Creating an Implicit Measure of Cognition More Suited to Applied Research: A Test of the Mixed Trial – Implicit Relational Assessment Procedure (MT-IRAP)

Michael E. Levin, Steven C. Hayes & Thomas Waltz

Abstract

The Implicit Relational Assessment Procedure (IRAP) is a promising tool for measuring implicit cognitions in applied research. However, the need for training and block effects can limit its capacity to assess effects with individual stimuli and participants, both of which are important for applied research. We developed a modified IRAP, the Mixed Trial – IRAP (MT-IRAP), in an attempt to correct for these problems. The MT-IRAP was tested with 58 undergraduate students using conventional good/bad words, emotion words, and words describing substance abusers. We found consistent, significant MT-IRAP effects at both a word list and individual word level and somewhat consistent effects at an individual participant level. The applied utility of the measure was supported by observed relationships between MT-IRAP effects and self-reported experiential avoidance and attitudes towards substance abusers. The MT-IRAP may provide an implicit cognition assessment tool that can be used with less training, and that provides consistent effects for specific stimuli.
Keywords: implicit measures, implicit attitudes, Implicit Association Test, Implicit Relational Assessment Procedure, Relational Frame Theory, experiential avoidance, substance abuse

Implicit cognition measures such as the Implicit Association Test (IAT; Greenwald, McGhee, &,Schwartz, 1998) have become increasingly common in many areas including research on prejudice, consumer preferences, political attitudes, psychopathology, and personality traits (Greenwald, Poehlman, Uhlmann, & Banaji, 2009), in part to avoid the problems of self- report such as susceptibility to self -presentation biases and introspective limits (Greenwald & Banaji, 1995). Implicit measures provide important additional information to explicit assessments, particularly in domains heavily affected by social desirability (Greenwald et al., 2009) or with automatic and spontaneous behaviors (e.g., Asendorpf, Banse, & Mücke, 2002). These measures have not become common in clinical settings, however, due to their procedural characteristics.

The IAT, currently the most popular method, relies on the finding that individuals are generally faster at sorting stimuli based on two concepts to the same response key when these concepts are associated than when they are not. For example, an individual may be faster at sorting words related to "flower" and "good" to the same key than at sorting words related to "flower" and "bad" to the same key. There are hundreds of studies on the IAT (Greenwald et al., 2009), but it only assesses the relative strength of target concepts (De Houwer, 2002), which greatly limits its applied use. For example, if the IAT shows faster responding with flower-good/insect-bad trials than flower-bad/insect-good trials, it is unclear whether the effect is due to a flower-good association, insect-bad association or some relative contribution of both. The IAT design also limits its applicability to domains that go beyond simple associations and bipolar categories, which is often the case with the kinds of beliefs and attitudes applied issues present. Researchers have been working on a variety of alternative IAT designs (e.g., Cohen, Beck, Brown, & Najolia, 2010; Karpinski & Steinman, 2006; Nosek & Banaji, 2001), but none yet overcome these problems.

Relational Frame Theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001) researchers have developed an alternative measure, the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006). Participants are asked to select a relation (e.g., similar/different) between a target stimulus (e.g., substance user words) and a label stimulus (i.e., good/bad) in a series of trials. Two types of trial blocks are used, one in which the verbal relations are consistent with the participants' history of relating stimuli (e.g., addict is similar to bad) and the other where the responses are inconsistent (e.g., addict is similar to good). Participants are trained to emit these two opposing types of sorts (i.e., consistent and inconsistent responses) through an alternating series of practice trials. The difference in response latency between consistent and inconsistent trial blocks in subsequent testing is used to detect the implicit effect. The IRAP is more flexible than the IAT, particularly as it can be used to examine specific implicit relations with a target concept, rather than only relative associations, and to assess a broad range of relations beyond associations.

The IRAP demonstrates predicted differences between known groups in a wide variety of areas including some of applied relevance such as self-esteem (Scanlon, Barnes-Holmes, Barnes-Holmes, & Stewart, under review; Vahey, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009), attitudes towards different nationalities (Power, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009), and sexual attitudes (Dawson, Barnes-Holmes, Gresswell, Hart, & Gore, 2009) among many others. The IRAP diverges from explicit self- reports in predicted ways (Barnes-Holmes, Murphy, Barnes-Holmes, & Stewart, 2010; Power et al., 2009), and is sensitive to variables that go beyond explicit reports (Roddy, Stewart & Barnes-Holmes, 2010). It is difficult to fake (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007), is internally consistent (Barnes-Holmes et al., 2009; Barnes-Holmes, Murtagh et al., 2010), and can be used to assess the effects of interventions (Cullen, Barnes-Holmes, Barnes-Holmes, & Stewart, 2009). Furthermore, in some contexts the IRAP is superior to the IAT in predicting behavioral intentions above and beyond explicit attitudes (Roddy, Stewart, & Barnes-Holmes, 2010).

An analytic strength of the IRAP is that it can distinguish the individual components of an overall relational network. For example, in "fat bias" the IRAP can distinguish the implicit effect for skinny-good and fat-bad separately, not just skinny-good/fat-bad as an entire set. A study by Barnes-Holmes, Murtagh, and colleagues (2010) demonstrates this methodological strength, finding that both vegetarians and meat eaters demonstrate a provegetable bias on the IRAP, but that vegetarians also have a significant antimeat bias, while meat eaters do not have a promeat bias.

While this is progress, more needs to be done to make the IRAP fully useful in applied settings. Both IAT and IRAP studies have exclusively focused on detecting effects at a group level, but for applied use most participants need to show the effect at the level of the individual. The IRAP shows a practice effect where differences in response latency between trial types change over time in the test (e.g., Power et al., 2009) and IRAP effects differ depending on whether testing begins with a consistent or inconsistent trial block (e.g., Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008). These features are undesirable as they introduce additional sources of variance that make it more difficult to identify differences in response latency attributable to implicit effects. This is particularly the case if it is important to give the assessment repeatedly, as it might be in an applied setting, because it is difficult to determine whether and to what degree changes in IRAP effects across assessments is attributable to these alternative sources of variance. IRAP studies that have failed to find order and practice effects have at times been very underpowered, such as conducting a 2X3X2 MANOVA with a sample of 16 participants (Barnes-Holmes et al., 2008), and thus are unconvincing. The IRAP also performs better with sets of stimuli relating to a concept than individual stimuli, but many applied uses require information at the individual stimulus level. IRAP researchers have pointed to some of these issues as important areas for research and measure development (Barnes-Holmes, Barnes-Holmes, Stewart ,& Boles, 2010).

These limitations with the IRAP seem to emerge from the need to compare differences in response latency between blocks of consistent trials and blocks of inconsistent trials. Comparing blocks of trials mean that differences by trial type in response latency (i.e., the implicit effect) can be confounded with other sources of variance such as changes in response latency over time and the order of trial blocks completed. Although randomizing the block sequence somewhat protects against confounds due to order and practice effects in group designs, it does not solve the problem at an individual level. Corrective feedback must also be given to train the test during practice trials and to maintain the block effect during testing trials, which may confound results with training effects as it is unclear whether observed differences in response latency emerge due to corrective feedback or would naturally occur. Furthermore, to give feedback the researcher predetermines the relation between stimuli such that individual stimuli within a conceptual category have the same relevant relational functions. However, for many individuals, certain stimuli could have unique functions (i.e., alcohol is good, but heroin and cocaine are bad). These features may reduce the sensitivity and reliability of the IRAP in detecting individual stimulus and participant effects.

In principle, behavioral approaches are ultimately focused on functional stimulus classes, not individual stimuli, but ironically it is harder to get to that level with methods that are based on list by list comparisons. If individual stimuli evoke different responses or are impacted by different contextual conditions they are not fully members of the same functional class. However, determining that requires methods that allow the impact of contextual conditions on individual stimuli to be known. Thus, it is a mistake to think of lists as functional classes merely by the demonstration of an IRAP effect and implicit research would benefit from a measure capable of examining implicit effects with individual stimuli.

The current study sought to develop and test a modified version of the IRAP that corrected these potential limitations in order to enhance the capacity to detect individual stimulus and participant effects. What we refer to here as the Mixed Trial-IRAP (MT-IRAP), combines consistent and inconsistent trial types into each test block using the conventional contextual cues “truth” and “lie” to indicate whether a participant should make a consistent or inconsistent relation for a given trial. Comparisons between consistent and inconsistent blocks can thus be made continuously rather than presenting a complete block of one trial type and then a complete block of the other. In addition, the use of “truth” and “lie” for indicating trial type removes the need for practice training with test stimuli or corrective feedback during testing trials and the direction of the response does not need to be specified beforehand by the experimenter.

The current study examined the utility of the MT-IRAP in detecting participants’ implicit cognitions at both a group and individual level and with both overall list and individual stimulus effects. The MT-IRAP was tested using conventional good and bad words as well as with two applied problems: the detection of stigmatizing words related to substance abuse and positive/negative evaluations of emotions. The validity of the MT-IRAP was examined in relation to explicit self-report questionnaires assessing attitudes towards substance abusers and how individuals’ relate to their emotions.

Method

Participants

A convenience sample of undergraduate psychology students were recruited from the University of Nevada, Reno. A total of 58 undergraduate students participated in the study. The sample included 38 females (65.5%) and 20 males (34.5%).

Measures

MT-IRAP. The MT-IRAP was adapted from the standard IRAP developed by Barnes-Holmes and colleagues (2006). This measure consists of a series of trials where participants select the relation between two stimuli. The label stimulus is usually a conventional dichotomous variable that has common functions (i.e., good/bad, pleasant/unpleasant). The target stimulus usually represents the concept(s) of particular interest to the experimenter, (i.e., emotions, descriptions of substance abusers). For each trial participants use a keyboard to select one of two relational cues specifying the relation of the target stimulus to the label stimulus (i.e., similar/different). Half of the trials ask participants to tell the "truth" (make a consistent relation) and the other half to "lie" (make an inconsistent relation). For each target stimulus there are four potential combinations of label stimuli and trial types (i.e., "Addict"/"Good"/Truth, "Addict"/"Good"/Lie, "Addict"/"Bad"/Truth, "Addict"/"Bad"/Lie). The sequence of trial presentations for each label stimulus, target stimulus, and trial type combination is random within each block. Each trial begins with a 1 second presentation of the trial type ("truth" or "lie") followed by the presentation of the label stimulus, test stimulus, and relational response options. A 400 millisecond pause occurs after a response is made, followed by the next trial. When a participant demonstrates inconsistent responding (i.e., sorting "Addict" as similar to "Good" in both truth and lie trials), the less frequent response direction is counted as an error. For example, the determination of which response is the correct response may be "Addict" as similar to "Good" in truth trials or "Addict" as similar to "Bad" in truth trials depending on that participants' pattern of responding, rather than being predefined by the experimenter. The difference in response latency, and potentially error rate, between truth and lie trials can be used to infer implicit verbal relations with an overall target concept as well as specific individual stimuli. A graphical depiction of the task is presented in Figure 1. The MT-IRAP program is available upon request from the primary author.

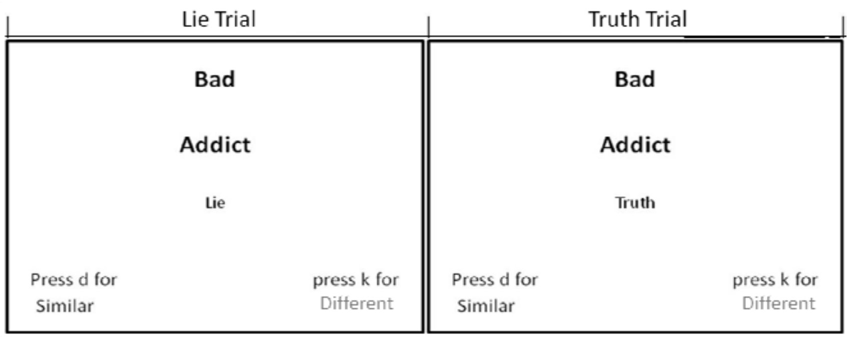

Figure 1. MT-IRAP Example.

Three sets of target stimuli were used in the current study. These sets were tested separately in three sequential test blocks, with the same order across participants. The same label stimuli, good and bad, were used for each set. The first test examined verbal relations of good/bad with conventional positive and negative valenced words. Words were selected based on past semantic research (Osgood, Suci, & Tannenbaum, 1957; Toglia & Battig, 1978) in order to ensure that the vast majority of participants would relate stimuli in the expected direction. The second test examined verbal relations of good/bad with positive and negative emotion words. The third test used descriptions of substance abusers selected from a previous study examining words commonly used in substance abuse treatment (Waltz et al., in preparation). A list of the stimuli used in the study is provided in Table 1.

Table 1 - *List of Stimuli*

Valenced Words (Test 1)		Emotion Words (Test 2)		Substance Abuser Words (Test 3)
Beautiful Foul		Happy	Sad	Drug Addiction
Freedom Awful		Cheerful	Anxious	Drug Problem
Nice Ugly		Love	Hate	Substance Abuse
				Alcoholic
				Addict
				Drug User

Two different response option sets were tested in the study. The majority of participants (n = 32) were given the response options similar/different and the remainder (n = 26) were given the options yes/no. Similar/different has been commonly used in IRAP studies (Barnes-Holmes, Barnes-Holmes et al., 2010). The response options were changed to yes/no midway through the study out of concern that evaluating stimuli on the basis of them being "similar" or "different" to "good" or "bad" would be more ambiguous than responses of "yes" and "no." There were no significant differences on response latencies or error rates at the list level between these two versions of the MT-IRAP ($p > .05$) so participants were combined for all analyses.

Explicit Self-Report Measures. A stigma measure, the Community Attitudes towards Substance Abusers (CASA; Hayes, Wilson et al., 2004), was included to assess the convergent validity and applied utility of the MT-IRAP when used with descriptions of substance abusers. The CASA assesses positive and negative attitudes towards substance abusers on four subscales; Benevolence, Social Restrictiveness, Community Approach, and Authoritarianism. The scale consists of 40 items rated on a 7 point scale ranging from 1 ("very strongly disagree") to 7 ("very strongly agree"). Studies using the CASA have found adequate reliability and validity for the scale (Hayes, Wilson et al., 2004; Vilardaga et al., under review).

To further examine the validity and applied utility of the MT-IRAP, the study examined differences in MT-IRAP effects[1] on emotion words based on a mean split of individuals higher and lower in experiential avoidance. Experiential avoidance is the rigid and inflexible engagement in behaviors to avoid, escape, or otherwise control aversive thoughts, feelings, and sensations, despite the negative consequences of doing so (Hayes, Wilson, Gifford, Follette, & Strosahl, 1996). Part of this process involves the tendency to become cognitively entangled in evaluations of one's emotional experiences, which should lead to observed differences between groups on MT-IRAP effects with positive and negative emotions words.

Experiential avoidance was measured with the Acceptance and Action Questionnaire – II (AAQ-II; Bond et al., under review; Hayes, Strosahl et al., 2004), which is a 10 item scale with responses ranging on a 7 point scale from 1 ("never true") to 7 ("always true"). Studies have found adequate reliability and validity with this scale in college populations (Bond et al., under review; Hayes, Strosahl et al., 2004). In addition, studies have found that groups high and low on the AAQ demonstrate predicted differences in laboratory-based behavioral measures (e.g., Feldner, Zvolensky, Eifert, & Spira, et al., 2003; Zettle et al., 2005; Zettle, Petersen, Hocker, & Provines, 2007).

Procedure

Participants completed all study procedures in a private room in a research laboratory. After entering, they were asked to complete the series of self-report questionnaires followed by the MT-IRAP. The experimenter then left the room and remained absent during completion of the measures.

The MT-IRAP program began with a series of animated instructions to familiarize participants with the procedure. These instructions included a description of the sorting task with examples, the importance of responding as quickly and accurately as possible, and the criteria for completing the practice and test trials. Participants were then asked to complete two series of practice blocks. In the first practice block series, only the truth contextual cue was presented in order to familiarize participants with the standard IRAP procedure, prior to introducing the mixed trial method. Four strongly valenced words, which were not used in any of the test blocks, were presented in random order during the practice trial ("pleasant", "excellent", "rotten", and "terrible") along with the label stimuli good/bad. These were the only trials where the experimenters determined which relations were "correct" a priori. This served to ensure that participants could quickly learn fast and consistent responding in the task. If a participant responded incorrectly a red "x" would appear and could be removed by making the correct response. In order to proceed to the second practice block series participants had to sort the four stimuli four times each (16 trials total per block) with an accuracy rate of at least 80% and an average response latency of 2 seconds or faster. This practice criterion was selected based on previous IRAP research, which found that requiring a high accuracy and short latency leads to stronger IRAP effects during test blocks (Barnes-Holmes, Murphy et al., 2010). Feedback regarding average response latency and accuracy was given after each practice block. Participants were given six attempts to pass the first practice phase, after which those failing to meet the criteria were excused from the study. The second phase used an identical procedure except that half of the trials presented the truth contextual cue and the other half presented the lie cue. If participants passed both phases of the practice trial they then proceeded to the test phase.

In the test phase participants were instructed to try to maintain the same speed and accuracy achieved in the practice trials. Similar to other IRAP studies, the speed and accuracy criterion were no longer required and participants did not receive any feedback regarding performance on the test blocks. Participants were also told that they would no longer receive corrective feedback, but that if they responded inconsistently (i.e., sorting a word as bad in some truth trials and as good in others) the study would take longer to complete. Falling below 75% consistency on a given word caused the test block to reset so that the participant had to start again at the beginning for that block of trials. The test phase consisted of six test blocks, with two identical blocks for each word set. Each block consisted of 72 sorting trials, with six truth and six lie trials for each of the six stimuli. Test blocks were completed in the same order for each participant.

Results

Data Preparation

Prior to analyses the data were transformed to remove extreme outliers and consistency errors. Participants with error rates above 25% (n = 4) or an average response latency above 3 seconds (n = 1) were removed from subsequent analyses. These criteria were based on similar procedures used in previous IRAP studies (Barnes-Holmes, Murphy et al., 2010; Barnes-Holmes et al., 2009) in order to remove participants who do not appear to follow the basic guidelines of responding quickly and accurately. Trials with response latencies over 10 seconds were removed as extreme outliers based on recommended procedures (Barnes-Holmes, Barnes-Holmes et al., 2010).

IRAP and IAT researchers commonly include error trials in analyses, adding a natural occurring penalty score to the measured response latency by the additional time required for the participant to emit the subsequent correct response after receiving an error message. However, some researchers have raised concerns about this method since a penalty score confounds response latency differences with accuracy differences (Gavin, Roche, & Ruiz, 2008). Thus, the current study excluded error trials (e.g., the infrequent response pattern for a given participant and stimulus) from response latency analyses and only used correct trials (i.e., the dominant response pattern for a given participant and stimulus). In addition, if a stimulus was not sorted consistently at least 65% of the time it was excluded from analyses ("Awful" was excluded for two participants and "Anxious" for three participants).

Ten of the 58 participants did not pass the practice phase due to consistently slow or incorrect responding and were excluded from further test phases. Of those who passed the practice phases, it took on average 1.25 attempts to pass phase 1 (SD = .54, Mode = 1) and 2.55 attempts to pass phase 2 (SD = 1.62, Mode = 1). An additional 5 of the remaining 48 participants were removed due to high error rates or response latencies. Thus, the final sample consisted of 43 participants (74.1% of the original 58 participants). These rates do not appear to differ significantly from reported dropout rates for similar practice phase and test performance criteria in previously published IRAP studies (e.g., Barnes-Holmes, Murphy et al., 2010; Vahey et al., 2009). For example, Vahey and colleagues (2009) found that 6 out of 30 undergraduates (20%) did not pass a 70% accuracy criterion in the test phase. Another study by Barnes-Holmes, Murphy and colleagues (2010) found that 7 out of 38 (18%) did not pass a 3,000 ms average response latency and 80% accuracy criterion for practice or test phases and 5 out of 24 (21%) did not pass a stricter 2,000 ms and 80% accuracy criterion.

Based on recommendations for data transformation with the IRAP (Barnes-Holmes, Barnes-Holmes et al., 2010) and IAT (Greenwald, Nosek & Banaji, 2003), a MT-IRAP score was calculated as the difference between response latencies on truth and lie trials for each word list and individual word using the Cohen’s *d* formula; *d* = (mean of lie trials – mean of truth trials)/ pooled standard deviation for truth and lie trials. Direction of responding was set so that a positive MT-IRAP score indicated that responses on the lie trials took longer than the truth trials.

Testing for the MT-IRAP Effect

The mean and standard deviation for response latencies on correct truth and lie trials for each word list and specific word are provided in Table 2. In order to test for the MT-IRAP effect, planned one-sample *t*-test analyses were conducted to test whether MT-IRAP scores were significantly different from 0 for each word list and individual word. Significant effects were observed with each word list and for 16 of the 18 individual words, (See Table 2). In all cases lie trials had significantly longer response latencies than truth trials. No MT-IRAP effect was observed for two words (“drug user” and “hate”).

Table 2 - Planned One-Sample *t*-tests of the IRAP Score

	Truth trials	Lie trials	IRAP Score	
Comparison	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*t*-score (*df*)
All Words	1797.37 (302.47)	2057.76 (379.87)	.24 (.17)	8.93 (39)***
Good Words	1744.14 (291.46)	2098.23 (376.37)	.36 (.20)	11.83 (42)***
Bad Words	1844.57 (344.65)	2056.69 (422.36)	.19 (.20)	6.03 (39)***
Test 1	1907.96 (344.06)	2250.09 (433.56)	.33 (.21)	10.50 (42)***
Test 2	1788.48 (312.72)	2039.09 (349.09)	.28 (.25)	7.34 (42)***
Test 3	1766.29 (479.71)	1957.35 (585.38)	.16 (.25)	4.08 (39)***
Test 1 Good Words	1791.62 (330.13)	2230.94 (483.51)	.42 (.25)	10.92 (42)***
Test 1 Bad Words	2029.99 (389.75)	2275.06 (412.02)	.26 (.28)	5.95 (42)***
Test 2 Good Words	1691.61 (327.81)	1979.27 (365.69)	.34 (.32)	6.99 (42)***
Test 2 Bad Words	1901.27 (327.79)	2105.91 (377.89)	.22 (.31)	4.68 (42)***
Awful	1993.66 (413.02)	2220.15 (451.95)	.25 (.39)	4.03 (40)***
Beautiful	1804.32 (455.72)	2212.13 (494.41)	.43 (.36)	7.77 (42)***
Ugly	2012.04 (477.98)	2287.33 (534.86)	.32 (.53)	3.94 (42)***
Foul	2083.44 (482.16)	2293.94 (448.95)	.22 (.39)	3.67 (42)**
Nice	1769.51 (336.75)	2190.84 (523.27)	.43 (.42)	6.82 (42)***
Freedom	1807.10 (404.58)	2313.21 (631.48)	.47 (.43)	7.12 (42)***
Anxious	1982.96 (522.26)	2215.37 (442.96)	.26 (.59)	2.81 (39)**
Cheerful	1759.20 (371.49)	2046.93 (433.48)	.32 (.36)	5.81 (42)***
Happy	1676.74 (506.52)	1915.70 (431.67)	.37 (.45)	5.39 (42)***
Hate	1924.02 (414.15)	2050.96 (525.48)	.12 (.52)	1.45 (42)
Love	1642.68 (389.76)	1970.47 (416.10)	.44 (.47)	6.15 (42)***
Sad	1793.17 (347.17)	2061.49 (377.28)	.30 (.49)	4.01 (42)***
Addict	1691.81 (563.72)	1957.70 (582.96)	.27 (.44)	3.80 (39)***
Alcoholic	1690.18 (597.18)	1859.86 (613.69)	.15 (.43)	2.21 (39)*
Drug User	1823.51 (567.23)	1894.02 (746.08)	.06 (.49)	0.72 (39)
Drug Addiction	1757.66 (407.08)	1946.94 (696.66)	.16 (.43)	2.32 (39)*
Drug Problem	1754.89 (517.09)	1976.80 (748.17)	.18 (.44)	2.58 (39)*
Substance Abuse	1794.60 (549.81)	2116.35 (828.60)	.21 (.42)	3.23 (39)**

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

The consistency of the MT-IRAP effect across participants was examined for each word list and individual word. Participants consistently related stimuli to the expected valence for truth and lie trials. Good words in tests 1 and 2 were related as good in truth trials, bad words in test 1 and 2 as well as the substance abuser words in test 3 were related as bad in truth trials. Thus, positive MT-IRAP scores are always in the expected valence relation.

A box plot of the MT-IRAP scores for each word list is presented in Figure 2 and for each word in Figure 3. Between 97.7% and 75.0% of participants demonstrated an MT-IRAP score in the expected direction at the word list level. At the individual word level, between 86.0% and 45.0% of participants demonstrated an MT-IRAP score in the expected direction.

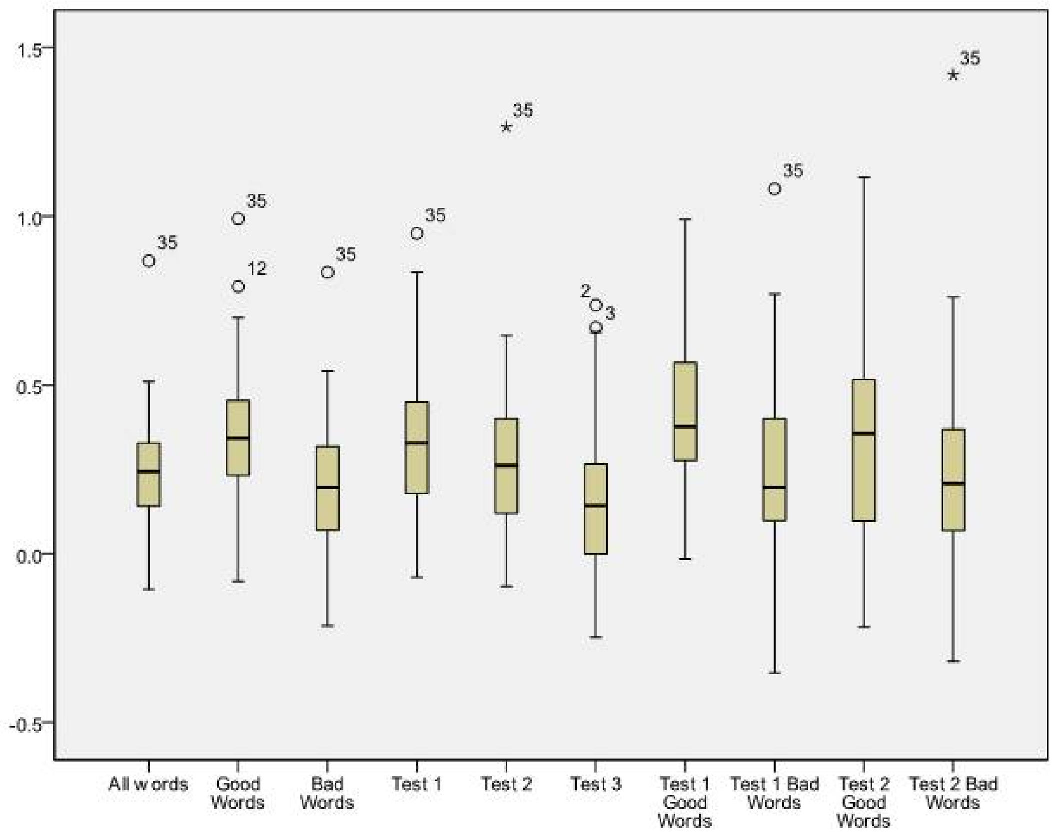

Figure 2. Box plot of IRAP Scores on Word Lists by Participant. Positive effect sizes demonstrate expected IRAP effect with longer response latency on lie trials.

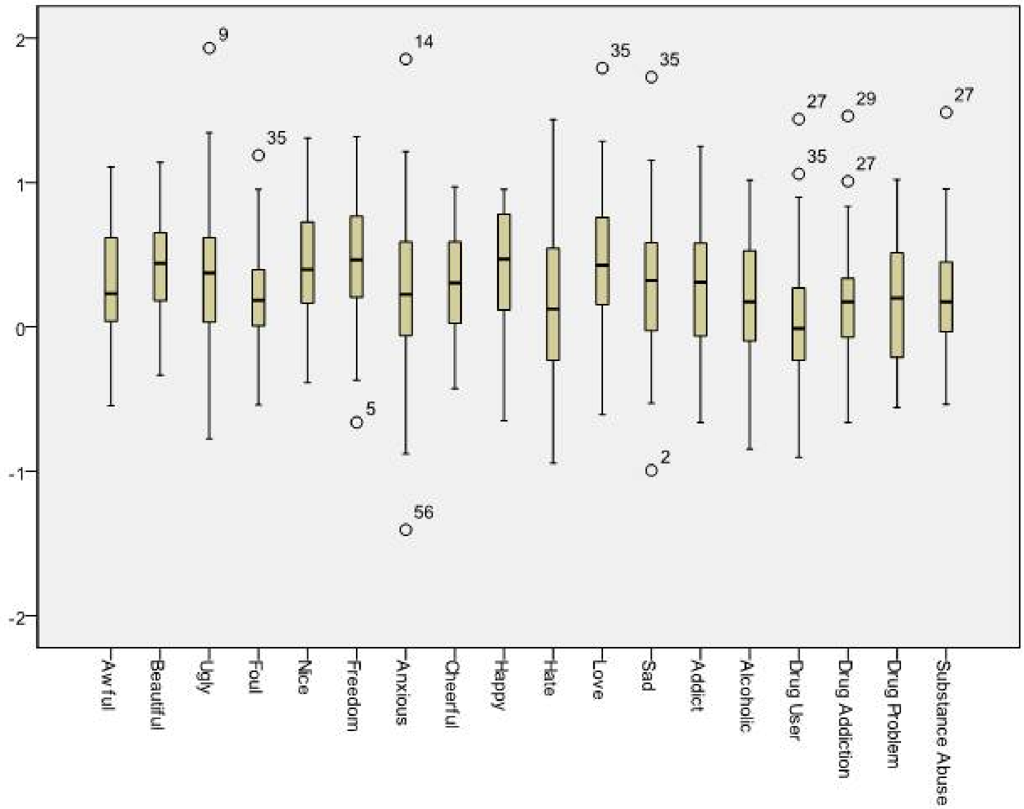

Figure 3. Box plot of IRAP Scores on Individual Words by Participant. Positive effect sizes demonstrate expected IRAP effect with longer response latency on lie trials.

Error Rates

One potential concern with the MT-IRAP procedure is that the method might produce a high error rate due to its relative complexity. However, the rate of participants who were unable to meet practice and test block performance criteria appeared relatively similar to previous IRAP studies with dropout rates around 20% (e.g., Barnes-Holmes, Murphy et al., 2010; Vahey et al., 2009). In addition, the mean overall error rate was only 9.38% (SD = 5.72), suggesting the measure can be completed with a reasonable level of accuracy.

Based on suggestions by Gavin, Roche, and Ruiz (2008), differences in error rates between truth and lie trials were examined as another potential measure of implicit verbal relations. Means and standard deviations for error rates on truth and lie trials are reported in Table 3. Planned paired t-test analyses were run comparing the error rate between truth and lie trials for each word list and individual word (See Table 3). Significant effects were found for each word list, except negative valenced emotion words, and 9 of the 18 individual words, with 4 other words approaching significance ($p < .10$). All significant effects were such that there was a higher error rate with lie trials than truth trials. Five words ("Hate", "Sad", "Addict", "Drug Addiction", and "Drug Problem") did not show the expected MT-IRAP effect with error rates.

Table 3 - *Planned Paired t-tests Comparing Error Rates on Truth and Lie Trials*

Comparison	Truth trial error rate	Lie trial error rate	*t*-score
All Words	.08 (.05)	.11 (.07)	7.18 (39)***
Good Words	.05 (.04)	.10 (.08)	6.24 (42)***
Bad Words	.09 (.05)	.11 (.07)	4.58 (39)***
Test 1	.08 (.06)	.14 (.09)	6.40 (42)***
Test 2	.08 (.06)	.10 (.07)	3.29 (42)**
Test 3	.06 (.05)	.08 (.07)	3.10 (39)**
Test 1 Good Words	.05 (.06)	.11 (.09)	5.07 (42)***
Test 1 Bad Words	.11 (.08)	.16 (.11)	4.26 (42)***
Test 2 Good Words	.04 (.05)	.08 (.08)	5.24 (42)***
Test 2 Bad Words	.12 (.09)	.12 (.10)	.31 (42)
Awful	.11 (.10)	.16 (.11)	2.38 (40)*
Beautiful	.04 (.06)	.08 (.09)	3.31 (42)**
Ugly	.10 (.10)	.16 (.13)	2.79 (42)**
Foul	.09 (.09)	.14 (.11)	2.15 (42)*
Nice	.05 (.07)	.10 (.10)	3.79 (42)***
Freedom	.05 (.06)	.13 (.12)	3.67 (42)**
Anxious	.09 (.14)	.12 (.15)	1.81 (39)+
Cheerful	.03 (.04)	.08 (.09)	3.22 (42)**
Happy	.03 (.06)	.07 (.07)	3.32 (42)**
Hate	.12 (.11)	.10 (.09)	-1.63 (42)
Love	.04 (.06)	.08 (.09)	2.62 (42)*
Sad	.10 (.11)	.11 (.10)	0.57 (42)
Addict	.07 (.07)	.07 (.08)	0.43 (39)
Alcoholic	.05 (.09)	.08 (.09)	1.89 (39)+
Drug User	.05 (.07)	.08 (.09)	1.70 (39)+
Drug Addiction	.07 (.08)	.09 (.10)	1.02 (39)
Drug Problem	.05 (.06)	.07 (.10)	1.52 (39)
Substance Abuse	.06 (.08)	.09 (.11)	1.84 (39)+

Note. + p < .10, * $p < .05$, ** $p < .01$, *** $p < .001$

Split-Half Reliability

The reliability of the MT-IRAP was examined by conducting a split-half reliability analysis. Pearson correlations were conducted between MT-IRAP scores for even and odd trials overall and for each test. Significant large correlations were observed between even and odd trials for the overall MT-IRAP score, $r(41) = .54$, $p < .001$, and test 2 emotion words MT-IRAP score $r(41) = .55$, $p < .001$. Medium correlations approaching significance were observed between even and odd trials for the test 1 good/bad words MT-IRAP score, $r(41) = .26$, $p = .09$, and test 3 substance use words MT-IRAP score, $r(38) = .31$, $p = .05$. The overall split-half reliability appeared adequate, particularly given a response latency measure (Nosek et al., 2006), although reliability varied more for each test.

Relationship Between Explicit and Implicit Attitudes Towards Substance Abusers

Pearson correlations were conducted examining the relation between the CASA and MT-IRAP scores on substance abuser words in test 3 (See Table 4). Two of the CASA subscales (Benevolence and Community Approach) would be expected to have negative correlations with MT-IRAP scores; while two

(Authoritarianism and Social Restrictiveness) would be expected to have positive correlations. Summing across the six IRAP stimuli, 21 of the 24 correlations fit that pattern (Fisher's exact, $p < .001$), but the correlations were generally weak. MT-IRAP scores for "Alcoholic" correlated significantly with the Social Restrictiveness subscale and showed a trend ($p < .10$) with the Community Approach subscale. MT-IRAP scores for "Drug User" correlated significantly with the Benevolence subscale. At the list level the CASA total, Benevolence subscale, and Social Restriction subscale scores showed a trend toward correlation with the overall test 3 MT-IRAP score.

Table 4 - *Correlations Between IRAP Scores and Explicit Ratings Towards Substance Abusers*

IRAP		CASA Total		Authori-		Benevolence		Social	Co		mmunity
Score	ta			rianism		Re		strictiveness		A	pproach
All Test 3 Stimuli		.275+		.039		-.303+		.306+			-.246
Addict	.0	88	-.0	13	-.0	44	.1	61			-.082
Alcoholic		.286+	.1	70	-.0	96	.3	73*			-.282+
Drug User		.231	-.0	49	-.3	90*	.1	45			-.262
Drug Addiction	.0	96	-.0	56	-.1	91	.0	42	-.1		29
Drug Problem		.170	.0	94	-.2	18	.1	42	-.1		22
Substance Abuse		.146	.0	60	-.2	01	.1	13	-.1		17

Note. + $p < .10$, * $p < .05$

Differences Between Higher and Lower Experiential Avoiders

To further test the validity of the MT-IRAP we compared MT-IRAP scores on test 2 emotion words between higher and lower experiential avoiders as measured by the AAQ-II. The mean AAQ-II score in the current sample was 56.37 (SD = 7.50). Based on this mean, participants were split into higher experiential avoiders (56 and below, n = 20) and lower experiential avoiders (57 and above, n = 23) groups.

Independent sample t-tests were conducted comparing test 2 emotion word MT-IRAP scores between higher and lower experiential avoiders (see Figure 4). There were no significant differences between groups on overall test 2 MT-IRAP scores, $t(41) = 1.59$, $p = .12$, test 2 good emotion words, $t(41) = 1.10$, $p = .28$, or test 2 bad emotion words, $t(41) = 1.62$, $p = .11$. However, significant effects were observed for "Hate", $t(41) = 2.16$, $p < .05$, and "Love", $t(41) = 2.53$, $p < .05$, such that high experiential avoiders demonstrated a stronger bias towards relating "Love" as good and "Hate" as bad. No significant differences were observed with MT-IRAP scores for the other four individual stimuli or any of the error rates between truth and lie trials ($p > .10$).

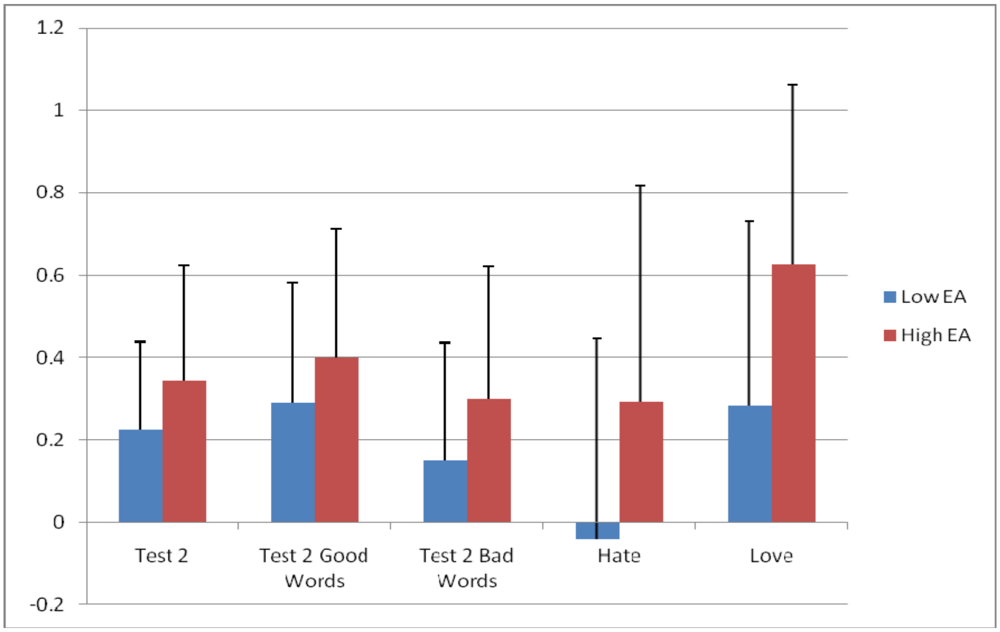

Figure 4. Mean Differences on IRAP Score Between Higher and Lower Experiential Avoiders

A linear regression analysis was conducted to test the predictive validity of emotion word MT-IRAP scores on experiential avoidance group status. A regression was first conducted with “Hate” and “Love” MT-IRAP scores due to the observed group differences on these two stimuli. A significant effect was observed predicting group status, $R^2 = .18$, $F(2, 37) = 4.09$, $p < .05$. When a second step was conducted including the other four emotion words there was no significant increase in predictive validity, R^2 change = 0.01, $F(4, 33) = 0.13$, $p = .97$.

Discussion

The current study sought to develop and test a modified version of the IRAP that could better meet the needs of applied research, particularly the requirement to assess implicit cognitions at both an individual stimulus and participant level. The results supported the utility of the mixed trial design, demonstrating consistent differences between truth and lie trials on response latency, and to a somewhat lesser extent on error rates, at both a word list and individual stimulus level, such that participants took longer to respond and made more errors on lie trials. The measure demonstrated adequate split-half reliability, error rates and dropout rates. The validity and applied utility of the measure was further supported by observed correlations between explicit attitudes towards substance abusers and MT-IRAP scores on substance abuser words as well as differences between higher and lower experiential avoiders on MT-IRAP scores with emotion words. Overall, these preliminary results suggest that the MT-IRAP could be an effective measurement tool for assessing implicit cognitions in an applied context.

Applied researchers are often interested in implicit effects for specific stimuli, rather than just for the overall target concept. Stimuli that show particularly strong or weak effects may be used to inform interventions, such as motivational statements for exercise (Jackson, 2008), or for reducing stigmatization, such as identifying descriptions of substance abusers that evoke the least negative bias (Waltz, Drossel, Hayes, Roget, & Fisher, in preparation). Focusing only on the overall target concept may lead to missing important data. For example, the current study found that higher experiential avoiders have a stronger positive bias with love and negative bias with hate. If only the overall target concept was examined, no differences would have been observed between higher and lower experiential avoiders.

Applied settings also often need implicit measures that are reliable and accurate with individual participants, rather than just at a group level. For example, an implicit measure that could reliably identify individuals high and low in experiential avoidance could have important applications for assessment and treatment. This would also enhance the use of implicit measures in research, such as for designs requiring high levels of precision (e.g., process of change research) and studies examining predictors of ideographic implicit effects.

The IRAP and IAT designs somewhat limit the capacity to examine individual level effects. In particular, the use of block by block comparisons between consistent and inconsistent trials can confound implicit effects with other sources of variance such as practice and order effects. In addition, providing feedback during test trials may confound results with training effects by forcing a trained association rather than assessing an implicit one.

The results of this study suggest that the modifications made with the MT-IRAP provide the necessary precision to examine individual stimulus effects, at least to a degree. It is less clear though whether the MT-IRAP is sufficiently reliable and accurate to detect individual participant effects. This measure may have been an improvement as compared to the IRAP and IAT based on the high percent of participants showing effects in the expected direction with strongly valenced words, but there were still a significant number showing effects in the opposite direction or at too small of a magnitude to be detectable individually. Additional modifications may be necessary to refine the MT-IRAP and future studies are needed that directly compare the MT-IRAP to the IRAP and IAT in detecting individual stimulus and participant effects.

The observed relationships between explicit self-report questionnaires and substance abuser and emotion words provide support for the validity of the MT-IRAP and its utility in assessing applied domains. The observed correlations between MT-IRAP scores on substance abuser words and explicit attitudes towards substance abusers suggests that the larger the difference in response latency between consistent and inconsistent trial types, the stronger the implicit bias. The tendency for participants who are higher on experiential avoidance to relate “love” more to good and “hate” more to bad is consistent with theoretical models of experiential avoidance in which the tendency to become excessively cognitively entangled with evaluations of emotions, both positive and negative, becomes prominent (Hayes et al., 1996). It is unclear whether the lack of effect with other emotion words is attributable to error variance with the MT-IRAP or if the observed effects are unique to these specific stimuli. This finding highlights why having data at the individual stimulus level is valuable. We do not know in this study why “love” and “hate” evoked different responses than other emotion words, but we do know now to ask that question.

There are some possible limitations with the study. First, a number of *t*-tests were conducted to compare response latency and error rates without using an adjusted alpha to correct for type I error (e.g., Bonferroni correction). However, these analyses were planned, examining theoretically-driven predictions in every case. Thus, the risks of type I error inflation are much less significant as in the case of data “fishing” and adjusting the alpha was deemed to be too conservative given the piloting nature of the study.

Although the dropout rates from performance criteria in the practice and test trials appear relatively equivalent to published studies using the IRAP (e.g., Barnes-Holmes, Murphy et al., 2010; Vahey et al., 2009), it is still a significant number of participants (around 25%). This may indicate the performance criterion was too high for the given sample and preparation, there were not enough practice trials, the instructions were insufficient, or additional features are needed to adequately motivate engagement in the task. Future research can manipulate and test these factors in order to enhance participant retention and adherence to the procedures.

The analyses conducted in the current study focused on averaged response latency across a test block. All of the correct trials were combined as long as responses were under 10 seconds. However, response latency data typically does not compose a normal distribution and averaged response latencies can still be significantly affected by outliers. Researchers have pointed out that examining response latency effects in this way can significantly reduce power to detect effects (Whelan, 2008). In addition, these analyses fail to examine important subsets of responses (i.e., immediate responses, average responses, delayed responses) and changes over time within test blocks. Thus, future analyses would benefit from examining the whole distribution of response latencies (Whelan, 2008) and conducting analyses that are more sensitive to these dynamic properties such as mixed regression models.

This study used relatively unambiguous stimuli for the MT-IRAP as evidenced by every participant relating the stimuli in the same way for truth and lie trials. It is possible that the MT-IRAP may not function as well with ambiguous stimuli as participants may encounter difficulties in determining which response is true and which a lie. The use of truth and lie to indicate trial type, rather than relying on training and corrective feedback, is an important difference from the standard IRAP and could be a strength or weakness with the MT-IRAP. Only more research will make that determination. It could be that the MT-IRAP will only be useful with concepts that are polarized, since within participant consistency is needed. Conversely, the MT-IRAP can detect an implicit effect even if the participant's preference is idiosyncratic as compared to other participants. In the traditional IRAP, such an effect is there only in the statistical noise in the data. Further research can examine this potential limitation by using stimuli with multiple or ambiguous relations to a label stimulus.

The current study differed from other IAT and IRAP studies in that only correct trials were used to examine response latency. The observed MT-IRAP effects with only correct trials serves to alleviate concerns that differences in response latency between trial types are really due to the effect of penalty scores attributed to incorrect trials (Gavin et al,, 2008). The study also found a difference on error rates between trial types, suggesting that error rate provides another method for detecting implicit relations. The combination of these two effects in the standard IRAP and IAT may serve to enhance their sensitivity. However, corrective feedback is not provided to participants in test trials with the MT-IRAP so the standard methods for including incorrect trials with the IAT and IRAP cannot be employed. Further studies are needed to determine how to combine these two effects with the MT-IRAP and whether their combination improves the effectiveness of the measure.

Overall, the MT-IRAP appears to be a promising implicit measure for applied research. The use of trial type cues, mixing trial types within blocks, and removing corrective feedback in test blocks may provide advantages over the standard IRAP and IAT. These modifications could reduce the potential impact of order, practice, training, and other method-specific effects that reduce the sensitivity of implicit measures in detecting effects at an individual level. The MT-IRAP does not appear to be significantly more difficult to complete than the standard IRAP and may even be easier to use as participants are given a direct cue for trial type, rather than relying only on training procedures. Continued research focused on developing and testing implicit measures that are sensitive to ideographic stimulus and participant effects, such as the MT-IRAP, could significantly improve the utility of implicit cognition measures in applied research.

References

Asendorpf, J.B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83*, 380–393.

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,* 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*(3), 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. *The Psychological Record, 58,* 497-516.

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.

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 towards meat and vegetables in vegetarians and meat-eaters. *The Psychological Record, 60*, 287-306.

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

Bond, F.W., Hayes, S.C., Baer, R.A., Carpenter, K.M., Orcutt, H.K., Waltz, T., & Zettle, R.D. (under review). Preliminary psychometric properties of the Acceptance and Action Questionnaire – II: A revised measure of psychological flexibility and acceptance.

Cohen, A. S., Beck, M. R., Brown, L. A., & Najolia, G. M. (2010). Decoupling implicit measures of pleasant and unpleasant social attitudes. *Journal of Behavior Therapy and Experimental Psychiatry, 41*, 24-30.

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

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. (2002). The Implicit Association Test as a tool for studying dysfunctional associations in psychopathology: Strengths and limitations. *Journal of Behavior Therapy and Experimental Psychiatry, 33*, 115-133

Feldner, M. T., Zvolensky, M. J., Eifert, G. H., & Spira, A. P. (2003). Emotional avoidance: an experimental test of individual differences and response suppression using biological challenge. *Behaviour Research and Therapy, 41*(4), 403-411

Gavin, A., Roche, B., & Ruiz, M. R. (2008). Competing contingencies over derived relational responding: A behavioral model of the implicit association test. *The Psychological Record, 58*, 427-441.

Greenwald, A.G., & Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

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

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

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 Academic/Plenum.

Hayes, S. C., Strosahl, K. D., Wilson, K. G., Bissett, R. T., Pistorello, J., Toarmino, D., … McCurry, S. M. (2004). Measuring experiential avoidance: A preliminary test of a working model. *The Psychological Record, 54*, 553-578.

Hayes, S. C., Wilson, K. G., Gifford, E. V., Follette, V. M., & Strosahl, K. (1996). Experiential avoidance and behavioral disorders: A functional dimensional approach to diagnosis and treatment. *Journal of Consulting and Clinical Psychology, 64*, 1152-1168.

Hayes, S. C., Wilson, K. G., Gifford, E. V., Bissett, R., Piasecki, M., Batten, S. V., . . . Gregg, J. (2004). A randomized controlled trial of twelve-step facilitation and acceptance and commitment therapy with polysubstance abusing methadone maintained opiate addicts. *Behavior Therapy, 35*, 667-688.

Jackson, M. (2008). *Examining the motivational effects of verbal stimuli: An application of motivational augmentals*. Unpublished dissertation, University of Nevada Reno, Reno, NV.

Karpinski, A.,& Steinman, R.B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*, 16–32.

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,* 253-268.

Nosek, B.A., & Banaji, M.R. (2001). The Go/No-go Association Task. *Social Cognition, 19*, 625-666.

Nosek, B.A., Greenwald, A.G., & Banaji, M.R. (2006). The Implicit Association Test at age 7: A methodological and conceptual review. In J.A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). New York: Psychology Press.

Osgood, C., Suci, G., & Tannenbaum, P (1957). *Measurement of meaning*. Urbana, IL: University of Illinois Press

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. *The Psychological Record, 59*, 621-640.

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*, 416-425.

Scanlon, G., Barnes-Holmes, Y., Barnes-Holmes, D., & Stewart, I. (under review). Implicit and explicit self-esteem in children with ADHD.

Toglia, M., & Battig, W. (1978). *Handbook of semantic word norms*. Somerset, NJ: Lawrence Erlbaum Associates.

Vahey, N. A., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). A first test of the Implicit Relational Assessment Procedure (IRAP) as a measure of self-esteem: Irish prisoner groups and university students. *The Psychological Record, 59*, 371-388.

Vilardaga, R., Hayes, S.C., Luoma, J.B., Pistorello, J., Levin, M.E., Hildebrandt, M., … Bond, F. (under review). Mindfulness and values processes predict burnout among addiction counselors above and beyond worksite factors.

Waltz, T.J., Drossel, C, Hayes, S.C., Roget, N., & Fisher, G. (in preparation). Implicit attitudes and words related to substance abuse and treatment

Whelan, R. (2008). Effective analysis of reaction time data. *Psychological Record, 58*(3), 475-582.

Zettle, R. D., Hocker, T. R., Mick, K. A., Scofield, B. E., Petersen, C. L., Hyunsung S., & Sudarijanto, R. P. (2005). Differential strategies in coping with pain as a function of level of experiential avoidance. *The Psychological Record*, *55(4)*, 511-524.

Zettle, R.D., Petersen, C.L., Hocker, T.R., & Provines, J.L. (2007). Responding to a challenging perceptual-motor task as a function of level of experiential avoidance. *The Psychology Record*, *57*, 49-62.

Footnotes

As part of the editorial review it was suggested that we speak of IRAP effects using the MT-IRAP as "MT-IRAP effects." We have done so for clarity, but the MT-IRAP is merely a form of the IRAP and thus we do not mean to imply by that term that the MT-IRAP is measuring a different effect or phenomenon.

Author Contact Information:

Michael Levin
Department of Psychology / MS 298
University of Nevada, Reno
Reno, NV 89557-0062
(541) 531-3892
Email: levinm2@gmail.com
