Journal of Contextual Behavioral Science
Official Journal of the Association for Contextual Behavioral Science
Editor-in-Chief
Joseph Ciarrochi, Ph.D.
ACBS

www.elsevier.com/locate/jcbs

The Implicit Relational Assessment Procedure: Emerging reliability and validity data

Nima Golijani-Moghaddam, Aidan Hart, David L Dawson

PII: S2212-1447(13)00020-3
DOI: http://dx.doi.org/10.1016/j.jcbs.2013.05.002
Reference: JCBS19

To appear in: *Journal of Contextual Behavioral Science*

Received date: 2 October 2012
Revised date: 27 March 2013
Accepted date: 16 May 2013

Cite this article as: Nima Golijani-Moghaddam, Aidan Hart, David L Dawson, The Implicit Relational Assessment Procedure: Emerging reliability and validity data, *Journal of Contextual Behavioral Science,* http://dx.doi.org/10.1016/j.jcbs.2013.05.002

The Implicit Relational Assessment Procedure: Emerging reliability and validity data

Nima Golijani-Moghaddam[a]

Email: nmoghaddam@lincoln.ac.uk

Aidan Hart[a]

Email: ahart@lincoln.ac.uk

David L Dawson[a]

Email: ddawson@lincoln.ac.uk

[a]University of Lincoln
Health, Life and Social Sciences
1st Floor, Bridge House
Brayford Pool
Lincoln, UK
LN6 7TS

Tel: (+44)1522 837733
Fax: (+44)1522 837390

Author note

Correspondence concerning this paper should be addressed to the first author: Nima Golijani-Moghaddam, Clinical Psychologist and Research Tutor, University of Lincoln, Health Life and Social Sciences, 1st Floor, Bridge House, Brayford Pool, Lincoln, UK, LN6 7TS. Tel: (+44)7866516646

Abstract

The Implicit Relational Assessment Procedure (IRAP) is a measure of ‘implicit cognition’ developed on the basis of a contemporary behavioural analysis of language and cognition. The IRAP has now been applied to a range of foci over five years of published research. A frequently-cited caveat in publications to date is the need for further research to gauge the reliability and validity of the IRAP as an implicit measure. This review paper will provide a critical synthesis of available evidence for reliability and validity. The review applies a multifaceted test-theory approach to validity, and reliability is assessed through meta-analysis of published data. The discussion critically considers reviewed IRAP evidence with reference to the extant literature on alternative implicit measures, limitations of studies to date, and consideration of broader conceptual issues.

Keywords: Implicit; Implicit Relational Assessment Procedure; Relational Frame Theory; validity; reliability

Over the last two decades increasing attention has been given to the concept of 'implicit cognition' within psychological research (Greenwald et al., 2002). Although there remains a lack of consensus regarding the definition and operationalisation of this concept, reflecting broader differences in the theoretical and epistemological orientation of researchers in this area, a number of measures of 'implicit cognition' have been developed. The most commonly used and extensively discussed of these measures is the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998). The IAT was designed to measure the relative strength of pairs of associations (e.g., insects-disgust vs. flowers-disgust) in a computerised categorisation task. As an example, relatively rapid responding to insects-disgust (in comparison with flowers-disgust) would be considered indicative that insects and disgust are more strongly associated in memory.

Much of the evidence-base for implicit cognition has been developed through applications of the IAT and this measure has formed a prototype for subsequent developments in implicit measurement (Nosek, Hawkins, & Frazier, 2011). Most implicit measures have thus been designed to target a basic association between stimulus-pairs under conditions of time pressure. This operationalisation promotes the assumption that implicit cognition reflects the activation of an underlying associative-memory network - an assumption that is evinced in the representational models outlined by key researchers in the cognitively-aligned fields. However, it is important to distinguish the associative procedures of implicit measures from inferences about underlying processes or representations (Hughes, Barnes-Holmes, & De Houwer, 2010).

The Implicit Relational Assessment Procedure

The Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006) represents an alternative measure of implicit responding, developed from the perspective of Relational Frame Theory (Hughes et al., 2010; Hughes, Barnes-Holmes, & Vahey, 2012) a behavioural-analytic account of language and cognition (Hayes, Barnes-Holmes, & Roche, 2001). From an RFT perspective it is not the extent to which two stimuli are 'associated in memory' that is of importance, but an individual's history of deriving specific relationships between stimuli and the contexts that control the behaviour. RFT suggests that humans can learn to relate stimuli in a functionally limitless number of ways, such that stimuli come to participate in a multitude of relational frames. These include frames of association (or coordination) but also frames involving oppositional, hierarchal, temporal, causal, and deictic relations (among others). Applied to implicit cognition, a relational – versus associative – account offers the potential for measuring implicit cognition with greater specificity and sensitivity to context. Whilst the IAT and other implicit measures may indicate the relative strength of stimulus associations, they cannot gauge the direction or nature of relations between stimuli. This may not be a limitation if human cognitive processes are fundamentally associative, but RFT provides a conceptually-coherent alternative account (Barnes-Holmes et al., 2006; Hayes, Barnes-Holmes, & Roche, 2001).

The IRAP is a computerised latency-based assessment tool that requires participants to respond to a set of specific stimulus relations in ways that are alternately consistent or inconsistent with their prior verbal learning. The basic IRAP hypothesis is that participants will give faster responses on trials where the stimulus and required response are consistent with their private verbal relations or beliefs (e.g., I am good – True) than on belief-inconsistent trials (e.g., I am good – False); termed the IRAP effect. It is assumed that participants are slower to respond overtly when the response required goes against their more probable private relational responses – i.e., relational responses that are more readily evoked because of historical reinforcement and current contextual factors. Within this framework, implicit cognition as captured by the IRAP is a brief, unelaborated relational-response, measurable under specific contextual conditions (e.g., time pressure to respond).

The basic IRAP effect has now been demonstrated across a range of stimuli, settings, and samples (see Table 1).

Detailed descriptions of the IRAP procedure are available elsewhere (e.g., Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010) and will not be restated here. However, two aspects of the procedure will be described further for the purposes of the review: the scoring method and exclusionary practice criteria. Familiarity with these features will be beneficial for understanding details of the studies in Table 1, and both procedural aspects will form foci for discussion of limitations and suggestions for future research.

Scoring Method and Exclusionary Practice Criteria

In studies to date, raw IRAP response latency data (time in milliseconds between trial onset and participant response) are commonly transformed into *D*-IRAP scores[1]. These are normalised indices of response-latency differences between consistent and inconsistent blocks of IRAP tasks and form the primary outcome measure in most IRAP studies to date. Transformation to *D*-IRAP scores controls for individual variability in response speed relating to extraneous factors (such as differences in cognitive ability). The *D*-IRAP transformation is based on a scoring algorithm used in IAT research (Greenwald, Nosek, & Banaji, 2003).

Before they can contribute data to the study, IRAP respondents are typically required to complete practice trial-blocks until they emit 80% correct responses with a median response time of < 3000ms. Individuals who fail to meet criterion performance are normally excluded from further participation and analysis. It has been argued that the strict practice requirements of the IRAP may be necessary to obtain meaningful responses from participants (Roddy, Stewart, & Barnes-Holmes, 2010). Indeed, a more stringent (e.g., 2000ms) criterion has been suggested to increase the 'implicitness' of captured responses (Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2011).[2] This is consistent with the notion that implicit-explicit divergence can be accounted for in terms of a single process of relational responding over time, as articulated by Hughes, Barnes-Holmes and De Houwer (2011) in the Relational Elaboration and Coherence (REC) model. However, it is notable that comparable practice requirements are generally not imposed on the IAT or other implicit measures – a procedural distinction that may obfuscate contrasts (Roddy et al., 2010).

The Present Review

A common caveat of early IRAP research has been a need to establish the validity and reliability of the procedure as a measure of implicit cognition (e.g., Barnes-Holmes, Hayden,

[1] Although this is the scoring method that has been used most frequently, and is one of the outputs produced automatically by the IRAP software, the IRAP is not contingent on this scoring method. Critique of the *D*-IRAP scoring method is therefore not *de facto* critique of the IRAP procedure, but is relevant to interpretation of studies that have used this method. We offer one alternative suggestion for analysing IRAP responses in the discussion of this paper and it is possible to conceive of many other ways of comparing response-times between consistent and inconsistent trials with appropriate adjustment for individual variability.

[2] It should be acknowledged that there are no absolute recommendations regarding response-latency. Instructions from the authors of the IRAP (available at irapresearch.org) recommend use of the lowest latency criterion that is feasible for the current population and stimulus-set – recognising, for example, that more complex stimuli (e.g., statements versus single words) may require a longer response window. Study-specific pilot testing would be required to define optimal criteria, balancing accuracy and latency constraints.

Barnes-Holmes, & Stewart, 2008). Use of the IRAP is increasing (Nosek et al., 2011) and the procedure has now been applied in more than 20 empirical studies. It would seem timely to review the accumulated research-to-date for evidence of psychometric quality.

Given this, the goal of the present paper was to critically review available evidence for the validity and reliability of the IRAP. It is crucial to note, however, that concepts of reliability and validity have emerged from classical test theory (that rely on theoretical phenomena such as underlying attributes) and it is recognised that such terms are not entirely compatible with a functional-contextual account of human language and cognition. We acknowledge that there are broader aims regarding the development of the IRAP in terms of RFT and the REC model, which are more closely aligned to their underpinning philosophical assumptions (i.e., functional contextualism) but these are outside the scope of this review; interested readers are referred to Hughes, Barnes-Holmes, and Vahey (2012). For the purposes of this paper we adopt a pragmatic account of reliability and validity, informed by classical test theory, but without the concomitant ontological assumptions (Borsboom et al., 2004). In this way, it is hoped that the review will be of practical use for clinicians and applied researchers, who may be interested in the psychometric properties of the IRAP, how IRAP responses relate to other so-called implicit measures, how they may be useful for predicting future overt behaviours in context (including behaviours of clinical interest), and how they may aid group discrimination.

Search Strategy

To identify relevant articles the search string "implicit relational assessment procedure" was entered into three online databases on 20th March 2013: PsycINFO, Medline, and EMBASE. The search was limited to peer-reviewed articles and identified 30 unique references. After screening the full-text of these articles, four were excluded: two were review/commentary articles that did not provide empirical research evidence, one was concerned with a novel measure derived from the IRAP, and one described a parallel implicit measure with common behaviour-analytic foundations.

Five additional articles were identified through hand-searching references of retrieved papers and the online repository of IRAP articles at irapresearch.org. In total, searching identified 31 articles with empirical evidence pertaining to the validity and/or reliability of the IRAP (see Table 1).

[Table 1 about here]

Validity of the IRAP as a Measure of Implicit Cognition

The following sub-sections examine the IRAP in terms of (1) convergent, (2) contrasted groups, (3) experimental, (4) discriminant, and (5) criterion-related types of validity evidence. The section addressing criterion-related validity is separated into (i) concurrent validity and (ii) predictive validity. Two final sub-sections discuss considerations of (6) content and (7) face validity. It is acknowledged that 'types' of validity may be defined in different ways, dependent on the theoretical framework being applied, and that the same data could be read as evidence for different types of validity. For example, a finding of correlation between IRAP and explicit responses might be classed as confirmatory evidence for concurrent validity or disconfirmatory evidence for discriminant validity, depending on hypotheses/interpretation[3]. Categorisation of validity evidence in this review article largely reflects post-hoc interpretations of available empirical data, from the perspective of the authors; the reader is invited to apply their own framework to discern the legitimacy of evidence within each category.[4]

Convergent validity. High Inter-correlation of tests designed to measure the same attribute or process would be indicative of validity from a classical test theory perspective.

IRAP and IAT measures of cultural preferences in the same sample were not found to be significantly correlated (Barnes-Holmes, Waldron, Barnes-Holmes, & Stewart, 2009). Given that stimuli were maximally consistent between measures, and both measures purport to assess implicit cognition (on the basis of stimulus-response latencies), this may be taken as evidence against convergence of the IRAP with other implicit measures. It may be that differing features of the IAT and IRAP (e.g., relativistic versus absolute measurement; indirect-associative versus direct-relational responding) capture different aspects of the target attribute. However, comparisons *across studies* (and across independent samples; Chan, Barnes-Holmes, Barnes-Holmes, &

[3] Definitions of 'discriminant' and 'concurrent' types of validity in this article are aligned with previous reviews attending to these types of validity in implicit measures (e.g., Greenwald & Nosek, 2001).

[4] Moreover, whilst shorthand references are made to the validity of the IRAP, it is recognised that validity is not a property of the IRAP procedure itself: rather, we review evidence suggesting that the IRAP can support valid inferences about measured responses (obtained scores). One logical consequence of this is that summative impressions of validity will be somewhat dependent on study-specific operationalisations, theoretical predictions, and author interpretations.

Stewart, 2009) have shown similar patterns of findings between implicit measures: the IRAP appears to operate like the IAT and other implicit measures (Barnes-Holmes et al., 2006).

In two studies of implicit body-size bias (Roddy et al., 2010; Roddy, Stewart, & Barnes-Holmes, 2011) the authors found inconsistent evidence for a relationship between IAT and IRAP indices of the same attribute. In the former study, Roddy et al. (2010) found that the *D*-scores for the IAT and overall IRAP were not significantly correlated (r = .18, p = .15), although one of the four IRAP trial-type *D*-scores was correlated with *D*-IAT (r = .27, p = .03)[5]. In the latter study, Roddy et al. (2011) observed a significant correlation between *D*-IAT and overall *D*-IRAP scores. However, the relationship was of modest strength (r = .26, p = .04) and none of the specific IRAP trial-type *D*-scores were related to the *D*-IAT score. Taken together, these studies suggest weak and/or unreliable inter-correlation of IAT and IRAP measures designed to assess body-size bias, despite some commonality in target stimuli (images of ‘average weight’ and ‘overweight’ persons) and labels (‘good’ versus ‘bad’).

In contrast to the studies considered above, Barnes-Holmes, Murtagh, Barnes-Holmes, & Stewart (2010) reported significant and relatively strong correlations between IAT and IRAP measures (overall and trial-type *D*-scores) in a study of implicit attitudes towards meat and vegetables (*r*s .43-.54, *p*s < .02). The disparity between this finding and other reports of IAT-IRAP correlations could reflect distinct properties of the stimulus-set, sample, or attitude object. However, it is not clear why the IAT and IRAP measures in this study appeared particularly convergent.

Correlations may be attenuated by the limited reliability of compared measures; limited (internal and test-retest) reliability has been a concern for all implicit measures to date (Nosek, Greenwald, & Banaji, 2007). The IAT and IRAP appear to compare well with other implicit measures such as the GNAT, EAST, and evaluative priming measures (which have been found to have split-half reliabilities as low as -.05; Nosek et al., 2007). More direct comparisons would bolster this suggestion.

The IRAP is a recently developed measure and more research is required to examine overlap with other implicit measures (for matched targets/stimuli). Evidence for convergent validity

[5] It is not clear whether this analysis corrected for multiple testing. If not, this relationship would not remain significant following, for example, Bonferroni-adjustment.

among other response-latency implicit measures is mixed. Bosson, Swann, and Pennebaker (2000) examined relationships among seven implicit measures (of self-esteem): of 21 possible zero-order correlations between these measures only two reached significance. Two of the most established measures – the IAT and evaluative priming – have failed to converge in a number of studies (Fazio & Olson, 2003). However, Cunningham, Preacher, and Banaji (2001) found that correcting for inter-item inconsistencies improved convergence between implicit (IAT and priming) measures.

In future assessment of IRAP convergence with other implicit measures, precision of correlational analyses may be enhanced by: (i) maximising reliability within measures; (ii) correcting for remaining measurement error (low reliability) using latent variable analysis (following Cunningham et al., 2001) to circumvent impact on inter-measure correlations; (iii) increasing the similarity of stimuli/task demands between measures (Olson & Fazio, 2003); and (iv) using large samples (Lane, Banaji, Nosek, & Greenwald, 2007). The third of these recommendations may require particular attention given procedural differences between the IRAP and other measures, such as the IAT (Roddy et al., 2010).

Convergent validity may also be examined in terms of particular target attributes (e.g., spider fear). Such examination might look at correlations between multiple measures of the specific target attribute (e.g., explicit and implicit measures of spider fear): the point being to establish target-specific convergence (e.g., does this measure tap a common spider fear attribute?) rather than convergence supporting a general attribute (e.g., 'implicit cognition') and its accessibility (operationalisation) by the IRAP. The present focus is on the notion of a general implicit attribute and the validity of the IRAP as a tool for measuring this attribute.

Contrasted groups. Another approach to measuring validity is to examine differences in test scores between groups of people who would be expected to score differently on the test.

Barnes-Holmes et al (2009) found that IRAP effects distinguished known social groups (based on cultural preferences), outperforming the IAT with matched stimuli[6]. The IRAP has also been found to distinguish between self-reported meat-eaters and vegetarians (based on food

[6] However, neither implicit measure predicted group membership over and above an explicit measure in this study.

preference), matching IAT performance (Barnes-Holmes, Murtagh, et al., 2010). Notably, the IRAP was more informative about the nature of contrasts: whereas the IAT (as a relativistic measure) could not distinguish pro-vegetable from anti-meat preferences, the IRAP assessed values for each target separately. IRAP responding has also been shown to discriminate between groups categorised as high or low in spider fear when relevant stimuli are presented (Nicholson & Barnes-Holmes, 2012). This finding could arguably be considered supportive of concurrent (rather than contrasted groups) validity, as groupings were defined by scores on the applied explicit measure of spider-fear.

In a study conducted by Dawson and colleagues (Dawson, Barnes-Holmes, Gresswell, Hart, & Gore, 2009), the IRAP distinguished between individuals who had been convicted of a sexual offence against a child and a non-offender group (based on child-sexual classifications). However, sensitivity was moderate (68.8%) and specificity low (56.3%). Gray et al. found higher sensitivity (78%) and specificity (58%) in a comparable IAT study (Gray, Brown, MacCulloch, Smith, & Snowden, 2005). This study compared individuals who had been convicted of a sexual offence against a child with other (sexual/violent) offenders, so discriminant findings may be considered more impressive: the control group may have matched the experimental group more closely than in the IRAP study (where a university-based control group was used). However, the IAT and IRAP studies used different stimuli, obfuscating comparison of contrasted-groups validity.

A recent study by Hussey and Barnes-Holmes (2012) compared groups with ‘normal’ versus ‘mild/moderate’ depression scores and observed expected differences in IRAP-assessed emotional reactions (after a mood-induction procedure). Another recent study (Parling, Cernvall, Stewart, Barnes-Holmes, & Ghaderi, 2012) found that IRAP performance varied between a clinical group (individuals with identified features of Anorexia Nervosa) and a matched control group, observing a stronger self-directed anti-fat bias in the clinical group.[7] Similarly, Stockwell, Walker, and Eshleman (2010) showed expected performance differences between groups self-identifying with dissimilar sexual interests (stronger pro-BDSM bias in IRAP responses of individuals with

[7] The clinical group also demonstrated a stronger pro-fat bias towards others. Although this finding was not expected (i.e., not a ‘known’ difference between groups), the authors speculate that it may reflect a preference for favourable social comparisons. Whilst the present focus is on validity based on known group differences, an arguable strength and expectation of implicit measures is that they could yield insights – and may in fact contradict expectations developed on the basis of explicit responses alone.

BDSM/fetish interests)[8]. An earlier study found that the IRAP distinguished between prisoner and undergraduate groups on the basis of self-esteem (in accordance with known group differences; Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009). The IRAP has also proven sensitive to predicted gender differences in responding (Nolan, Murphy, & Barnes-Holmes, in press).

Finally, Vahey, Boles, and Barnes-Holmes (2010) reported data from a small pilot sample of adolescent smokers and non-smokers. This study showed a different pattern of responding between groups (consistent with expectations) but failed to show a significant difference when directly comparing groups (although the study was likely underpowered for this comparison).

Experimental. Experimental validity is evident when manipulation of relevant variables produces theoretically consistent effects on the measures that should be influenced. For example, effects of a self-esteem intervention on an implicit measure of self-esteem may provide evidence of attribute validity – especially if the intervention has specificity and does not simultaneously affect theoretically unrelated outcomes. Because less is known about influencing implicit versus explicit cognition – and changes in these attributes have been dissociated – interventions that have theoretically/empirically been shown to influence explicit responses may not affect implicit responses in the same way.

Cullen, Barnes-Holmes, Barnes-Holmes, and Stewart (2009) showed experimentally-manipulated malleability of IRAP effects (indexing ageist attitudes) between groups. A general anti-old IRAP bias was completely reversed in a group that was exposed to pro-old exemplars prior to testing. Effects were specific to the implicit measure (explicit attitude measures were unaffected – supporting discriminant validity, as discussed below). Barnes-Holmes and colleagues tested the effect of public versus private context on IRAP responding, hypothesising that evaluative responses to race may show greater bias in a private context (Barnes-Holmes et al., 2011). However, this study did not show the expected effects of context-manipulation. The authors attributed this to low reliability in IRAP responding and demonstrated that reliability could be improved by reducing the response window from 3000ms to 2000ms.

[8] Cullen and Barnes-Holmes (2008) report some preliminary data suggesting a similar difference between heterosexual and homosexual groups on the basis of homonegativity.

Hughes and Barnes-Holmes (2011) demonstrated that IRAP biases could be induced through relational training and verbal instruction. In this study, the attitude-induction procedures affected both implicit and explicit measures. The generality of effect (in contrast to the finding of Cullen et al., 2009) may partly reflect the novel and arbitrary nature of attitude objects (fictitious words) used in this study. Sensitivity to training was also reported by Bortoloti and De Rose (2012).

More recently, Hussey and Barnes-Holmes (2012) demonstrated effects of a mood-induction procedure on IRAP responding (indexing depressive emotional reactions). Sensitivity to mood induction was specifically evident for those who were higher in depression, or lower in 'psychological flexibility', and this was consistent with theoretical predictions regarding differential susceptibility to mood-induction.

One study to date has looked at the effects of clinical treatment-analogues on IRAP responding. Hooper, Villate, Neofotistou, and McHugh (2010) reported that a mindfulness intervention produced predicted changes in IRAP-responding (in terms of acceptance versus avoidance of negative emotion).

Discriminant validity. Discriminant validity may be looked at in terms of non-correlation with theoretically distinct explicit (versus implicit) responses. Evidence below suggests that the IRAP taps variance that is not captured by explicit measures: this may be interpreted more generally as the ability of the IRAP to capture a differential response-class.

Power, Barnes-Holmes, Barnes-Holmes, and Stewart (2009) reported discriminant implicit versus explicit preferences for nationalities. IRAP responses were found to diverge from explicit responses in a theoretically coherent way: implicit preferences were consistent with predictions from in-group theories of perceived social similarity, whereas explicit preferences were considered to reflect 'socially desirable' (politically sensitive) responding. Lack of correlation between IRAP and explicit responses was observed in studies looking at experiential avoidance (Hooper et al., 2010), age-related biases (Cullen et al., 2009), cocaine-use beliefs (Carpenter, Martinez, Vadhan, Barnes-Holmes, & Nunes, 2012), and body-size evaluations (Parling et al., 2012; Roddy et al., 2010).

In an early IRAP study, Barnes-Holmes et al. (2006) reported distinct explicit (positive) versus implicit (negative) responses towards individuals with autism in professionals working with this population. More recently, Chan and colleagues (2009) reported divergent patterns of implicit (IRAP) versus explicit responses to work and leisure in a sample of Irish and North American participants. Similar patterns of divergence have been reported elsewhere (Dawson et al., 2009; Roddy et al., 2010; Roddy et al., 2011); with studies showing IRAP response-patterns that were not evident in explicit responses. More qualified evidence was reported by Barnes Holmes et al (2011) who observed emergence of discriminant implicit versus explicit racial attitudes, but only under a condition of more stringent time pressure (i.e., 2000ms rather than 3000ms). Earlier studies had demonstrated divergence with a less stringent time-criterion (e.g., Dawson, et al., 2009).

The discriminant findings of Roddy et al. (2010, 2011) were arguably contradicted by a more recent study in the same domain (body-size biases; Nolan et al., in press) which found that IRAP responding converged with explicit responses. This finding could be classed as evidence for concurrent validity, but given the hypothesis of the authors, likely social-sensitivity of the evaluative domain, and previous findings by Roddy and colleagues it may instead be interpreted in terms of inconsistent discriminant validity. However, the observation of convergence versus divergence, whilst unexpected, does not represent a failure of replication *per se*. Conflicting results may be partly accounted for by differences in the stimulus-sets used (a factor that generally complicates synthesis of IRAP research): Nolan et al (in press) specifically focussed on intelligence-evaluative terms (e.g., ‘Clever’ versus ‘Foolish’) in contrast to the broader evaluations (‘Good’ versus ‘Bad’) used by Roddy et al (2010, 2011). Similarly, the pattern of divergence reported by Barnes-Holmes et al. (2006) was not reproduced in a more recent study of professional responses to autism (Kelly & Barnes-Holmes, 2013); again, although this was not an expected finding, the authors offered an explanation in terms of population differences between the two studies, and could point to evidence of divergence between IRAP responding and secondary self-reports.

Evidence for discriminant validity may further be inferred from studies that have shown differential effects of an intervention on IRAP versus explicit measures of the same target (Cullen et al., 2009; Hooper et al., 2010; Scheel, Fischer, McMahon, Mena, & Wolf, 2011).

Discriminant validity can also be assessed in terms of dissociation between theoretically distinct targets. Nicholson and Barnes-Holmes (2012b) showed that high- versus low-spider fear participants differed on IRAP trials measuring aversive bias towards spiders but not on IRAP trials measuring approach bias towards pleasant scenes. Such a difference suggests that the IRAP has target specificity and does not simply pick up on a propensity to show IRAP (i.e., response-time bias) effects (Lane et al., 2007). Similarly, Hussey and Barnes-Holmes (2012) found that IRAP responding to depressive relations was sensitive to group differences when group membership was defined by self-reported depression, but not when defined by anxiety or stress. The potentially nuanced specificity of the IRAP is perhaps best evinced by results showing differential responding on IRAPs designed to measure (primary) disgust propensity versus (secondary) disgust sensitivity (Nicholson & Barnes-Holmes, 2012a).

Criterion validity. Criterion validity refers to how strongly IRAP scores are related to other behaviours and psychological attributes.

Concurrent validity. Here, concurrent validity is considered in terms of the relationship of the IRAP to other established (primarily explicit) measures of targeted attributes (when measures are taken contiguously). Thus, a valid implicit measure should assess the same attribute as an explicit measure whilst also demonstrating dissociation: referring back to considerations of convergent and discriminant validity, it is evident that implicit measures must demonstrate an unusual balance of shared and unique variability (Greenwald & Nosek, 2009).

A number of findings in the available literature support concurrent validity of the IRAP. Domain-relevant IRAP responses have been found to correlate in the expected direction with (concurrently administered) established measures of spider fear (Nicholson & Barnes-Holmes, 2012), self-esteem (Vahey et al., 2009), and obsessive-compulsive tendencies (Nicholson & Barnes-Holmes, 2012a; Nicholson, McCourt, & Barnes-Holmes, in press). The literature to date has also shown significant correlations between IRAP responses and explicit evaluations of sexual practices (Stockwell et al., 2010), lifestyle (Barnes-Holmes et al., 2009), and diet (Barnes-Holmes, Murtagh, et al., 2010).

Two further articles report significant correlations between IRAP and explicit indices of the same targets (Barnes-Holmes et al., 2011; Timko, England, Herbert, & Forman, 2010). However, these studies examined many relationships (e.g., 100 correlations across two studies in Timko et al., 2010) without adjusting for multiple testing and findings could partially reflect an inflated Type I error-rate.

In a preliminary study (Barnes-Holmes et al., 2008), IRAP performance was found to be correlated with concurrent event-related-potentials (ERP) measures: inconsistent trials produced a more negative ERP waveform than consistent trials. Stimuli and response actions were equivalent across trials so differences may reflect automatic (well-established/high-probability) response processing versus low-probability response processing. A more recent study also showed a concordant relationship between IRAP responding and a psychophysiological indicator [facial electromyography (EMG); Roddy et al., 2011].

Finally, the speed and flexibility of IRAP relational-responding has been shown to demonstrate a positive relationship with general IQ (as theoretically predicted; O'Toole & Barnes-Holmes, 2009)[9].

Predictive validity. Juarascio and colleagues (2011) reported that an IRAP measure designed to assess implicit idealisation of thinness was prospectively predictive of changes in weight, disordered eating, and body dissatisfaction over the subsequent year. The IRAP measure demonstrated incremental predictive validity above applied explicit measures.

Carpenter and colleagues (2012) found that IRAP responding to statements about cocaine was associated with subsequent treatment outcome in a clinical sample of treatment-seeking cocaine-users. No parallel association was found for explicit measures, although this may reflect a lack of power. The sample size (n = 25) was not sufficient to allow multiple regression analyses, which may have supported more robust conclusions about predictive validity.

Nicholson and Barnes-Holmes (2012b) showed that IRAP-assessed spider fear predicted subsequent spider approach behaviour. The IRAP measure did not demonstrate incremental predictive validity over an explicit measure of spider fear, although the study was not designed or powered to detect any such effect. Nicholson and Barnes-Holmes (2012a) also demonstrated that IRAP-assessed sensitivity to disgust predicted subsequent avoidance behaviour (and was predictive over and above self-reported anxiety).

Roddy and colleagues (2010; 2011) report two studies showing that IRAP responding was predictive of behavioural intention towards an overweight person (in a hypothetical scenario). The

[9] Note that this relationship was found with raw IRAP responses and can be controlled for by using the *D*-IRAP transformation (Vahey et al., 2009). In this way, IRAP studies generally control for possible effects of individual differences in cognitive ability on speed/flexibility of responding.

conjectural nature of this scenario distinguishes these studies from those recording overt behaviour, and the findings of these studies could be considered more indicative of concurrent validity.

Content validity. Content validity is a qualitative type of validity (although quantitative approaches have been proposed; Haynes, Richard, & Kubany, 1995) concerned with the extent to which an instrument measures the important aspects of the attribute under assessment. Judgement of content validity is made with reference to a theoretical definition of the attribute to be assessed. Content validity could be judged for specific IRAPs in terms of the stimuli used (e.g., do IRAP-presented spider-fear stimuli adequately capture the attribute of spider fear?). More generally, content validity of the IRAP can be examined in terms of the extent to which the IRAP possesses the functional properties of an implicit measure (De Houwer, 2006; Power et al., 2009).

In the present discussion, validity pertains to inferences drawn from IRAP scores (thus extending beyond the IRAP itself, with implications for understanding implicit responding more broadly). Content validity is more limited in that it specifically looks at the appropriateness of the IRAP for measuring implicit responses. Content validity may be informative about the quality of the IRAP as an instrument but not the implicit attribute it was developed to measure (Sireci, 1998).

Drawing on available theoretical and empirical literature, De Houwer (2006) argued that a measure can be considered implicit if it meets one or more of the following criteria: (1) the participant is unaware of their cognition; (2) the participant is unaware that the outcome reflects their cognition; or (3) the participant has no control over the outcome. These criteria are considered below.

The IRAP is a relatively direct measure of ‘cognition’; the relations between presented stimuli are made clear: as a relational statement (e.g., I do not fear the spider). This means that, in contrast to disguised priming measures, and basic stimulus-pairing (associative) measures (such as the IAT), IRAP respondents are likely to be aware of the target being assessed. That is, IRAP respondents will probably be aware of what the IRAP outcome is supposed to reflect. Their insight into the target ‘cognition’ itself is more questionable: there may be private relational responses (cognitions) that they are unaware of, and these responses may diverge from the elaborated responses that are available to self-report. Because the criterion of cognitive unawareness (criterion 1) is difficult to assess/demonstrate and the criterion of outcome naivety (criterion 2) is likely not met for the IRAP, the remaining criterion (criterion 3) may be considered a critical test of content validity (according to current understanding of this attribute in the field of implicit cognition, as articulated by De Houwer, 2006).

McKenna, Barnes-Holmes, Barnes-Holmes, and Stewart (2007) studied the effects of instructing participants to 'fake' performance on the IRAP (having explained how the measure operates). Results showed no evidence of faking, suggesting that the outcome of the IRAP cannot be easily controlled by the respondent. IRAP responses may be harder to control than IAT responses: a study by Kim (2003) found that participants could fake the IAT when given explicit instructions. It appears that the IRAP meets criterion 3 of implicit measurement, although further empirical inquiry is merited.

Face validity. Face validity is considered a less important aspect of validity, indicative of whether a measure looks like it will measure the thing it purports to.

Considered from the perspective of the participant, implicit measures may have little face validity – in fact, as discussed above, participant naivety to the purpose of measurement is one criterion for considering a measure to be 'implicit.' The IRAP is exceptional among implicit measures in the directness of its stimulus presentations, so the participant may be relatively clear about the targets under examination (although they may not see how their responses will be measured).

From the perspective of experts in the field, the IRAP has face validity as an implicit measure (LeBel & Paunonen, 2011). It resembles established implicit measures (such as the IAT) in its basic structure and response-latency-based scoring.

Reliability

Considerations of validity are contingent on estimates of reliability. A measure cannot 'have validity' (or support valid conclusions) if scores on the measure reflect error variance and fail to capture the true score of the underlying attribute. Reliability of a measure thus represents an upper-bound for validity.

Internal consistency. The literature search identified seven studies reporting internal reliability for the IRAP in nine independent samples (total *n* = 316; mean *n* = 35.1, SD = 21.9;

range 15-79). Six samples used the 3000ms criterion at practice, two samples used the 2000ms criterion, and the criterion for one sample was not reported. One study compared two samples on this criterion (all other aspects of the task were held constant) and found an increase in reliability from .44 at 3000ms to .81 at 2000ms.

Meta-analysis was conducted using Hedges' method for a random-effects model (correcting mean estimates of effect-size for sample-size; Field & Gillett, 2010). Individual IRAP reliability estimates and relevant sample sizes are shown in Table 1. Reported estimates ranged from .23 to .85, with a weighted mean *r* of .653 (95% CI: .542-.742). The sample effect sizes used to calculate the weighted mean appeared homogenous ($Q(8) = 9.68$, $p = .29$).

Mean reliability is below the suggested minimum value for acceptable internal consistency in behavioural science measures (i.e., .70; Nunnally, 1978)[10]. Notably, the two samples that used a 2000ms criterion had acceptable consistency values (.72 and .81).

Table 2 shows stem and leaf plots for published IRAP reliability estimates. It is acknowledged that reliability is a property of test scores rather than the tests themselves, such that some variation in reliability from one sample to another may be expected – particularly given the variable stimulus-sets and procedural paradigms that have been used with the IRAP and the various iterations of the IRAP software to date.

[Table 2 about here]

As discussed in relation to convergent validity, reliability of implicit measures is generally lower than for parallel self-report measures. Indeed, the IRAP compares well to most implicit measures (LeBel & Paunonen, 2011) – such as the Extrinsic Affective Simon Task ($\alpha < .30$) and Go/No-go Association Task ($r_{sh} < .20$). However, evidence to date suggests that the IAT tends to demonstrate more acceptable reliabilities ($\alpha = .60$-$.90$; Gawronski, Deutsch, & Banse, 2011).

Stability. In terms of stability over time, Cullen et al. (2009) have reported test-retest reliability of .49. This is comparable to other implicit measures; for example, Nosek et al. (2007) reported that median test-retest reliability for the IAT was .56, with little variation as a function of retest intervals. Partly, expectations for test-retest reliability may differ according to conceptual

[10] This criterion was to be used for calculation of publication bias/failsafe-N statistics; given that published values fall below criterion, these statistics were not computed.

understanding of the attribute being measured: is implicit cognition trait-like or more state-dependent? Within the REC model, IRAP responses are expected to be highly context-dependent, but some conceptualisations within cognitively-aligned fields have suggested that implicit responses may be stable. In terms of potential pragmatic utility (e.g., use of the IRAP for monitoring treatment progress) responsivity to change may be prioritised over constancy across administrations. However, repeatability of measurement is desirable if the IRAP is to be used to identify intended effects over and above retest effects.

Evidence from IAT studies indicates that scores reflect both trait- and occasion-specific variation (Schmukle & Egloff, 2004). A similar finding for the IRAP might be expected (reflecting both historical- and current-contextual influences) but it is clear that more data on IRAP test-retest reliability within and across similar contexts is required – it would be premature to draw strong conclusions from the single study to date. Furthermore, the test-retest finding reported by Cullen et al. (2009) was subject to systematic effects of active intervention (implemented between test and retest for all participants). Intervention effects likely led to an underestimation of stability over time relative to retesting without intervention.

Although the preceding sections have attempted to synthesise available evidence for validity, reliability data to date suggests that findings could prove difficult to replicate – simulation has shown that the probability of replicating an effect is reduced for measures with lower internal consistency (LeBel & Paunonen, 2011). However, reliability data is currently in a state of infancy, and data from earlier studies may not reflect recent (and ongoing) developments in administrative practices (e.g., around optimising response-latency) which may affect IRAP reliabilities. In addition, as indicated above, understandings of reliability and validity of measures such as the IRAP cannot be entirely divorced from underlying philosophical assumptions. More simply, differences in reliability may provide researchers from a functional contextualist perspective with key data on the contextual factors impacting on responding, rather than indicating measurement ‘error’.

Discussion

Reviewing the inchoate IRAP literature there is accumulating evidence for multiple facets of validity. However, research to date is limited by the reliability (and therefore replicability) of IRAP scores and conceptual issues remain. The ensuing discussion summarises review findings and implications, considers limitations of research to date, and makes recommendations for future work. Comparative references are made to the extant literature for other implicit measures – principally the IAT, which might be considered the current ‘gold standard’ in the field.

The IRAP does seem to show the peculiar balance of shared and unique variability expected of implicit measures (if they are to differentially complement self-report). There are multiple examples of non-correlation with parallel explicit measures. Findings of discriminant validity are potentially undermined by the low reliability of the IRAP, as dissociations may be attributable to measurement ‘error’. However, in the context of evidence for concurrent and other criterion-related validity, this suggests that the IRAP captures an attribute that is independently informative about other behaviours and processes. Notwithstanding this, although associations and non-associations can be understood post-hoc as evidence for concurrent and discriminative validity respectively, it will be important to identify contextual factors that moderate the degree of association and to develop and test a priori predictions. The evidence to date indicates that brief and elaborated relational responses are more divergent for socially sensitive concepts (e.g., evaluations of race and age versus self-identified lifestyle preferences). However, it is notable that two recent studies (Kelly & Barnes-Holmes, 2013; Nolan et al., in press) did not show the divergence that may have been expected given the likely social-sensitivity of targets (bodyweight, autism) and previous IRAP studies that had shown implicit-explicit divergence for evaluations of these targets.

In terms of convergence with other implicit measures, available evidence is more limited and findings to date have been inconsistent: the IRAP has been shown to correlate with alternative implicit measures on some occasions but not others. In this respect the IRAP is not alone: in general, implicit measures have been observed to demonstrate limited convergent validity in relation to other implicit measures (LeBel & Paunonen, 2011).

However, evidence for validity from contrasts of known groups is relatively strong, with supportive findings reported in all (ten) studies to date. Fewer studies have evinced validity pertaining to contextual manipulation (experimental validity) although supportive findings were evinced in five of the six papers that examined this (Bortoloti & de Rose, 2012; Cullen et al., 2009; Hooper et al., 2010; Hughes & Barnes-Holmes, 2011; Hussey & Barnes-Holmes, 2012). The one study that did not show the expected effects of contextual manipulation (Barnes-Holmes et al., 2011) identified the possibility that low reliability masked these effects.

Early findings for the pragmatic validity of the IRAP are promising, indicating that the IRAP has potential to usefully complement self-report measures in predicting future behaviours – although few published studies to date have reported findings on behavioural prediction-over-time.

Limitations of IRAP research so far

Although the IRAP has a number of potential advantages over the IAT in terms of non-relativistic measurement and precision (testing specified relations rather than mere associations) studies to date suggest that the IAT may be a more accessible implicit measure (from the perspective of the participant). The IAT does not typically require participants to meet specific practice criteria before they can proceed to testing; and even when criteria are applied, it seems that participants find the IAT easier to complete. For example, Chan and colleagues (2009) reported that 16 of 55 participants (29%) failed to meet a criterion for data inclusion (80% correct responses) on the IRAP. In contrast 13 of 76 participants (17%) failed to meet this criterion on a matched IAT procedure. Clearly, both procedures excluded a sizable proportion of prospective participants, but the IRAP task appeared more challenging, and excluded more participants than the IAT. The exclusionary difficulty of the IRAP potentially limits its applications and the generalisability of reported findings. This factor could have a systematic influence on IRAP findings: those who have a stronger tendency to emit brief relational responses in a particular direction (i.e., individuals with a more pronounced implicit bias) may struggle to respond quickly and accurately on trials requiring responses that are inconsistent with their bias. Against this, it has been argued that the performance criteria of the IRAP are important to capture brief relational responses (as congruent with the REC model; Roddy et al., 2010). A possible implication of this reasoning is that the (less stringent) IAT would more likely capture a mixture of brief and elaborated responses – potentially undermining its specificity as an implicit measure. More direct comparisons of these measures, maximising procedural similarity, would be informative about their relative merits and associative versus relational accounts more generally. However, it is crucial to note that refinement of the IRAP procedure is ongoing: current guidance for conducting IRAP research (available at irapresearch.org) emphasises the importance of facilitating access to the procedure – e.g., through bespoke instructions to participants – and calibrating response criteria according to stimulus-difficulty and participant ability.

Irrespective of systematic error variance (validity issues) the utility of the IRAP may be limited by random measurement error (reliability issues). The reliability of implicit measures is generally relatively low in comparison with explicit measures of the same targets. All else being equal, low internal consistency may result in more Type II errors being made, reducing the probability of identification and replication of effects (LeBel & Paunonen, 2011). This could be taken as supportive of studies that do find significant effects (i.e., effects may be larger) although there may be a publication bias that masks important negative (and false-negative) results.

Comparing within implicit measures, the IRAP appears to be more reliable than most alternatives, but less reliable than the IAT – although many IRAP studies have not reported estimates of reliability and procedural differences (between studies and across iterations of the IRAP technology) again obfuscate comparability. LeBel and Paunonen (2011) recommend routine reporting of reliability estimates and suggest that it is imperative that researchers work to improve the reliability of implicit measures to bolster confidence in findings from studies using such measures.

Certainly, there is recognition within the IRAP literature of a need to further develop and refine the IRAP procedure. For example, Barnes-Holmes et al. (2011) demonstrated that internal consistency could be improved by decreasing permitted response latency from 3000 to 2000ms (split-half reliability increased from .44 to .81). This raises the question of whether there is an invisible ceiling to the reliability (and so validity) of this procedure and/or implicit measures more generally. While attempts to refine implicit methodologies are welcomed, the current authors would echo the recommendations of Barnes-Holmes et al. (2010) by urging caution over how such refinement is implemented. Reducing response latencies could exclude relevant participants erroneously. For example, as discussed above, those who emit strong implicit responses may find it particularly difficult to respond quickly in their counter-conditioned trials – potentially leading to the exclusion of participants who may be of most interest to the research being carried out. Certainly, current instructions to researchers (at irapresearch.org) stress that criteria should be calibrated on the basis of pilot-testing so as to avoid unnecessary exclusion.

It is apparent that the most common approach to computing IRAP scores (as D-scores) introduces another source of unreliability, as the transformation involves the use of difference-scores (e.g., Chiou & Spreng, 1996). The reliability of these scores decreases as the correlation between components increases. One alternative approach to scoring and analysis may be to use multilevel modelling of raw reaction-time data (E. Ferguson, Moghaddam, & Bibby, 2007). Multilevel modelling would retain variability both between and within individuals, by examining trial-by-trial reaction times as nested within participants. Such an approach affords greater power and modelling possibilities – and avoids aggregating response-transformations, which substantively change the outcome of interest.

IRAP and other implicit effects may be highly sensitive to context (M. J. Ferguson & Bargh, 2007) and the relative sensitivity of brief versus elaborated (explicit) responses is predicted by the REC model. To be sure, such sensitivity may contribute to unreliability and complicate replication across contexts, irrespective of attempts to maximise continuity in stimuli and response criteria. Investigation of contextual (historical and situational) moderators of IRAP effects will inform understanding of any systematic effects on reliability.

In another sense, the IRAP is a measurement procedure with variable stimulus-sets – and different stimulus-sets have been applied to study similar relations (e.g., Timko et al., 2010; Vahey et al., 2009). This flexibility is a strength of the IRAP but may also limit research consistency. Where possible, it may be useful to identify and reuse operationalisations of the IRAP that have proved reliable and able to support valid inferences.

It is important to note that some of the above-described limitations of IRAP research to date are not intrinsic limitations of the IRAP. One recent study (Scheel et al., 2011) has employed the core procedure of the IRAP – capturing brief responding to stimulus relations that are alternately consistent or inconsistent with a given propensity – without using commonly-applied practice criteria or scoring algorithms (or indeed the standard IRAP software). This study serves to illustrate that some of the foregoing issues are secondary administrative considerations, independent of the IRAP procedure and its basic hypothesis: as administrative practices are refined over time, the reliability and validity of the IRAP – i.e., the ability of this procedure to support valid inferences about ‘implicit’ responses – will likely become clearer.

Conceptual issues

Theoretically, evidence for IRAP effects suggests that implicit responses can be more complex than an associative account would allow for (Hughes, Barnes-Holmes, & De Houwer, 2011).This challenges assumptions that underpin the IAT and suggests that the IAT captures (artificially) limited information about the underlying attribute of interest (implicit responding). Some studies to date (e.g., Parling et al., 2012) have shown how relational-specificity of the IRAP can be informative, permitting insights that the IAT and other implicit measures would not allow.

It seems that implicit responses may be conceptualised as relational or propositional (Hughes & Barnes-Holmes, 2011) and the IRAP is likely more suited to the assessment of such responses than the IAT. Thus, returning to questions of validity, and content validity in particular, if implicit cognition is relational (versus purely associative) the IRAP may be considered superior to other available implicit measures.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, 32*(7), 169-177.

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., 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. *Psychological Record, 58*(4), 497-515.

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2011). The Implicit Relational Assessment Procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. *Psychological Record, 60*(1), 57-79.

Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes toward meat and vegetables in vegetarians and meat-eaters. *Psychological Record, 60*(2), 287-305.

Barnes-Holmes, D., Waldron, D., Barnes-Holmes, Y., & Stewart, I. (2009). Testing the validity of the Implicit Relational Assessment Procedure and the Implicit Association Test: Measuring attitudes toward Dublin and country life in Ireland. *Psychological Record, 59*(3), 389-406.

Bortoloti, R., & de Rose, J. C. (2012). Equivalent stimuli are more strongly related after training with delayed matching than after simultaneous matching: A study using the Implicit Relational Assessment Procedure (IRAP). *The Psychological Record., 62*(1), 41-54.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*(4), 631-643.

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.

Chan, G., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). Implicit attitudes to work and leisure among North American and Irish individuals: A preliminary study. *International Journal of Psychology and Psychological Therapy, 9*(3), 317-334.

Chiou, J., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction Dissatisfaction and Complaining Behaviour, 9*, 158-167.

Cullen, C., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The Implicit Relational Assessment Procedure (IRAP) and the malleability of ageist attitudes. *Psychological Record, 59*(4), 591-620.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12*(2), 163-170.

Dawson, D. L., Barnes-Holmes, D., Gresswell, D. M., Hart, A. J., & 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*(1), 57-75. doi: 10.1177/1079063208326928

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11-28). Thousand Oaks, CA: Sage Publishers.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition. research: their meaning and use. *Annual Review of Psychology, 54*, 297-327. doi: 10.1146/annurev.psych.54.101601.145225

Ferguson, E., Moghaddam, N. G., & Bibby, P. A. (2007). Memory bias in health anxiety is related to the emotional valence of health-related words. *Journal of psychosomatic research, 62*(3), 263-274.

Ferguson, M. J., & Bargh, J. A. (2007). Beyond the attitude object: Automatic attitudes spring from object-centered-contexts. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 216-246). New York, NY: Guilford.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694.

Gawronski, B., Deutsch, R., & Banse, R. (2011). Response interference tasks as indirect measures of automatic associations. In K. C. Klauer, C. Stahl & A. Voss (Eds.), *Cognitive methods in social psychology*. New York, NY: Guilford.

Gray, N. S., Brown, A. S., MacCulloch, M. J., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology, 114*(2), 304-308.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*(1), 3-25.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48*(2), 85-93.

Greenwald, A. G., & Nosek, B. A. (2009). Attitudinal dissociation: what does it mean? In R. E. Petty, R. H. Fazio & P. Briñol (Eds.), *Attitudes: Insights from the New Implicit Measures* (pp. 65-82). Hillsdale: Erlbaum.

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

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

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Hooper, N., Villatte, M., Neofotistou, E., & McHugh, L. (2010). The effects of mindfulness versus thought suppression on implicit and explicit measures of experiential avoidance. *The International Journal of Behavioral Consultation and Therapy, 6*(3), 233-244.

Hughes, S., & Barnes-Holmes, D. (2011). On the formation and persistence of implicit attitudes: New evidence from the Implicit Relational Assessment Procedure (IRAP). *The Psychological Record, 61*(3), 391-410.

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2010). The Dominance of Associative Theorising in Implicit Attitude Research: Propositional and Behavioral Alternatives. In M. Valverdre & M. Alvaerz (Eds.), *Current perspectives in human learning*.

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

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science, 1*(1–2), 17-38. doi: http://dx.doi.org/10.1016/j.jcbs.2012.09.003

Hussey, I., & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice., 19*(4), 573-582.

Juarascio, A. S., Forman, E. M., Timko, C. A., Herbert, J. D., Butryn, M., & Lowe, M. (2011). Implicit internalization of the thin ideal as a predictor of increases in weight, body dissatisfaction, and disordered eating. *Eating Behaviors, 12*(3), 207-213. doi: 10.1016/j.eatbeh.2011.04.004

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

Kim, D. Y. (2003). Voluntary controllability of the implicit association test (IAT). *Social Psychology Quarterly, 66*, 83-96.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and Using the Implicit Association Test: IV: Procedures and validity. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes: Procedures and controversies* (pp. 59-102). New York: Guilford Press.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37*(4), 570-583.

McKenna, I. M., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2007). Testing the fake-ability of the implicit relational assessment procedure (IRAP): The first study. *International Journal of Psychology and Psychological Therapy, 7*, 123-138.

Nicholson, E., & Barnes-Holmes, D. (2012a). Developing an implicit measure of disgust propensity and disgust sensitivity: Examining the role of implicit disgust propensity and sensitivity in obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry., 43*(3), 922-930.

Nicholson, E., & Barnes-Holmes, D. (2012b). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record, 62*(2), 263-277.

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

Nolan, J., Murphy, C., & Barnes-Holmes, D. (in press). Implicit perceptions of intelligence in slim and overweight individuals: Effects of gender of participants and target stimuli. *The Psychological Record*.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265-292). London: Psychology Press.

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.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Toole, C., & Barnes-Holmes, D. (2009). Three chronometric indices of relational responding as predictors of performance on a brief intelligence test: The importance of relational flexibility. *Psychological Record, 59*(1), 119-132.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*, 36-39.

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.

Power, P., 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. *Psychological Record, 59*(4), 621-639.

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2010). Anti-fat, pro-slim, or both? Using two reaction-time based measures to assess implicit attitudes to the slim and overweight. *Journal of Health Psychology, 15*(3), 416-425. doi: 10.1177/1359105309350232

Roddy, S., Stewart, I., & Barnes-Holmes, D. (2011). Facial reactions reveal that slim is good but fat is not bad: Implicit and explicit measures of body-size bias. *European Journal of Social Psychology, 41*(6), 688-694.

Scheel, M. H., Fischer, L. A., McMahon, A. J., Mena, M. M., & Wolf, J. E. (2011). The Implicit Relational Assessment Procedure (IRAP) as a measure of women's stereotypes about gay men. *Current Research in Social Psychology, 18*(2).

Schmukle, S. C., & Egloff, B. (2004). Does the Implicit Association Test for assessing anxiety measure trait and state variance? *European Journal of Personality, 18*(6), 483-494.

Sireci, S. G. (1998). The construct of content validity. *Social indicators research, 45*(1), 83-117.

Stockwell, F. M. J., Walker, D. J., & Eshleman, J. W. (2010). Measures of implicit and explicit attitudes toward mainstream and BDSM sexual terms using the IRAP and questionaire with BDSM/Fetish and student participants. *The Psychological Record, 60*(2), 307-324.

Timko, C. A., England, E. L., Herbert, J. D., & Forman, E. M. (2010). The Implicit Relational Assessment Procedure as a measure of self-esteem. *The Psychological Record, 60*(4), 679-698.

Vahey, N., 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. *Psychological Record, 59*(3), 371-387.

Vahey, N., 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*(3), 453-474.

Table 1

Reviewed studies with empirical data pertaining to the reliability or validity of the Implicit Relational Assessment Procedure

Author/Date/Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Barnes-Holmes et al. 2006 Ireland Report of 3 studies (S1 and S3 republished elsewhere)	[S2]. Relation of evaluative terms to Autistic Spectrum Disorder (ASD)	[S2]. Professionals with varying experience of working with ASD. 3 groups: no experience (n=16); <6 months (12); and >6 months (16)						[S2]. Between group differences in explicit but not IRAP responses	
Barnes-Holmes et al. 2008 Ireland Report of two experiments (=[S1] from 2006 paper)	[e1] & [e2]. Relational evaluation of words as pleasant vs. unpleasant	[e1]. 28 UGs. 2 groups: experimental (n=16) and control (n=12) [e2]. 11							[e2]. IRAP responding correlated with ERP (different waveform for consistent vs. inconsistent trials)
Barnes-Holmes et al. 2011 Ireland Report of two experiments	[e1] & [e2]. Relation of evaluative terms to race	[e1]. 31 UGs. 2 groups: public context IRAP (n=16) and private context IRAP (n=15) [e2]. 19 UGs (public context only)	[e1]. With 3000 ms limit, reliability was .23 in private and .44 in public [e2]. With 2000 ms limit, reliability was			[e1]. Contrary to experimental predictions, public/private context did not have an overall significant effect on IRAP responding. Attributed to lack of time-pressure		[e2]. IRAP responses (pro-white) diverged from explicit (pro-black) bias, as predicted. More equivocal results were seen in [e1] under less-stringent respons	[e1] & [e2]. Some sig. correlations between (5) IRAP indices and (3) explicit measures. *Relationships not consistent and may be an artefact*

.81 (in public)	(3-second response-time criterion)	e time criterion	*of multiple testing*

Author/Date/ Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Barnes-Holmes et al. 2010 Ireland	Relation of evaluative terms to meat vs. vegetables	32 students. 2 groups: vegetarians (n=16) and meat-eaters (n=16)	.72 (overall D) (.76 for IAT in same study)	IRAP and IAT scores correlated significantly (*r* = .54) and performed in similar ways (e.g., predicting group membership)	IRAP responding discriminated between vegetarians and meat-eaters (incrementally explained additional 14% of the variance, over explicit report alone)				IRAP responding correlated weakly with some features of explicit rating measures. Similar finding for parallel IAT.
Barnes-Holmes et al. 2009 Ireland	Relation of evaluative terms to Dublin and country life in Ireland	26 individuals. 2 groups: Dublin dwellers (n=13) and rural dwellers (n=13) Originally 31; data from 5 was excluded	.41 (.46 for IAT in same study)	IRAP and IAT scores did not demonstrate significant correlation (*r* = .08)	IRAP responding differed significantly between groups, Although neither implicit measure (IRAP or IAT) incrementally added to ability to discriminate groups over explicit report alone				IRAP responding correlated with explicit measures (weak to moderate correlations). Parallel IAT did not demonstrate correlation.
Bortoloti & De Rose 2012 Brazil	Relation of happy and angry faces to novel stimuli (nonsense words)	19 UGs. 2 training conditions: simultaneous matching to sample (n=10) or delayed matching to sample (n=9)				IRAP responding was sensitive to training via delayed matching. Responding was partially sensitive to training via simultaneous matching (IRAP effect for happy target only)			

Author/Date/ Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Campbell et al. 2011 Ireland	Relational evaluation of words as pleasant vs. unpleasant	48 UGs Originally 60; 12 failed to meet IRAP criteria	.64 (n=47)						
Carpenter et al. 2012 USA	Relation of self-descriptions to cocaine vs. no cocaine	25 individuals seeking treatment for active cocaine-use	.85 (overall D)				IRAP responding predicted subsequent treatment outcome	No correlation between IRAP and explicit measures	
Chan et al. 2009 Ireland/Canada	Relation of evaluative terms to work vs. leisure	39 participants (19 Irish, 11 US, 9 Canadian) Originally 55; 16 (1; 7; 8) failed to meet IRAP criteria		IRAP responding was concordant with IAT responding in a parallel experiment (independent sample within the same study); both showed divergent explicit (pro-work) and implicitly (pro-leisure) responses				IRAP responding diverged from explicit responding overall, with negative correlations between IRAP and explicit responses in a subgroup of (US) participants	
Cullen et al. 2009 Ireland	Relation of evaluative terms to age	[e1]. 12 Irish nationals [e2]. 23 participants Originally 24; 1 lost to follow-up	Test-retest r = .49			[e2]. Pro-old exemplar condition reversed anti-old bias in IRAP responding		[e1] & [e2]. No correlation between IRAP and explicit measures, although both demonstrated pro-young preference. [e2]. Selective effects of intervention on IRAP versus explicit responses	
Dawson et al. 2009 UK	Relation of sexual terms to	32 males. 2 groups: sex offenders			IRAP responding discriminate			Groups differed on IRAP	

	children	(n=16) and non-offenders (n=16)			d groups, correctly classifying 69% of offenders and 56% of non-offenders		responding but not explicit self-report
Drake et al. 2010 USA	Relation of evaluative terms to race, religion, and obesity. Relation of chores and occupations to gender [Participants first completed a practice IRAP relating shapes and colours]	58 UGs. 4 conditions: race IRAP (n=15); religion IRAPs (n=14); gender IRAPs (n=16); and obesity (n=13) Originally 67 (17; 17; 17; 16); 9 failed to meet criteria on either practice or condition-specific IRAPs	.60		Divergence in responding on a race IRAP between African-American and Caucasian participants – although subgroup analyses were not planned		
Hooper et al. 2010 UK	Relation of acceptance vs. avoidance terms to negative emotions	24 UGs. 2 groups: mindfulness (n=15) and thought suppression (n=9) Originally 50; 26 failed to meet criteria on either pre- or post-induction IRAP				One intervention (mindfulness) produced a change in IRAP responding, consistent with expectations	Differential sensitivity of IRAP vs. explicit measure to intervention. No correlation with explicit measure of experiential avoidance at pre- or post

Author/Date/Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Hughes & Barnes-Holmes 2011 Ireland	Relation of positive and negative words to novel stimuli (arbitrary nonsense words)	64 UGs. 3 conditions: relational training, verbal instruction, or combined training				IRAP responding was sensitive to training in all conditions (relational, instructive, and combined)			Post-training explicit responses were concordant with IRAP responses
Hussey & Barnes-Holmes 2012 Ireland	Relations among positive vs. negative antecedents and responses (of form: 'When X happens … I feel Y')	30 UGs. 2 groups: 'normal' range of depression scoring (n=15) and 'mild/moderate' range of depression scoring (n=15).			After negative mood induction, IRAP responding differed between groups (more positive emotional bias in 'normal' versus 'mild/moderate' depression group).	IRAP responding was sensitive to mood induction in the 'mild/moderate' depression group, but not the 'normal' group, as predicted.			
					Parallel findings when participants were grouped as high (n=12) versus low (n=18) in psychological flexibility (substantial overlap b/w 'low flexibility' and 'mild/moderate' depression groups).				
Juarascio et al. 2011 USA	Relation of self to images of fat and thin women	79 newly enrolled female UGs.19 lost to follow-up Originally 80; 1 failed to meet IRAP criteria	.72				IRAP responding prospectively predicted changes in body-image dissatisfaction, disordered eating, and weight – over and above an explicit measure		

Author/Date/ Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Kelly & Barnes-Holmes 2013 Ireland	Relation of evaluative terms to Autistic Spectrum Disorder (ASD)	32 professionals. 2 groups: tutors trained in applied behaviour analysis (n=16) and mainstream school teachers (n=16)						Some evidence of implicit-explicit divergence between groups (on secondary variables) but discriminant pattern in Barnes-Holmes et al (2006) was not replicated	
Nicholson & Barnes-Holmes 2012 Ireland	Relation of fear- vs. approach-type responses to spiders vs. pleasant scenes	30 UGs Originally 40; 6 failed to meet IRAP criteria, 4 were not included on basis of explicit spider-fear			IRAP responding discriminated high- and low-spider fear groups, correctly classifying 70% [could be considered concurrent validity, as groups were defined by score on the explicit measure]		IRAP responding predicted performance on a spider-approach task. However, the IRAP did not show incremental validity over and above an explicit measure in this study		IRAP responding was correlated with explicit spider fear
Nicholson & Barnes-Holmes 2012 Ireland	2 IRAPs. Relation of disgusting vs. pleasant images to: 'primary' disgust responses (IRAP1) and 'secondary' appraisals of distress (IRAP2)	33 UGs and PGs					IRAP2 responding predicted performance on a behavioural approach task, over and above explicit anxiety and IRAP1 responses, as predicted		For both IRAPs, responding was correlated with an explicit measure of obsessive-compulsive tendencies

Author/Date/Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Nicholson et al. In press Ireland	Relation of disgusting vs. pleasant images to positive vs. negative appraisals	44 UGs and PGs							IRAP responding (disgusting-negative bias) was correlated with 2 of 3 explicit measures of obsessive-compulsive tendencies (associations survived when anxiety was controlled for in regression model)
Nolan et al. In press Ireland	Relation of images of overweight vs. thin people to intelligence-evaluative terms	18 UGs and individuals known to researcher Originally 21; 3 failed to meet IRAP criteria			Male pro-slim bias, not evident in females.			Unexpectedly, and in opposition to discriminant findings of Roddy et al (2010, 2012), IRAP responding was correlated with explicit responses (pro-slim/anti-fat bias)	
O'Toole & Barnes-Holmes 2009 Ireland	Relation of words in terms of before/after and similar/different	55 UGs Originally 62; 6 failed to meet IRAP criteria, 1 was excluded due to dyslexia							IRAP performance (faster response speed and smaller difference-score) was positively correlated with IQ, as predicted

Author/Date/ Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Parling et al. 2012 Sweden	Relations among self/other, thin/fat, and evaluative terms. Relations of self-thin and self-fat terns to evaluative terms ('good' vs. 'bad')	34 individuals. 2 groups: clinical AN (n=17) and age-/gender-matched controls (n=17)			IRAP responding differed between clinical and control groups (stronger self-fat-bad bias in the clinical group)			Generally little correlation between IRAP and explicit responses.	Exceptionally, there was a correlation between responding on one trial-type of the IRAP and a parallel explicit response
Power et al. 2009 Ireland (=[S3] from 2006 paper)	Relative likeability of social groups	12 Irish participants						IRAP responses diverged from explicit preferences, as predicted	
Roddy et al. 2010 Ireland	Relation of images of average- vs. over-weight persons to evaluative terms	64 UGs and PGs Originally 80; 16 failed to meet IRAP criteria		IRAP responding was concordant with IAT responding (overall pro-slim/anti-fat bias). Overall *D*-IAT and *D*-IRAP scores were not correlated, but *D*-IAT was correlated with slim-bad trial-type of the IRAP			IRAP responding had marginal incremental validity for predicting behavioural intention towards an overweight person (hypothetical scenario). IRAP predicted more variance than parallel IAT	IRAP responding indicated a pro-slim bias that was not evident in explicit responses. No significant correlations between IRAP and explicit measures	
Roddy et al. 2011 Ireland	Relation of images of average- vs. over-weight persons to evaluative terms	64 UGs Originally 78; 14 failed to meet IRAP criteria		IRAP responding was concordant with IAT responding (overall pro-slim/anti-fat bias)			IRAP responding predicted behavioural intention towards an overweight person (hypothetical scenario)	IRAP responding indicated a pro-slim bias that was not evident in explicit responses	IRAP responding was concordant with facial EMG responding
Author/Date/ Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Scheel et al. 2011	Relation of male	68 female UGs.						Differential sensitivity	

USA	sexuality (gay vs. straight) to stereotypical trait-descriptive terms	No practice criteria used.						of IRAP vs. explicit measure to order effects, in line with predictions.	
Stockwell et al. 2010 USA	Relation of sexual terms (mainstream vs. BDSM) to evaluative terms (‘sick’ vs. ‘healthy’)	17 participants. 2 groups: individuals with BDSM/fetish interests (n=9) and graduate students (n=8) Originally 22; 5 failed to meet IRAP criteria			IRAP responding differed between BDSM/fetish and graduate groups (stronger pro-BDSM bias in the BDSM group)				IRAP responding was correlated with explicit measures
Timko et al. 2010 USA Report of two studies	[s1]. Relation of self to body-shape terms [s2]. Relation of self to evaluative descriptors (in terms of intelligence, appearance, and friendliness)	[s1]. 50 female UGs Originally 54; 3 failed to meet IRAP criteria (65% accuracy), 1 encountered computer error [s2]. 93 UGs Originally 100; 6 failed to meet IRAP criteria; 1 encountered computer error							[s1]. IRAP responding (overall and trial-specific) was correlated with a number of explicit measures [s2]. IRAP responding to one trial-type (‘I am [positive word]’) correlated with some explicit measures. No relationships with overall *D*-IRAP or other trial-type *D*-scores.

Author/Date/Location/Notes	IRAP	Participants	Reliability	Convergent	Contrasted groups	Experimental	Predictive	Discriminant	Concurrent
Vahey et al. 2010 USA	Relation of smoking to words indicating social acceptance or rejection [Participants first completed a practice IRAP, relationally evaluating words as pleasant vs. unpleasant]	16 11-19 year-old students. 2 groups: smokers (n=8) and non-smokers (n=8)			No significant difference in direct comparison but different patterns of responding were evident (e.g., significant smoker-acceptance coordination in smokers but not non-smokers)				
Vahey et al. 2009 Ireland	Relation of participant's own name to (positive and negative) evaluative terms	43 participants. 3 groups: UGs (n=24), main block prisoners (n=13), and open area prisoners (n=6). Originally 51 participants (30; 15; 6); 8 failed to meet IRAP criteria (70% accuracy on test blocks, further to 80% practice)			IRAP responding differed between the main block prisoners and other groups, consistent with expected group differences (in terms of self-esteem)				IRAP responding was correlated with an explicit measure (collapsing groups). Concordant group differences in IRAP and explicit responses

Note. Reliability is (Spearman-Brown corrected) split-half reliability unless stated otherwise.

Table 2

Stem and leaf plots of split-half IRAP reliability coefficients (k=9)

Stem	Leaf
.8	1, 5
.7	2, 2
.6	0, 4
.5	
.4	4, 1
.3	
.2	3

- The IRAP seems to demonstrate shared and unique variability expected of implicit measures
- Important questions of pragmatic validity require further investigation
- Meta-analysis showed that reliability was comparable with other implicit measures
- However, evidence for validity is partially obscured by questions around reliability
- Demonstration of IRAP effects may elucidate understanding of implicit responses