dataset (D_{full}) or after excluding outliers (D_{excl}). Right Panel: Figure 1b; Scatterplot of the obtained PI scores using the full dataset (PI_{full}) or after excluding outliers (PI_{excl}). Solid lines represent equal scores, dotted lines represent plus or minus one standard deviation of the sample scores (using the full datasets).

The PI: An alternative standardized effect size measure

Thas, et al., (2012) recently introduced a new class of semiparametric regression models called Probabilistic Index Models (PIMs). In the current context, a Probabilistic Index (PI) can be interpreted as the probability that a randomly selected inconsistent trial has a larger RT than a randomly selected consistent trial. As an expression of probability, the PI can therefore range between 0 and 1, where; 0 would refer to situations where RTs for all consistent trials were faster than all inconsistent trials, 1 refers to situations where RTs for all inconsistent trials are faster than all consistent trials, and 0.50 refers to situations where there is no systematic difference between the two. Importantly, the PI treats data as ordinal rather than interval, thus “faster” here refers to the fact that one reaction time (e.g., 1000) is (simply) “faster” than another (e.g., 1100), rather than being “faster by 100 ms”. This is the key difference from other effect size measures that serves to minimize the influence of outliers.

In the context of the IRAP, the PI can be calculated in an easy way that immediately illustrates its interpretation (see Table 1); the reader is referred to Appendix 1 for a mathematical definition of the PI and its application to the IRAP. Suppose we observed three RTs related to consistent trials (500, 600, 700) and three RTs related to inconsistent trials (550, 650, 750). By creating the set of “pseudo-observations” (i.e. all possible pairs between consistent and inconsistent trials, in this case 3x3 = 9), we count the number of RTs faster for consistent trials compared to inconsistent trials. In this example, there are 6 pairs for which the RTs of consistent trials are faster (thus smaller). Dividing this sum by the total number of comparisons, it follows that PI = 6/9 = 0.67. We would therefore conclude that the probability that inconsistent trials have larger RTs than consistent trials was 0.67. As such, we would reformulate the original basic hypothesis of the IRAP as the probability of observing faster reaction times on trials that are consistent with a participant’s behavioral history when compared to reaction times on trials that are inconsistent with that history.

Table 1. Calculation of the PI in a simple setting using a set of pseudo-observations for three consistent trials (500, 600, 700) and three inconsistent trials (550, 650, 750).

RT Consistent	RT Inconsistent	Inconsistent > Consistent? (if Yes = 1; If No = 0)
500	550	1
500	650	1
500	750	1
600	550	1
600	650	0
600	750	0
700	550	1
700	650	1
700	750	0
		Sum = 6 Number of pairs = 9 **PI = 6/9 = 0.67**

To illustrate the impact of using the PI, instead of the D score, with the IRAP we calculated the PI for each participant from the gender stereotyping dataset. Here again, the PI is calculated only for the men-masculine trial type. We calculated scores from both the full data set (PI_{full}) and from the dataset after removing outliers using the same criteria as before

(PI_{excl}). Note that the latter is included only to explore the influence of outliers on the PI, and not as a recommendation that outliers should generally be defined and excluded when calculating the PI. As illustrated in Figure 1b, the mean absolute difference between these score equals *M* = 0.014 with *SD* = 0.012. More specifically, this graph illustrates that, for the PI, there were no participants whose PI_{full} and PI_{excl} scores deviated from the regression line (maximum deviation = 0.062) by more than the standard deviation of the PI_{full} score (*SD* = 0.105). Additionally, PI_{full} and PI_{excl} were found to correlate almost perfectly (*r* = .99), which was significantly higher than the correlation between D_{full} and D_{excl} (*r* = .93, r_{dif} = 0.06, 95%CI = [.05, .09]). As such, the PI was demonstrated to be less influenced by outlier data - and thus the arbitrariness of the rules to define outliers - than *D*. As we have previously discussed, this may be particularly important when working with reaction time data, in which outliers are very common.

Lastly, to assess the presence of a linear relation between scores and general responding speed (GRS), which is, itself, often correlated with spurious variables such as age, we calculated the correlation between the absolute mean difference between consistent and inconsistent blocks and GRS (*r* = .55), D_{full} and GRS (*r* = .01) and PI_{full} and GRS (*r* = .01). Clearly, both D_{full} and PI_{full} reduce the unwanted correlation between the non-standardized effect size measures and the general response speed, as is typically desirable. In sum, the preceding analyses therefore suggest that the PI effect size measure is more robust to outlier data than is the *D* score.

An additional scoring algorithm for the IRAP: PI_{IRAP}

In the previous section we introduced an alternative standardized effect size measure for expressing the difference in performance on consistent and inconsistent trials. In this section, we will propose a new scoring algorithm making use of the PI. Readers familiar with the classic D_{IRAP} scoring algorithm will notice that the PI_{IRAP} scoring algorithm does not differ in many aspects from previously proposed measures (see for instance Barnes-Holmes et al., 2010; Hussey et al., 2015). For example, we chose to calculate one $\mathrm{PI}_{\mathrm{IRAP}}$ score for each pair of test blocks and then combine them (i.e., rather than calculate a single $\mathrm{PI}_{\mathrm{IRAP}}$ using all consistent blocks vs all inconsistent blocks). This is purposeful, given that the primary aim of the current article is to consider an additional effect size measure rather than its implementation within particular scoring algorithms, which is beyond the scope of the current article. We propose the following steps in calculating the $\mathrm{PI}_{\mathrm{IRAP}}$ scores: (1) Use only RTs from test blocks; (2) Remove those participants with at least 10% response latencies faster than 300ms; (3) Calculate for each participant four $\mathrm{PI}_{\mathrm{IRAP}}$ scores, one $\mathrm{PI}_{\mathrm{IRAP}}$ for each of the four trial types (e.g., men-masculine, men-feminine, women-masculine, and women-feminine). Each $\mathrm{PI}_{\mathrm{IRAP}}$ is calculated by defining a set of pseudo-observations, conditional on the block pairs. That is, all consistent trials are compared with all inconsistent trials from the same pair of test blocks. The final set combines these three different sets of pseudo-observations of the three pairs of consistent and inconsistent test blocks into one single set, and $\mathrm{P}(Y < Y' \mid X = 0, X' = 1)$, with X = 0 for consistent trials and X = 1 for inconsistent trials, is calculated. Note that we do not exclude any outlier observations prior to the calculations.

To illustrate a number of points regarding the two algorithms, the D_{IRAP} scores and $\mathrm{PI}_{\mathrm{IRAP}}$ scores obtained for the men-masculine trial type from all participants in the dataset are presented in Figure 2. To take some specific examples, a $\mathrm{PI}_{\mathrm{IRAP}}$ = 0.75 is obtained for the men-masculine trial type for participant 1. Thus, when selecting a random consistent trial and a random inconsistent trial, it is more likely that the RTs from the inconsistent trial is larger (probability = 0.75). Participant 1's D_{IRAP} score = 0.82. Thus, the ratio between the difference in average RTs between the blocks and the variance of the pooled reaction times was 0.82. Participant 3 represents a second example, whose $\mathrm{PI}_{\mathrm{IRAP}}$ score for the men-masculine trial type = 0.32. For this participant, it is more likely that larger RTs are observed for consistent trials compared to inconsistent trials (probability = 0.68, i.e., 1.0 – 0.32). This participants D_{IRAP} score = -0.42, which is the ratio between the difference in average RTs between the blocks and the variance of the pooled reaction times.

Interestingly, the direction of the scores sometimes differs. For instance, for participant 19 it is more likely that the RTs from the inconsistent trial is larger (probability = 0.65), while the difference in average RTs between blocks is negative (D_{IRAP} score = -0.21, indicating a larger average RT for consistent trials compared to the average RT for inconsistent trials). At this point, it should also be apparent that the interpretation of the PI_{IRAP} score is therefore clearer than that of the D_{IRAP} score in terms of our original question regarding probabilistic responding speeds.

Although the direction and magnitude of individual scores might differ depending on which effect size measure is used, we do not expect large differences in the patterns observed at the group level. For instance, although substantial differences between D_{IRAP} scores and PI_{IRAP} scores are observed, for the current data set the two scores correlate highly ($r = .88$).

Nevertheless, due to its relative insensitivity to outlier data, the PI_{IRAP} should also demonstrate higher internal reliability. We therefore split our dataset in two halves (subsets) based on an odd/even (trial index) split and calculated Cronbach's alpha.[6] A modest improvement is observed for the PI_{IRAP} scoring algorithm ($\alpha_{\text{PI}} = .37$) compared to the D_{IRAP} scoring algorithm ($\alpha_D = .29$).

Finally, it is also useful to note that quantifying data using a PIM allows for the calculation of more than just a single point estimate (i.e., PI_{IRAP}). For example, we can easily obtain a 95% confidence interval for each individual score. For instance, a $\text{PI}_{\text{IRAP}} = 0.75$ is

[6] Spearman-Brown corrections, rather than Cronbach's alpha, are sometimes reported in IRAP publications, but essentially they yield the same results (Bentler, 2009).

obtained for the men-masculine trial type for participant 1. Here, the 95% confidence interval is given by [0.59, 0.86]. Additionally, the PIM allows us to test an alternative hypothesis $\mathrm{H_a}: \mathrm{PI_{IRAP}} \neq 50\%$ (i.e., presence of an IRAP effect) against a null hypothesis $\mathrm{H_0}: \mathrm{PI_{IRAP}} = 50\%$ (i.e., no IRAP effect). In this case, we can reject the null hypothesis with $p = 0.003$. That is, one can easily determine, without the need to simulate (as would be required with the D_{IRAP} score), whether individual participants produced statistically significant $\mathrm{PI_{IRAP}}$ effects, as well as the magnitude and confidence interval of these effects. This may be useful for, among other things, single case designs. Relatedly, PIMs also allow the researcher to add covariates to the PI formula. Among other things, this may be useful for examining the influence of specific stimuli on the IRAP effect.

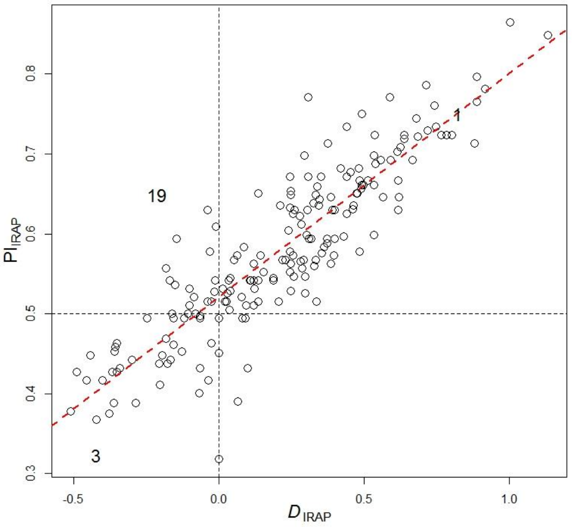

Figure 2. Scatterplot of the D_{IRAP} and $\mathrm{PI_{IRAP}}$ scores respectively for the entire sample. Three data points are highlighted for illustration using their participant numbers rather than a circle: "1" indicates the scores for Participant 1 and "3" indicates the scores for Participant 3, and "19" indicates the scores for Participant 19. The linear trend between D_{IRAP} and $\mathrm{PI_{IRAP}}$ is illustrated by a regression line.

Discussion and conclusion

In this article, we have introduced the PI as an alternative effect size measure to the frequently used *D* effect size measure, and then implemented it within (one possible version) of the $\mathrm{PI}_{\mathrm{IRAP}}$ scoring algorithm for use with the IRAP. Although $\mathrm{PI}_{\mathrm{IRAP}}$ and D_{IRAP} share similar steps, they do differ at their core: the proposed effect size measure. While the D_{IRAP} scoring algorithm defines a scaled point-biserial correlation coefficient to reflect the difference in reaction times between consistent and inconsistent trials as a proportion of the variance in all reaction times, the $\mathrm{PI}_{\mathrm{IRAP}}$ expresses this difference in performance as the probability that reaction times are higher in one context (inconsistent blocks) relative to another (consistent blocks).

As illustrated, reaction time data tends to be both heavily skewed to the right and also include outliers. Statistics, such as the sample mean and sample variance are sensitive to outliers, or as formulated by Greenwald and colleagues (1998) they "distort means and inflate variances" (p.1467). Using the sample mean and the sample variance might not be the best option, even when response latencies larger than 10000ms are omitted. In contrast, the proposed PIM-framework is much more robust to deviations from normality and to outliers. By analyzing the men-masculine trial type of the gender IRAP dataset, we have shown that (1) Substantial differences were observed between individual D_{IRAP} scores and $\mathrm{PI}_{\mathrm{IRAP}}$ scores; (2) A high correlation between D_{IRAP} and $\mathrm{PI}_{\mathrm{IRAP}}$ scores of the entire sample was obtained; and (3) A (moderately) higher reliability estimate was recorded.

In order to aid researchers in implementing the PI generally and $\mathrm{PI}_{\mathrm{IRAP}}$ more specifically, we have included R code for a minimal implementation of the PI in Appendix II. Additionally, we produced an R Shiny web app that researchers can use to calculate $\mathrm{PI}_{\mathrm{IRAP}}$ scores, which can be accessed at http://datapp.ugent.be/shiny/irap/. The source code for this app, and all code employed within the current article, can also be found on the Open Science Framework (http://osf.io/4cmsm).

We recognize that other measures for calculating effects for reaction time measures that are robust to heterogeneity and the presence of outliers have been proposed (see Richetin, Costantini, Perugini, & Schönbrodt, 2015). For instance, the Gaussian rank latency difference, or *G* score, offered by Sriram, Nosek, and Greenwald (2006) is closely related to the PI. However, the two differ with respect to their interpretation. While the PI is a direct probability, *G* scores reflect scores on a Gaussian distribution obtained by transforming fractional ranks. As such, the PI is arguably easier to interpret. In addition, given that the PI is a model-based measure, it allows researchers to calculate $\mathrm{PI}_{\mathrm{IRAP}}$ scores but also confidence intervals, *p* values, and the inclusion of additional covariates (e.g., the impact of specific stimuli). In our opinion, these properties make the $\mathrm{PI}_{\mathrm{IRAP}}$, and PIM models more broadly, an interesting and highly useful choice among effect size measures that may be of use in future research.

We should reiterate that the current article has focused on the choice of effect size measure (i.e., *D* vs. PI), but has not addressed broader questions concerning other aspects of the scoring algorithm beyond the effect size measure. In contrast, Greenwald et al. (2003) made comparisons between six scoring algorithms that employ the *D* effect size measure (i.e., D_1 to D_6), but which adjust other aspects of the algorithm. Future research should therefore compare variations in $\mathrm{PI}_{\mathrm{IRAP}}$ scoring algorithms that implement the PI effect size measure.

Finally, in closing it is important to recognize that it would, of course, be premature to conclude on the basis of one article, which employed only one data set and did not address the issue of predictive validity, that the $\mathrm{PI}_{\mathrm{IRAP}}$ should now be used instead of the D_{IRAP}. Working out the strengths and weaknesses of specific scoring algorithms for RT measures is a complex and difficult task (e.g., Richetin, et al., 2015), and it would be unwise for researchers to adopt the $\mathrm{PI}_{\mathrm{IRAP}}$ scoring method based simply on a “knee-jerk” reaction to the limited set of analyses we have presented here. Furthermore, we would advise against adopting a one-size-fits-all approach to selecting a scoring algorithm. Indeed, this may be particularly important for the IRAP because it has been used in fundamentally different ways (e.g., as a measure of implicit attitudes, as a measure of the relative strength of arbitrarily applicable relational responding, and even as a method for training and testing flexibility in relational responding).

Indeed, one might object to the use of the PI if the magnitude of the effect in milliseconds is of interest and if it is to be used as a proxy for a psychological construct. For example, if latency differentials are seen as a direct mapping onto the relative strength of association between two representations in memory (i.e., longer latency differentials indicating weaker associations), then a dichotomous interpretation of the PI will be problematic. On balance, if a researcher assumes that the probability that scores in one condition are larger than in another condition provides a proxy for the relative strength of association between representations in memory the problem disappears. As such, the preference for or against the PI, in this regard, appear to depend upon quite abstract theoretical assumptions that a researcher wishes to make about the relationship between latency differentials and underlying psychological constructs. In any case, the key purpose of the current work was simply to alert researchers to some of the benefits of the $\mathrm{PI}_{\mathrm{IRAP}}$ relative to the D_{IRAP} effect size measure in dealing with the influence of outliers and skew effects.

Appendix I

Mathematical definition of the PI and its application to the IRAP

The PI is defined as $\mathrm{P}(\mathrm{Y} < \mathrm{Y}')$, where Y and Y' denote two independent responses associated with one or more covariates. In the presence of ties a modified definition of the PI can be used. A PIM models the PI as a function of covariates. In contrast to the previously discussed effect size measures, the PI does not reflect a difference in terms of the mean. To illustrate how PIMs can be used in the context of the IRAP, let us consider the PIM with logit link and expit as the inverse of the logit function ($\mathrm{expit} = \exp(x)\,/[1 + \exp(x)]$): $\mathrm{P}(Y \le Y' \,|X, X') = \mathrm{expit}[\beta_1(\mathrm{X}' - \mathrm{X})]$, where X is a dummy variable, with X = 0 for consistent trials and X = 1 for inconsistent trials. The equation can now be written as: $\mathrm{P}(Y < Y' \,|X = 0, X' = 1) = \mathrm{expit}(\beta_1)$. For a detailed discussion of PIMs, the reader is referred to Thas et al. (2012) and De Neve (2013).

Appendix II

R code for a minimalist implementation of the PI effect size using the pim library.

```
## dependencies
#install.packages("pim")
library(pim)

# acquire data
rt = c(1000, 1001, 2004, 1003, 1004, 2000, 2001, 2002, 2003, 1004)
block_type = as.factor(c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2) )
my_data = data.frame(rt, block_type)

## PI
# 1. returns one overall PI, not separated by trial type
# 2. assumes that my_data is in long format, with the columns rt and block_type

# define pim model
pim_model <- pim(rt ~ block_type,
                data = my_data)

# calculate PI
PI <- plogis(coef(pim_model))

# return PI
PI

# returns
#>0.78
```

References

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I., & Luciano, C. (2016). Relational Frame Theory: Finding its historical and intellectual roots and reflecting upon its future development. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), The Wiley handbook of contextual behavioral science (pp. 129–178). New York, NY: Wiley-Blackwell. http://doi.org/10.1002/9781118489857.ch8

Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From the IRAP and REC model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable responding. *Journal of Contextual Behavioral Science*, http://dx.doi.org/10.1016/j.jcbs.2017.08.001

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*(3), 527.

Barnes-Holmes, D., Finn, M., McEnteggart, C., & Barnes-Holmes, Y. (2017). Derived stimulus relations and their role in a behavior-analytic account of human language and cognition. *The Behavior Analyst.*

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *The Psychological Record*, *58*(4), 497.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137-143.

Cartwright, A., Hussey, I., Roche, B., Dunne, J., & Murphy, C. (2017). An investigation into the relationship between the gender binary and occupational discrimination using the

Implicit Relational Assessment Procedure. *The Psychological Record, 67*(1), 121-130. http://doi.org/10.1007/s40732-016-0212-1

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cullinan, V. A., Barnes, D., & Smeets, P. M. (1998). A precursor to the relational evaluation procedure: Analyzing stimulus equivalence. *The Psychological* Record, 48(1), 121–145.

De Neve, J. (2013). *Probabilistic index models*. Unpublished doctoral dissertation. Ghent University. Faculty of Sciences, Ghent, Belgium.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*, 777-799.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M.S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74-97), Newburry Park, CA: Sage.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (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*, 197-216.

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

Hughes, S., & Barnes-Holmes, D. (2016). Relational Frame Theory: The basic account. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), The Wiley handbook of contextual behavioral science (pp. 129–178). New York, NY: Wiley-Blackwell. http://doi.org/10.1002/9781118489857.ch9

Hussey, I., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). From Relational Frame Theory to implicit attitudes and back again: clarifying the link between RFT and IRAP research. *Current Opinion in Psychology, 2,* 11–15. http://doi.org/10.1016/j.copsyc.2014.12.009

Hussey, I., Barnes-Holmes, D., & Booth, R. (2016). Individuals with current suicidal ideation demonstrate implicit “fearlessness of death.” *Journal of Behavior Therapy and Experimental Psychiatry*, *51*, 1–9. https://doi.org/10.1016/j.jbtep.2015.11.003

Hussey, I., 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

Kavanagh, D., Hussey, I., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2016). Using the IRAP to explore natural language statements. *Journal of Contextual Behavioral Science*, *5*(4), 247–251. http://doi.org/10.1016/j.jcbs.2016.10.001

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49(4)*, 764-766. http://doi.org/10.1016/j.jesp.2013.03.013

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *11*, 361-365.

O'Toole, C., & Barnes-Holmes, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligent test: the importance of relational flexibility. *The Psychological Record*, *59*, 621-640.

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should we stop looking for a better scoring algorithm for handling Implicit Association test data? Test of the role of errors, extreme latencies treatment, scoring formula, and practice trials on reliability and validity. PLoS ONE 10(6): e0129601. doi:10.1371/journal.pone.0129601

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*, 19-30. http://doi.org/10.1037/1082-989X.13.1.19

Sriram, N., Nosek, B. A., & Greenwald, A. G. (2006). Scale invariant contrasts of response latency distributions. Unpublished manuscript. http://dx.doi.org/10.2139/ssrn.2213910

Thas, O., De Neve, J., Clement, L., & Ottoy, J. P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74,* 623-671. http://doi.org/10.1111/j.1467-9868.2011.01020.x

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. http://doi.org/10.1016/j.jbtep.2015.01.004