J. Behav. Ther. & Exp. Psychiat. 48 (2015) 59–65

Contents lists available at ScienceDirect

Journal of Behavior Therapy and Experimental Psychiatry

journal homepage: www.elsevier.com/locate/jbtep

A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain

Nigel A. Vahey*, Emma Nicholson*, Dermot Barnes-Holmes [1]

Department of Psychology, Maynooth University, Maynooth, Co. Kildare, Ireland

ARTICLE INFO

Article history:
Received 24 September 2014
Received in revised form
12 December 2014
Accepted 14 January 2015
Available online 7 February 2015

Keywords:
IRAP
Meta-analysis
Criterion effects
Clinically-focused
Statistical power

ABSTRACT

Background and objectives: The Implicit Relational Assessment Procedure (IRAP) is a technique that is attracting a substantial body of research literature, particularly within the clinical domain.
Method: In response, the present paper outlines a meta-analysis of clinically-focused IRAP effects ($N = 494$) to provide the first estimate of how well such effects validate against their respective criterion variables in general.
Results: The meta-analysis incorporated clinically-focused IRAP effects from 15 studies yielding a large effect size, $\bar{r} = .45$, with a desirably narrow 95% credibility interval (.23, .67). The funnel plot and subsequent sensitivity analyses indicated that this meta-effect was not subject to publication bias.
Limitations: The present meta-effect is an estimate based upon an IRAP literature that is still evolving rapidly in the clinical domain, and so as per its accompanying credibility interval, all conclusions that follow are necessarily provisional even if bounded. Apart from the fact that the current meta-effect might be subject to inadvertent under- and/or over-estimations of the current literature, the present meta-effect might strengthen with further refinements of the IRAP.
Conclusions: The current meta-effect provides the means to calculate what sample size would be required to achieve a statistical power of .80 when testing the criterion validity of clinically-focused IRAP effects using a given parametric statistic. For example, first-order Pearson correlations would hypothetically require an N of 29–37 for such purposes depending upon how conservatively over-estimation of the present meta-effect is controlled for. Overall, the IRAP compares favourably with alternative implicit measures in clinical psychology.

1. Introduction

Implicit measures constitute one of the most significant contributions to social psychology's literature during the past 15 years (Nosek, Hawkins, & Frazier, 2011). What has distinguished implicit measures from more traditional measures is their potential for developing insights into behaviour beyond, and including, those that are explicitly reported or observed (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Gawronski & Bodenhausen, 2011). Over recent years this feature of implicit measures continues to gather interest in applied domains even beyond the borders of social psychology. In particular, given that most forms of psychopathology are characterised by a lack of rational, conscious control over one's own behaviour, researchers in this domain are increasingly incorporating implicit measures into their research (Teachman, Cody, & Clerkin, 2010). The key hope is that implicit measures may offer novel insights into clinical phenomena which may not be detectable with traditional measures.

The Implicit Relational Assessment Procedure (IRAP), as a method, is based on a behaviour-analytic theory of human language and cognition, known as Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001). Given that the basic assumption of RFT is that the fundamental components of human language and cognition are relational, the IRAP focuses on stimulus relations and relational networks (e.g., Power, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009). The IRAP is a computer-based procedure, which requires participants to respond quickly and accurately in a manner that is consistent or inconsistent with their previous learning history. Responding is viewed as occurring on a continuum that ranges from “brief and immediate” to “extended

* Corresponding authors.Tel.: +353 1 708 6086.
E-mail addresses: nigelvahey@gmail.com (N.A. Vahey), Emma.Nicholson@nuim.ie (E. Nicholson), Dermot.Barnes-Holmes@nuim.ie (D. Barnes-Holmes).
[1] Tel.: +353 1 708 4765.

http://dx.doi.org/10.1016/j.jbtep.2015.01.004
0005-7916/

and elaborated". That is, when a response is elicited by a stimulus, it may be followed by another relational response, which may occur in response to the stimulus or the response itself. Thus, relating is a behavioural probability rather than a representation or a mental construct (Hayes, Barnes-Holmes & Wilson, 2012). Typically, participants respond by way of selecting the response options 'True" or 'False', which serve to indicate the extent to which the relations between the stimuli are either relationally coherent or incoherent (see Hayes et al. 2001, p. 66).

The core assumption is that responding on the procedure should be quicker on bias-consistent relative to bias-inconsistent trials, an effect that has been explained within RFT in terms of the Relational Elaboration and Coherence model (REC; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010). What makes the IRAP unique among other implicit measures is its ability to differentiate relatively complex implicit responses. From the perspective of more cognitively oriented researchers this means that the IRAP is uniquely equipped to measure implicit cognition in terms of propositions rather than mere associations (see Gawronski & De Houwer, 2014). The IRAP is designed to provide four effects in the form of four individual D_{IRAP} scores, with each measuring a distinct aspect of the topic of interest via one of four corresponding IRAP trial-types. For instance, an IRAP about body image might include four trial-types that respectively assess how readily people confirm versus deny whether "I want to be thin", whether "I want to be fat", whether "I don't want to be thin" and whether "I don't want to be fat" (see Parling, Cernvall, Stewart, Barnes-Holmes, & Ghaderi, 2012). In turn, researchers also have the option of averaging different trial-type D_{IRAP} scores together to achieve a *compound* D_{IRAP} score (see Barnes-Holmes, Barnes-Holmes, et al., 2010, p. 535).

At the time of writing, there were 41 empirical publications about the IRAP distributed across 14 different peer-review journals. Psychopathology is by far the most popular research domain with almost half (i.e. 18) of the publications being devoted to topics ranging from self-esteem (Remue, De Houwer, Barnes-Holmes, Vanderhasselt, & De Raedt, 2013; Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009), obsessive-compulsive disorder (Nicholson & Barnes-Holmes, 2012b; Nicholson, Dempsey, & Barnes-Holmes, 2014; Nicholson, McCourt, & Barnes-Holmes, 2013), body image (Parling et al., 2012), cocaine dependence (Carpenter, Martinez, Vadhan, Barnes-Holmes, & Nunes, 2012) and sexual attraction to children among sexual offenders (Dawson, Barnes-Holmes, Gresswell, Hart, & Gore, 2009), to name a few.

1.1. A meta-analysis of the clinical relevance of IRAP effects

The present meta-analysis aimed to quantify how much IRAP effects from clinically-relevant responding co-vary with corresponding clinically-relevant criterion variables, such as known group differences, self-reports of clinically-relevant psychological events (e.g., ratings on a disgust scale) and performances on behavioural approach tasks. Given that the validity and indeed utility of IRAP effects rely upon empirical relationships with such criterion variables (cf. Cohen, 1988, p. 25; Cumming, 2013), our meta-analysis sought to provide a first preliminary estimation of the clinical relevance of such effects. In addition, we also sought to provide this meta-analysis as a pre-requisite for any statistical power analyses being used for research planning with the IRAP in the clinical domain. Specifically, the estimated strength of the relationships using first-order correlations that researchers should expect to observe between an IRAP effect and clinically-relevant criterion variables will determine the sample size that would be required in order to replicate comparable effects with a statistical power of .80 (see Cumming, 2013).

2. Method

2.1. The literature search

We began searching for clinically-focused IRAP research by first examining all IRAP literature that was published, in press or under submission as of the 1st of April 2014. We cast our initial search widely in this manner so as to include even those clinically-relevant criterion effects that were mentioned only incidentally in the IRAP literature. Having first downloaded all 35 IRAP publications contained on the official IRAP website (see http://irapresearch.org/publications/), we then obtained a further 11 IRAP-related manuscripts by searching various online academic databases (i.e. *PsychInfo*, *Wiley Web of Science*, *ScienceDirect*, and *Google Scholar* using the Boolean search terms "Implicit Relational Assessment Procedure" OR "IRAP"). Finally, we canvassed an email listserv for any IRAP research under peer-review on the basis that the listserv targets the vast majority of the research community currently involved in IRAP research (i.e. the 'RFT' listserv at http://contextualscience.org/group/rft with approximately 640 members); no additional IRAP research manuscripts emerged. Thereafter, the first two authors set about independently reading through all of the aforementioned 46 IRAP manuscripts in search of criterion effects that adhered to specific standards.

2.2. The meta-analytic inclusion criterion

To be included within the current meta-analysis a given statistical effect must have described the co-variation of an IRAP effect with a corresponding clinically-focused criterion variable. To qualify as clinically-focused, the IRAP and criterion variables must have been deemed to target some aspect of a condition included in a major psychiatric diagnostic scheme such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, 2013). For example, consider IRAP effects designed to measure pro-smoking responding that participates in tobacco use disorder. In principle, such IRAP effects should co-vary, for instance, with various operational definitions of tobacco addiction such as number of cigarettes smoked per day or self reports of urges to smoke.

By applying this general inclusion criterion across the entire extant IRAP literature of 46 published articles, the first two authors ultimately identified only 15 articles reporting clinically focused IRAP effects that were accompanied by directly relevant criterion variables. Collectively, these 15 articles yielded 56 statistical effects between various clinically focused IRAP effects and their respective criterion variables. The authors arrived at this final selection of statistical effects by first independently reviewing the entire IRAP literature for qualifying statistical effects, before then critically discussing each other's rationales. From the outset, there was no disagreement between the authors about what statistical effects should be excluded from the meta-analysis; and of the 56 statistical effects only 8 (i.e. 14%) were not initially cited by both authors for inclusion.

In practice, the authors decided whether the responses measured by a given IRAP trial-type should co-vary with a specific criterion variable by consulting the relevant empirical literature. In the absence of such evidence, the authors strictly excluded even substantial statistical effects between IRAP effects and accompanying variables from the current meta-analysis. In other words, data were excluded even if the relevant statistical effects appeared to yield novel and/or strong support for the clinical relevance of the IRAP effects in question. For example, consider Nicholson and Barnes-Holmes (2012b) many correlations between IRAP effects about disgust versus self-reported clinical measures of obsessive compulsive disorder (OCD). These correlations confirm widely held suspicions in the literature that disgust processes are integral to

OCD, and yet we excluded them all from the current meta-analysis (i.e. six correlations in the range $.39 \leq r \leq .49$, and six in the range $.23 \leq r \leq .32$). The reason we did so is because the clinical measures employed by Nicholson and Barnes-Holmes (2012b) are documented as confounding the specific disgust processes being measured by the IRAP effects in that study and so they did not qualify as criterion variables for those IRAP effects (i.e. they confounded disgust propensity with disgust sensitivity; see Nicholson & Barnes-Holmes, 2012b).

2.3. Requests for supplementary information

Nine of the 56 effects that were eligible for inclusion in the current meta-analysis were alluded to in the literature without the information needed to calculate their respective magnitudes. We contacted the five relevant corresponding authors about this and in each case we successfully obtained the statistical information needed to calculate the missing effect(s) in question.

2.4. Converting qualifying effects for meta-analysis in terms of r

Many of the statistical effects eligible for inclusion in the meta-analysis were not originally reported in terms of *Pearson's r* and so we used the various conversion formula cited in Supplementary Table 1 to convert all such effects into corresponding *r* values (see Lakens, 2013; Rosenthal & DiMatteo, 2001; Rosnow, Rosenthal, & Rubin, 2000). We opted to conduct our meta-analysis in terms of *r* effects because these are readily interpretable, consistently bounded and preserve more information than other mainstream effect size metrics (see Field & Gillett, 2010; Lakens, 2013; Rosenthal & DiMatteo, 2001). In doing so we therefore maximised the number of different types of statistical effects that we could legitimately incorporate into our meta-analysis. In line with standard meta-analytic practice, we reverse scored any *r* values that involved a criterion variable that was measured in reversed terms. For transparency's sake, Supplementary Table 1 catalogues all of the aforementioned transformations in addition to detailing key aspects of the resulting *r* effects that were included in the current meta-analysis (i.e. the study they each originated from, the specific criterion relationship they each described and the original effect size statistics from which they were each calculated).

Having computed the 56 *r* effects described in Supplementary Table 1 we proceeded to combine any criterion effects that were statistically interdependent with each other[2] by computing weighted averages. Specifically, for each set of statistically interdependent criterion *r* effects we computed a weighted average *r* effect according to the degrees of freedom possessed by each respective effect within that set (cf. Field & Gillett, 2010, p. 670; Lakens, 2013; Olejnik & Algina, 2003; Rosnow et al., 2000). By calculating the degrees of freedom weighted average for each of the statistically independent sets of criterion effects we obtained the 15 criterion *r* effects described in Fig. 1 and these formed the basis of the present meta-analysis. In addition, Fig. 1 lists the sample size from which each of its criterion *r*s was obtained, and collectively these 15 effects incorporated the data from 494 participants.

3. Results

3.1. The meta-analysis

Meta-analyses generally come in one of two broad varieties: so called 'fixed-effect' versus 'random-effect' models (see Cumming, 2013; Field & Gillett, 2010; Vevea & Woods, 2005).[3] Whereas fixed effect models assume that all of its constituent effects are from the same sampling population as each other (i.e. with the same statistical parameters), random effect models can not only accommodate effects from different sampling populations but they can also estimate the extent to which those effects came from different populations (i.e. thereby subsuming fixed effect models; see Field & Gillett, 2010, pp. 672–675). As such, given that the 15 criterion effects in question originated from multiple clinical domains we decided to use a random effects model in order to detect any heterogeneity among the variables controlling each criterion effect. We chose Hunter and Schmidt's (2000) random effects model in particular because it readily allows the calculation of credibility intervals to qualify its resultant meta-effects,[4] and because it is a well established and versatile method (see Field & Gillett, 2010).

Applying Hunter and Schmidt's (2000) model to the 15 criterion effects in Fig. 1 yielded a meta-effect of $\bar{r} = .45$ with a desirably narrow 95% credibility interval going from $\bar{r} = .23$ to $\bar{r} = .67$. As such, our model indicated that we can expect 95% of clinically-relevant IRAP effects to co-vary with criterion variables with strength somewhere in the range $.23 \leq \bar{r} \leq .67$ (see Cumming, 2013). Furthermore, the observed variance of the 15 *r* effects was very small relative to that meta-effect (i.e. $s_r^2 = .006$ versus $\bar{r} = .45$). In other words, the effects in Fig. 1 were distributed very similarly to each other thus suggesting that they derived from populations that were very similar. To test this assertion we used a χ^2 test of the homogeneity of variance observed among the 15 *r* effects (i.e. based upon the sum of squared residuals around $\bar{r}$ versus the variance left unexplained by $\bar{r}$ [i.e. $1 - \bar{r}^2$]; see Field & Gillett, 2010, p. 676). The resulting $\chi^2(14) = 5.65$ was non-significant ($p = .97$). It therefore seems likely that the 15 *r* effects comprising the meta-analysis resulted from very similar controlling variables, despite the varied clinical domains from which they originated.

3.2. Publication bias analysis

We used the funnel plot in Fig. 2 to examine for the possibility that the current meta-analysis suffered a publication bias towards disregarding statistically non-significant effects. The core idea behind funnel plots is that effects derived from larger samples should, according to statistical first principles, tend to cluster more closely around their population average than effects derived from smaller samples. In particular, by quantifying the sample size of effects in terms of their standard errors it allows the calculation of

[2] It is a fundamental assumption of all meta-analytic methods that each must account for statistical interdependencies among its constituent effects in order to prevent any particular study from being over-represented in the final meta-analytic conclusions it provides (O'Sullivan, 2006; Field & Gillett, 2010; Rosenthal & DiMatteo, 2001; Rosnow et al., 2000). Generally data is considered to be statistically independent whenever it is derived from independent samples (i.e. samples that are collected without affecting each other; see Field & Gillett, 2010, p. 670; Rosenthal & DiMatteo, 2001, p. 67). As such, although it is possible for a between-groups study to yield as many independent effects as the number of independent groups it contains, nevertheless, those effects are merely independent of each other and not of any effects that span their respective groups (see Lakens, 2013; Olejnik & Algina, 2003; Rosnow et al., 2000).

[3] Although it is possible to combine fixed versus random effects models within a meta-analysis such improvised approaches are uncommon in practice (Cumming, 2013; Field & Gillett, 2010; Rosenthal & DiMatteo, 2001).

[4] Whereas confidence intervals only provide an estimation of the degree to which sampling variation contributes to a given set of effects, credibility intervals also account for whether other (unknown) variables might have moderated that set of effects (see Field & Gillett, 2010; Rosenthal & DiMatteo, 2001). As such, unlike confidence intervals, credibility intervals provide a means of generalizing beyond the peculiarities of whatever sample of effects are being examined. Incidentally, this means that credibility intervals are generally wider and thus more conservative than corresponding confidence intervals.

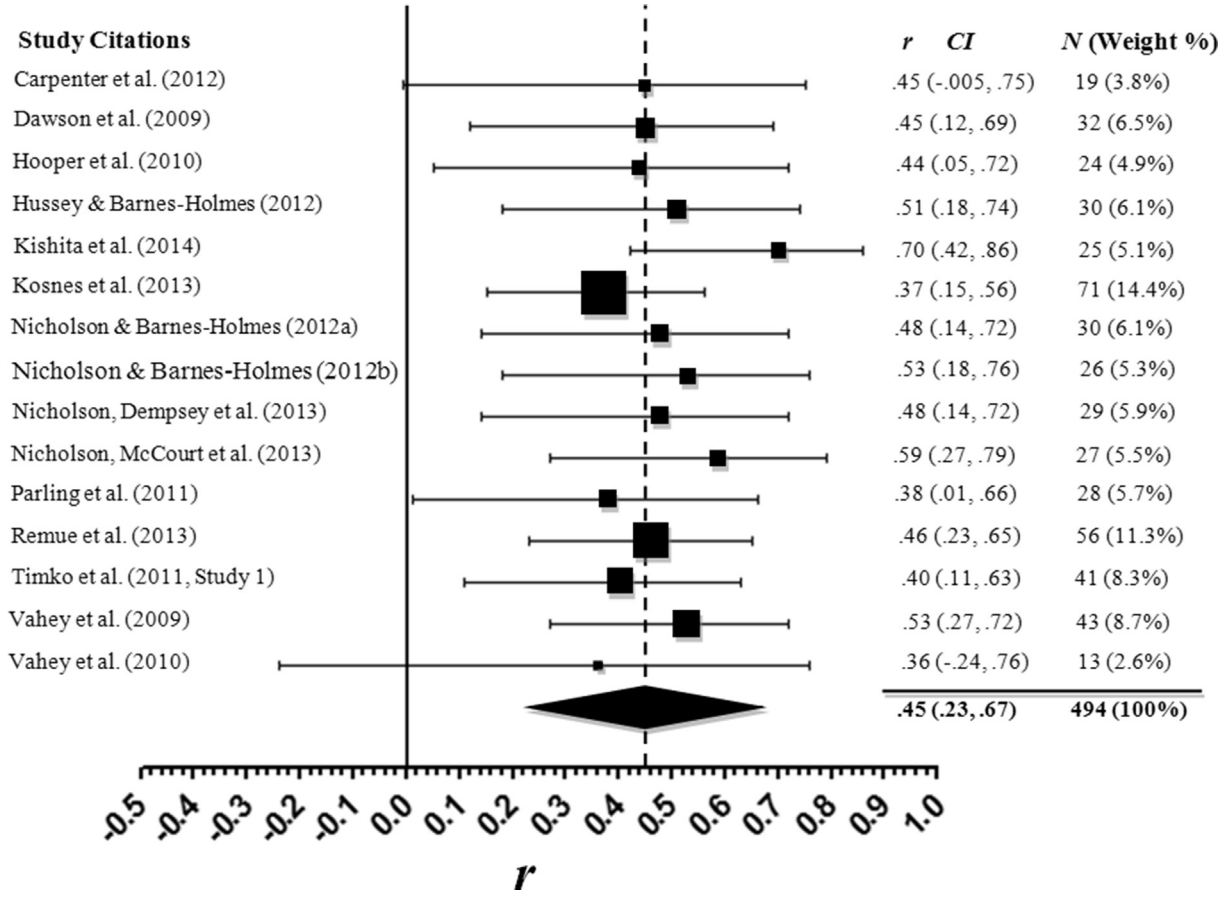

Fig. 1. A forest plot of the 15 criterion effects included in the present meta-analysis between various clinically-focused IRAP effects and their respective criteria. Each study cited is accompanied horizontally by a square whose centre indicates the strength of the criterion *r* effect the relevant study provided (i.e. as specified numerically in the column labelled '*r*') and whose size indicates the contribution that each criterion *r* made to the overall meta-effect in terms of its sample size *N* (i.e. as indicated by the column labelled 'Weight %'). In turn, the horizontal position of the centre of the diamond at the bottom of the plot represents the size of the overall meta-effect, $\bar{r} = .45$. Note that the width of this diamond indicates the 95% credibility interval (.23, .67) for the meta-effect, but that the width of the bars emanating from either side of each square above it represent the 95% confidence interval (CI) for each corresponding *r* value (see the column labelled 'CI' for exact values). Hooper, Villatte, Neofotistou & McHugh, 2010; Hussey & Barnes-Holmes, 2012; Kosnes, Whelan, O'Donovan, & McHugh, 2013; Nicholson & Barnes-Holmes 2012a.

confidence intervals (see Field & Gillett, 2010; Vevea & Woods, 2005). Thus, if the current meta-analysis was biased toward significant effects, there would be a lack of small effects at higher standard errors (towards the lower section of the funnel plot) compared to large effects with higher standard errors (i.e. a lack of effects in the lower left portion of a funnel plot).

Therefore, according to the relatively symmetric funnel plot in Fig. 2 our meta-analysis did not appear to prioritize consideration of statistically significant effects over non-significant effects (i.e. apart from the Z_r effect provided by Kishita, Muto, Ohtsuki, and Barnes-Holmes (2014) at the far right of the funnel plot all other Z_r effects clustered closely around the estimated population average). We confirmed this by conducting a Vevea and Wood's (2005) sensitivity analysis which systematically examined the impact that the following range of publication bias scenarios would have had on our random effects meta-analysis: 'moderate one-tailed selection', 'moderate two-tailed selection', 'severe one-tailed selection' and 'severe two-tailed selection'. Given that the meta-effect obtained with the current dataset varied by very little regardless of which form of publication bias we modelled with the Vevea and Woods method, it suggests that the current meta-analysis was not subject to publication bias (i.e. the various scenarios all reduced the size of the meta-effect $\bar{r}$ by the very small respective amounts of −.007, −.007, −.016, −.016). What is more, when we computed *Kendall's* τ as a measure of the rank correlation between Z_r and its associated standard error we found almost no relationship, $\tau(15) = .09$, $p = .66$ (two-tailed). Admittedly, the criterion effects displayed in Fig. 2 did not adopt a classic funnel plot pattern of distributing more widely about their average as their standard error increased (i.e. as one might expect at lower sample sizes). However, rather than indicating the prioritisation of statistically significant effects, instead the tight clustering of effects in Fig. 2 merely indicates that researchers have found that IRAP effects co-vary very similarly with their respective criterion variables across multiple clinical domains using a narrow range of consistently modest sample sizes (i.e. $\bar{x}_N = 32.9$, $s_N = 14.7$; also see Supplementary Table 1).[5]

[5] Note that the vertical axis in Fig. 2 is derived from the inverse of each effect's *N* as per Fisher's *Z*-transformation and so the narrow vertical range of observations in Fig. 2 is merely a reflection of the consistently modest sample sizes included in the current meta-analysis. Therefore, given that there are relatively few effects contained within Fig. 2's narrow vertical range, normal probability theory predicts that such a small unbiased sample of effects should usually congregate closely around their average as in Fig. 2 (i.e. by definition, samples from normally distributed populations are more likely to arise around that population's mean than around its tails). As such, when the foregoing facts are considered it is unsurprising that Fig. 2 did not display a clear funnel shape even assuming that its effects derived from an unbiased normal population (see Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; Vevea & Woods, 2005).

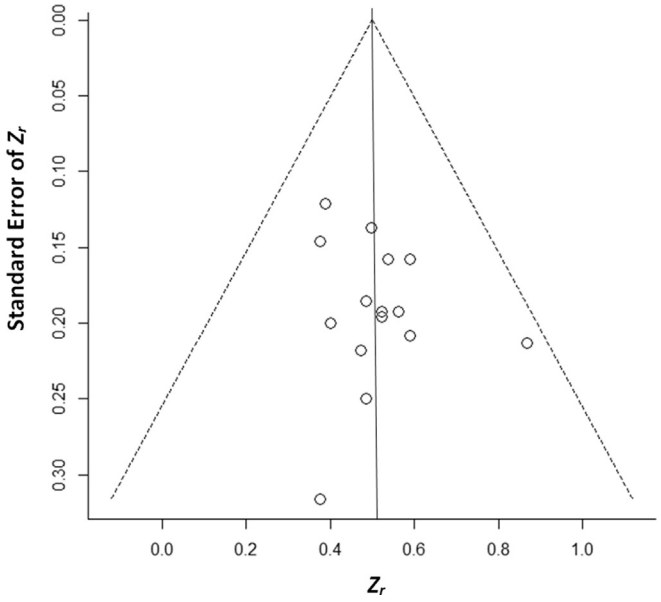

Fig. 2. A funnel plot of Fisher *Z*-transformed (standardised) versions of the 15 independent criterion *r* effects included in the meta-analysis (Z_r) versus their corresponding standard errors. The two diagonal dashed lines bound the 95% confidence interval for Z_r and the vertical line through the apex of these confidence intervals indicates the estimated population average for all 15 Z_r effects displayed.

4. Discussion

IRAP research has progressed steadily since the measure was first introduced into the literature with 18 clinically-focused research publications emerging which cover a wide range of topics in the clinical domain alone. And yet, critically, this substantial and rapidly growing body of clinically-focused IRAP research so far lacks any quantification of its collective findings and therefore its validity. The current meta-analysis produced a meta-effect of $\bar{r} = .45$ along with a desirably narrow credibility interval (.23, .67).

It may be useful to compare the present meta-effect with previous meta-analyses of alternative measures. In a recent meta-analysis examining the criterion validity of the Implicit Association Test (IAT) in relation to psychopathology meta-effects of $\bar{r} = .22$ and $\bar{r} = .3$ were obtained between clinically focused IAT scores and their respective criterion variables (i.e., the first for 18 IAT scores targeting addiction and the second for 19 IAT scores targeting psychopathology apart from addiction, see Greenwald, Poehlman, Uhlmann, & Banaji, 2009). Importantly, both of these criterion meta-effects for the IAT were based upon a similar number of 15 contributing effects as the present meta-analysis. Likewise, evaluative priming methods (including the Affective Misattribution Procedure) generally yield modest meta-effects between their scores and corresponding criterion variables ($.18 \leq \bar{r}$'s $\leq .28$; see Cameron, Brown-Iannuzzi, & Payne, 2012; Herring et al., 2013; Rooke, Hine, & Thorsteinsson, 2008). It therefore appears that the corresponding meta-effect of .45 produced by clinically-focused IRAP's compare favourably with leading alternative implicit measures, at least in the clinical domain.

A further motive for conducting the present meta-analysis was to provide a metric of clinical IRAP effects which would allow future researchers to calculate appropriate sample sizes in order to achieve the benchmark statistical power of .80 (see Cohen, 1988). For instance, based on the current meta-effect, a sample size of 29 participants would provide a study with a statistical power of .80 when examining the statistical significance of first-order *Pearson's r* correlations between clinically-focused IRAP effects and corresponding criterion variables. Alternatively, if one wished to use *t*-tests for such purposes the required sample size would be $N = 26$ for an independent samples *t*-test, and just $N = 8$ for a one-sample or repeated measures *t*-test. However, some caution is required in drawing such statistical conclusions.

First, these sample size estimations apply only to IRAP criterion effects as per the inclusion criteria employed in the current meta-analysis. For example, criterion effects are directional in nature and so the sample size estimation was based on one-tailed statistics. In contrast, exploratory research would require two-tailed statistics leading to a sample size estimation of 36 as being required for similarly sized effects. Second, a further caveat is that the above calculations of required sample size are only statistically sound given that the χ^2 test indicated that the variance observed among the 15 *r* effects proved to be homogeneous. It is therefore imperative that future meta-analyses of IRAP criterion effects continue to monitor the homogeneity of their constituent criterion effects. Third, the meta-effect is an estimate based upon an IRAP literature that is currently still evolving, and so as per its accompanying credibility interval (.23, .67), there is a degree of uncertainty about whether it might be subject to over- and/or under-estimation.

Fortunately, the literature has very recently suggested a number of ways of statistically accounting for the possibility of meta-effect under- and/or over-estimation when calculating the sample size required for a given statistical test at a given level of statistical power. Adopting a conservative approach in favour of controlling for overly optimistic publication biases, the most recent recommendation is to calculate sample size requirements not in terms of a given meta-effect, but rather in terms of the lower bound of its associated confidence interval (Perugini, Gallucci, & Costantini, 2014). Given that we obtained a confidence interval of (.40, .54) around the present meta-effect, Perugini et al.'s approach implies that a sample size of at least $N = 37$ would be required in order to achieve a statistical power of .80 when testing a continuous first-order correlation between a clinically-focused IRAP effect and a given criterion variable (i.e. as opposed to $N = 29$ without Perugini et al.'s correction). Likewise, Perugini et al.'s method implies that *N*s of at least 36 and 10 would respectively be required when using an independent samples *t*-test versus a one-sample/repeated-measures *t*-test for such purposes (i.e. as opposed to required *N*s of 26 and 8 without such corrections).

Another issue surrounding the replication of effects obtained across multiple studies, as is the case in any meta-analysis, concerns the likelihood of obtaining at least some non-significant effects. In particular, a publication bias in favour of significant findings would result in an unduly small number of non-significant effects being included within a given meta-analysis. Recently, Schimmack's (2012) has suggested that researchers employ an *incredibility index*. Schimmack's method involves using the profile of post hoc statistical power exhibited by a meta-effect's constituent criterion effects to calculate the probability that its sample of constituent effects could have involved more non-significant findings than it did. The current dataset yielded an incredibility index of .49 (according to Schimmack anything in the region of .9 or above would be deemed “incredible”). This indicates that there is only a 49% chance that *any* meta-analysis involving 15 effects with the profile of post hoc statistical power we observed, would include more than the two non-significant effects (see the confidence intervals in Fig. 1). In fact, Schimmack's model implies that, much as we found, on average one should expect to obtain 2.6 non-significant effects for a meta-analysis with the same profile as observed here (i.e., just .6 above what we found). Thus, overall, it does not appear that the

present meta-analysis was subject to a publication bias in favour of statistically significant findings.

The present paper demonstrates the potential of the IRAP as a tool for clinical assessment and it is hoped that the present meta-analysis will prove useful to clinical researchers who are considering using the IRAP as a measure. Given the increasing number of IRAP publications, particularly within the clinical domain, a systematic analysis seemed timely, particularly in terms of determining adequate sample sizes and appropriate statistical power for first-order correlations. The results of the current meta-analysis are to some extent reassuring in this regard given that the 15 study-level effects comprising the current meta-analysis had an average sample size of 32.9 participants (with the majority greater than or approaching $N = 29$, the minimum N hypothetically required for a single-comparison parametric test statistic to detect a meta-effect of the current magnitude). Furthermore, the meta-effect of $\bar{r} = .45$ compares reasonably well with the most widely used implicit measures presently in the literature. Indeed, it is possible that the present meta-effect might strengthen with further refinements of the IRAP that maximize the speed and/or accuracy with which participants complete its trials (see Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010, Experiment 2, pp. 54–62; Vahey, Boles, & Barnes-Holmes, 2010, pp. 469–471).

Acknowledgement

All three authors gratefully acknowledge the helpful and insightful recommendations provided by both Jan De Houwer and an anonymous colleague as part of the peer review process. The preparation of this article was supported by a postgraduate scholarship from the Irish Research Council awarded to the second author and financial support provided by National College of Ireland to the first author. The order of authorship for the first two authors was decided on the flip of a coin.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jbtep.2015.01.004.

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., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The implicit relational assessment procedure (IRAP): exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *The Psychological Record, 60*, 57–66.

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of social cognition: a meta-analysis of associations with behaviour and explicit attitudes. *Personality & Social Psychology Review, 16*, 330–350.

Carpenter, K. M., Martinez, D., Vadhan, N. P., Barnes-Holmes, D., & Nunes, E. V. (2012). Measures of attentional bias and relational responding are associated with behavioral treatment outcome for cocaine dependence. *The American Journal of Drug and Alcohol Abuse, 38*(2), 146–154.

Cohen, J. (1988). *Statistical power analysis for the behavioural science*. New Jersey, USA: Lawrence Erlbaum Associates.

Cumming, G. (2013). The new statistics: why and how. *Psychological Science, 25*, 7–29.

Dawson, D. L., Barnes-Holmes, D., Gresswell, D. M., Hart, A. J. P., & Gore, N. J. (2009). Assessing the implicit beliefs of sexual offenders using the Implicit Relational Assessment Procedure: a first study. *Sexual Abuse: A Journal of Research and Treatment, 21*, 57–75.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: a normative analysis and review. *Psychological Bulletin, 135*(3), 347–368.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*, 665–694.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44*, 59–127.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed.). (pp. 283–310). New York: Cambridge University Press.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.

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

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

Herring, D. R., White, K. R., Jabeen, L. N., Hinojos, M., Terrazas, G., Reyes, S. M., et al. (2013). On the automatic activation of attitudes: a quarter century of evaluative priming research. *Psychological Bulletin, 139*, 1062–1089.

Hooper, N., Villatte, M., Neofotistou, E., & McHugh, L. (2010). The effects of mindfulness versus thought suppression on implicit and explicit measures of experiential avoidance. *International Journal of Behavioral Consultation and Therapy, 6*, 233–245.

Hunter, J. E., & Schmidt, E. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292.

Hussey, I., & Barnes-Holmes, D. (2012). The IRAP as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice, 19*, 573–582.

Kishita, N., Muto, T., 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.

Kosnes, L., Whelan, R., O'Donovan, A., & McHugh, L. A. (2013). Implicit measurement of positive and negative future thinking as a predictor of depressive symptoms and hopelessness. *Consciousness and Cognition, 22*, 898–912.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 1–11.

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333*, 597–600.

Nicholson, E., & Barnes-Holmes, D. (2012a). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record, 62*, 263–278.

Nicholson, E., & Barnes-Holmes, D. (2012b). 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.

Nicholson, E., Dempsey, K., & Barnes-Holmes, D. (2014). The role of responsibility and threat appraisals in contamination fear and obsessive-compulsive tendencies at the implicit level. *Journal of Contextual Behavioral Science, 3*, 31–37.

Nicholson, E., McCourt, A., & Barnes-Holmes, D. (2013). The Implicit Relational Assessment Procedure (IRAP) as a measure of obsessive beliefs in relation to disgust. *Journal of Contextual Behavioral Science, 2*(1–2), 23–30.

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.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods, 8*, 434–447.

O'Sullivan, D. (2006). Meta-analysis. In G. M. Breakwell, C. Fife-Schaw, S. Hammond, & J. Smith (Eds.), *Research methods in psychology* (3rd ed.). (pp. 466–481). London, UK: Sage.

Parling, T., Cernvall, M., Stewart, I., Barnes-Holmes, D., & Ghaderi, A. (2012). Using the implicit relational assessment procedure to compare implicit pro-thin/anti-fat attitudes of patients with anorexia nervosa and non-clinical controls. *Eating Disorders, 20*(2), 127–143.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*, 319–332.

Power, P. M., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). *The Implicit Relational Assessment Procedure (IRAP) as a measure of implicit relative preferences: A first study*.

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.

Rooke, S. E., Hine, D. W., & Thorsteinsson, E. B. (2008). Implicit cognition and substance use: a meta-analysis. *Addictive Behaviors, 33*, 1314–1328.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect- size estimation. *Psychological Science, 11*, 446–453.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551–566.

Teachman, B., Cody, M., & Clerkin, E. (2010). Clinical applications of implicit social cognition theories and methods. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition* (pp. 489–521). New York, NY: Guilford Press.

Vahey, N. A., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). 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*, 371–388.

Vahey, N. A., Boles, S., & Barnes-Holmes, D. (2010). Measuring adolescents' smoking- related social identity preferences with the implicit relational assessment procedure (IRAP) for the first time: a starting Point that explains later IRAP evolutions. *International Journal of Psychology and Psychological Therapy, 10*, 453–474.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological Methods, 10*, 428–443.