Measuring Implicit Attitudes: A Positive Framing Bias Flaw in the Implicit Relational Assessment Procedure (IRAP)

Brian O'Shea

Derrick G. Watson

Gordon D. A. Brown

University of Warwick, UK

Abstract

How can implicit attitudes best be measured? The Implicit Relational Assessment Procedure (IRAP), unlike the Implicit Association Test (IAT), claims to measure absolute, not just relative, implicit attitudes. In the IRAP, participants make congruent (Fat Person-Active: *False*; Fat Person-Unhealthy: *True)* or incongruent (Fat Person-Active: *True*; Fat Person-Unhealthy: *False*) responses in different blocks of trials. IRAP experiments have reported positive or neutral implicit attitudes (e.g., neutral attitudes towards fat people) in cases where negative attitudes are normally found on explicit or other implicit measures. It was hypothesized that these results might reflect a Positive Framing Bias (PFB) that occurs when participants complete the IRAP. Implicit attitudes towards categories with varying prior associations (nonwords, social systems, flowers and insects, thin and fat people) were measured. Three conditions (standard, positive framing, and negative framing) were used to measure whether framing influenced estimates of implicit attitudes. It was found that IRAP scores were influenced by how the task was framed to the participants, that the framing effect was modulated by the strength of prior stimulus associations and that a default PFB led to an overestimation of positive implicit attitudes when measured by the IRAP. Overall, the findings question the validity of the IRAP as a tool for the measurement of absolute implicit attitudes. A new tool (Simple Implicit Procedure: SIP) for measuring absolute, not just relative, implicit attitudes is proposed.

Key Words: Implicit attitudes; Implicit Relational Assessment Procedure; Positive Framing Bias; Simple Implicit Procedure.

Measuring Implicit Attitudes: A Positive Framing Bias Flaw in the Implicit Relational Assessment Procedure (IRAP)

Implicit attitudes are automatic evaluations that occur outside conscious awareness and are measured without requiring respondents to introspect on their feelings. Explicit attitudes in contrast are the result of deliberate introspection and controlled evaluative judgment (Greenwald & Banaji, 1995). One reason for measuring implicit attitudes is that participants may use self-presentation tactics or respond in a socially desirable manner on explicit self-reports to avoid being perceived as prejudiced. Implicit measures can also be useful in areas where participants might be unwilling to reveal personal psychological attributes or are unaware of these psychological attributes (for a review of implicit attitudes and the tools used to measure them see Gawronski & De Houwer, 2014).

The current gold standard method for assessing implicit attitudes is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and this tool has been increasingly used in clinically relevant areas (see https://implicit.harvard.edu/implicit/pimh/). The IAT has shown promise in predicting self-harm (Randall, Rowe, Dong, Nock, & Colman 2013), social anxiety disorders (Teachman & Allen, 2007) and suicidal ideation (Harrison, Stritzke, Fay, Ellison, & Hudaib, 2014). The current study questions the validity of a recently developed implicit measure, the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Barnes-Holmes, Stewart, & Boles, 2010a), which has also been used with vulnerable populations.

Absolute vs. Relative Measures of Implicit Cognition

In a typical IAT (e.g., the Fat-Thin IAT) participants are successively presented with various pictures of thin and fat individuals and positive and negative words. In one of the two critical blocks, participants have to press the E key on a computer keyboard if a positive word or a picture of a thin person appears and press the I key if a negative word or a picture of a fat

person appears (congruent task). In the other critical block, participants have to press E if a positive word or a picture of a fat person is shown and they must press the I key for a negative word or a picture of a thin person (incongruent task). The basic idea underlying the IAT is that participants will make faster and more accurate responses when those responses are congruent with their current beliefs than when they are not.

Researchers using the IAT typically find a pro-thin/anti-fat bias (e.g., O'Brien, Hunter, & Banks, 2006). Importantly, the IAT can only measure relative attitudes (e.g., attitudes to fat people relative to attitudes to thin people). It is therefore impossible to determine using the IAT whether this bias is the result of a strong/weak pro-thin bias, a strong/weak anti-fat bias or some combination of the two (see Blanton & Jaccard, 2006; Roddy, Stewart, & Barnes-Holmes, 2011; 2012). Likewise, the IAT cannot be used to determine how interventions that aim to increase or reduce implicit biases have their effect (e.g., a difference in implicit attitudes could be reduced by acting on the 'Thin Person' category, the 'Fat Person' category, or both; see Lai et al., in press).

Another limitation of the IAT arises because some categories do not have an obvious comparison group. For example, when assessing implicit self-esteem, IAT researchers can measure the positive and negative associations a person has with the self in comparison to a specified/unspecified other (or with "me" in comparison to "not me"). The type of comparison category (i.e., specified vs. unspecified other) used affects implicit self-esteem results (Karpinski, 2004). Therefore, a more appropriate approach would measure only evaluative associations with the self without the need to use a complementary category.

To address these problems, reaction time tools which attempt to measure absolute attitudes and do not require a relative comparison to another group have been developed. For example, the Go/No-Go (Nosek & Banaji, 2001) and the Extrinsic Affective Simon Task (De Houwer, & De Bruycker, 2007) both claim to measure implicit attitudes non-

relatively/absolutely. However, both suffer from a number of problems ranging from a high level of task difficulty to low reliability (Bar-Anan & Nosek, 2013; De Houwer & De Bruycker, 2007). Variations of the IAT that could be described as single concept IATs, such as the Single Target/Category IAT (e.g., Bluemke & Friese, 2008; Karpinski & Steinman, 2006) have shown promise for measuring absolute implicit attitudes. As these names suggest, implicit attitudes towards only one concept can be measured during each test. As an example, implicit attitudes towards fat people would be measured by having participants press the I key when a picture of a fat person or a negative word appears (congruent task) and the E Key for anything else. On the next block of trials participants would press the E key for pictures of fat people or positive words (incongruent task), and the I key for everything else. The absolute results are obtained by measuring the mean latency difference between the congruent and incongruent block of trials. As in the standard IAT[1] participants are expected to make faster and more accurate responses on congruent tasks.

The IRAP is a recently developed alternative to the IAT that claims to measure attitudes in both absolute and relative terms. Like the IAT, the IRAP is based on latencies of participants' accurate and speeded responses to stimuli. However, rather than categorizing items with the appropriate key press (as in the IAT), participants instead have to press keys that correspond with *True* or *False* displayed on the computer screen. "Correct" responses are based on instructions given to participants before each block of trials begins. For example, participants might be presented with the picture of a 'Fat Person' and the adjective 'Active'. In one condition participants are instructed to respond with a *True* key press (an incongruent

[1] The Brief IAT (Sriram & Greenwald, 2009) is similar in methodology to the single concept IATs but, like the standard IAT, it still only measures implicit attitudes relatively.

response) and in another condition they are instructed to respond with a *False* key press (a congruent response).

The mean latency difference between pressing *True* and pressing *False* across the congruent and incongruent response conditions is claimed to reflect a participant's absolute positive implicit attitude towards 'Fat Person'. Similarly, an evaluation of absolute negative implicit attitudes can be obtained by measuring the latency difference to respond *True* to the association of 'Fat Person' with negative words (e.g., 'Fat Person'-'Unhealthy'), except that the *True* key is now a 'congruent response' and the *False* key an 'incongruent response' (see Figure 1)[2]. When the results from responding *True* or *False* to 'Fat Person' with positive words and 'Fat Person' with negative words are combined/averaged, the result reveals if the absolute attitude towards 'Fat Person' is positive, negative or neutral[3].

This kind of absolute attitude measurement cannot be achieved using the standard IAT, because the latency with which a person responds to positive words (e.g., 'Active') and pictures of fat people in the 'Fat Person-Positive' condition is uninterpretable in isolation. The latency only becomes useful if it can be compared with the latency from the condition in

---

[2] Separately measuring both a positive and a negative attitude towards a category can provide a more nuanced understanding of the attitude under investigation. For example a person can have both a strong positive attitude towards giving blood (i.e., like helping people) but also have a strong negative attitude towards giving blood (i.e., fear of needles). Traditional bipolar attitude measures (e.g., like-dislike) treat positive and negative evaluative processes as reciprocally activated, equivalent and interchangeable which is not always the case (see Cacioppo, Gardner, Berntson, 1997).

[3] The IRAP normally includes a comparison group to allow researchers to carry out relative comparisons between groups in order to compare results with those obtained from the IAT.

which negative words (e.g., ‘Unhealthy’) and pictures of thin people require the same response (i.e., the ‘Thin Person-Negtive’ condition).

The IRAP has been used in at least 20 empirical studies to date (Golijani-Moghaddam, Hart, & Dawson, 2013). These studies have examined a number of forensic/clinically relevant samples (for a review see Vahey, Nicholson, & Barnes-Holmes, 2015). Because the IRAP is being used with such vulnerable populations, it is important that the validity of this tool is tested thoroughly to ensure that the method is accurately measuring what it purports to. This is of particular importance because a number of IRAP publications have reported unexpected or contradictory results that cannot be accounted for easily.

For example, Roddy et al. (2011, 2012; see also Nolan, Murphy, & Barnes-Holmes 2013) reported that what was driving the pro-thin/anti-fat bias in the IAT was an extremely strong pro-thin bias with participants having a neutral implicit attitude towards fat people. The concept of culture-based socialization that glorifies a slender figure was used to account for these results. However, Bessenoff and Sherman (2000), using a non-relative/absolute lexical decision making task, obtained completely opposite results to those of Roddy et al. (2011, 2012; i.e., they found an anti-fat bias and a neutral/unbiased attitude towards thin people).

In another example, Barnes-Holmes, Murphy, Barnes-Holmes, and Stewart (2010b) found that participants associated pictures of a black person with a gun and pictures of a white person with a gun as ‘Safe’ (i.e., they were faster to press *True* rather than *False* to the joint presentation of a picture of a person with a gun and ‘Safe’). This was not what the authors had predicted based on prior evidence that a civilian holding a gun in a neutral context would normally be considered as dangerous rather than safe (i.e., participants should

press *False* quicker than *True*)[4]. Other counterintuitive results include findings from unpublished studies that a normative sample of participants had an unexpected pro-death implicit bias (Hussey & Barnes-Holmes, 2012), and one where religious participants and atheists had unpredicted positive attitudes towards both their in-group and out-group (O'Shea & Stewart, 2015)[5]. Importantly, when the data were compared relatively (like in the IAT) the typical pro in-group/anti out-group biases were obtained. In the current study we show that these results can be accounted for on the assumption that participants are adopting a simple default heuristic (a Positive Framing Bias; PFB) when performing the IRAP.

Framing effects and language biases

The power of positive and negative framing has been well documented, particularly in the area of risky decision making (for reviews, see Kuhberger, 1998). One of the best known examples of the framing effect is Tversky and Kahneman's (1981) 'Asian disease problem'. When presented with choices, if logically equivalent outcomes are phrased in terms of the *number of lives saved*, people will generally select the certain/safe option. In contrast, if outcomes are phrased in terms of the *number of lives lost*, people preferentially select the risky option. Building on this evidence McKenzie and Nelson (2003) found that participants had a bias towards describing a new treatment in terms of the percentage that survived rather

---

[4] It could be argued that people are generally exposed to guns in a safe environment (i.e., guns being carried by law enforcers), resulting in the pro-safe attitude. However, we think this is unlikely because the experiment used pictures of civilians wearing a white t-shirt and this is likely to evoke a fight or flight response especially for the black individuals (e.g., Correll, Park, Judd, & Wittenbrink, 2002).

[5] Atheists are generally distrusted in societies with a religious majority (e.g., Gervais, Shariff, & Norenzayan, 2011).

than died (i.e., a positively biased outcome description). Furthermore, participants were also biased towards describing a glass as half full rather than half empty (i.e., an optimism bias; e.g., Peterson, 2000). Importantly, it was possible to manipulate these positivity biases by describing the reference points of the treatment outcomes differently or stating where the liquid in the glass had been previously. For example, people were more likely to say the glass was half empty if it had previously been full.

Language both reflects and shapes our representations of the world (Boroditsky, 2011) and language can therefore provide important insights into thought processes and biases. English speakers appear to have a general tendency to describe changes in events or objects as increasing rather than decreasing. To clarify, if a person's weight has decreased, we might describe this as 'the person is thinner' (thinness has increased), not 'the person is less fat'. In contrast, a person who has gained weight might be referred to as 'fatter', but not as 'less thin'. Additionally there is no morpheme in English that is equivalent to the suffix 'er' to indicate that a dimension has decreased (McKenzie & Nelson, 2003). This implies that describing things in positive (increasing) rather than negative (decreasing) ways is a bias inherent in the English language and therefore our thought processes.

Another strand of psycholinguistic research that emphasizes the importance of positivity is based on markedness (see Haspelmath, 2006 for a critical review) where 'unmarked' is the classification for words used most frequently and which also have the most neutral meaning (Leech, 2006). 'Positive adjectives like *good* and *long* are stored in memory in less complex forms than…their opposites' (Clark, 1969, p.398) and therefore, are described as 'unmarked'. A speaker asking in a restaurant 'How good is the food?' is simply asking for an evaluation of the food and will be satisfied with a positive or negative appraisal. However, when asking 'How bad is the food?' essentially he/she is inquiring about the extent of the food's badness. In this context, "since *good* can be neutralized and *bad* cannot, *good* is

said to be ‘unmarked’ and *bad* ‘marked’” (Clark, 1969, p.398). Likewise describing a board as six meters long is acceptable to English speakers but describing it as six meters short is not. Therefore, short would be described as ‘marked’. Finally, a morphologically negative word is ‘marked’ as opposed to a positive one (e.g., honest vs. dishonest; happy vs. unhappy) and therefore, humans are less efficient (in terms of processing speed) at handling negative statements (e.g. Sherman, 1973).

Related studies (e.g., Mathews & Dylman, 2014) have shown that English speakers have a preference to use ‘larger’ (e.g., more, taller, higher) comparisons to describe the relationship between two magnitudes (e.g., one flag is _____ than the other). Other evidence shows that people make a concerted effort to dampen down, mute, and even erase negative experiences and that positive illusions promote psychological wellbeing (Taylor, & Brown, 1994). Peoples’ strong positivity biases have been described as the Pollyanna Principle (for review see Matlin, 2004). Overall, the evidence points towards humans having a bias towards positivity and larger or increasing descriptions (i.e., a Positive Framing Bias; PFB) and suggests that this bias can be manipulated by framing effects.

Present Study

The present study tests the hypothesis that the absolute estimates of implicit attitudes obtained from the IRAP are influenced by a PFB. Specifically, people might find it easier to response *True* to positive descriptions of stimuli than to press *False*. This effect is predicted to occur over and above any effect of congruence between the stimulus and the description that is presented with it, and is expected to lead to an overestimation of the positivity of absolute implicit attitudes. For example participants may be faster to respond *True* (rather than *False*) when responding whether a stimulus category (e.g., Thin Person/Flowers) is positive (e.g., “Good”) than when it is negative (e.g., “Bad”). If a PFB does influence IRAP

responding, the IRAP may not be able to measure absolute implicit attitudes in the way intended.

Three main hypotheses were therefore tested:

H1: By default, participants will be more likely to focus on positive rather than negatively framed associations and therefore they will be faster to respond *True* than *False* for categories presented with positive words in the standard IRAP condition.

H2: The way the task is framed will influence estimates of a person's absolute implicit attitudes. If absolute attitudes as measured by the IRAP are susceptible to PFB effects, directly manipulating how the task is framed to participants should influence the results in predictable directions. Specifically, encouraging participants to focus on whether positive associations are true or false should increase estimated 'implicit positive attitudes' towards any category. Conversely, encouraging participants to focus on whether negative associations are true or false should increase estimates of 'implicit negative attitudes'. The effect of framing is predicted to be modulated by the strength of pre-existing negative or positive associations. That is, robust positive or negative associations will be less likely to be influenced by framing effects. Tasks that involve weak or absent prior associations will be more strongly influenced by the framing manipulation than those with strong prior associations. To test this we ran four different IRAP tasks, using stimuli which had differing strengths of prior positive and negative associations.

H3: When the IRAP data are analyzed in such a way as to obtain relative attitudes (as is done for the IAT), the framing effect will have no influence. This is because any systematic biases will be cancelled out when the relevant conditions are combined. Tasks that use stimuli with weak or absent prior associations will show neutral attitudes; those with strong prior associations will in contrast show the expected pro/anti-bias (e.g., pro-thin/ anti-fat).

To test these hypotheses we manipulated the way in which tasks were framed (no frame-standard, positive frame, and negative framing conditions) for each of four IRAP tasks, which contained stimuli that varied in their strength of prior associations. If a PFB does exist (i.e., if the first two hypotheses were confirmed), the use of the IRAP as a measure of absolute, rather than just relative attitudes would be severely limited.

Method

Participants

The final sample consisted of 60 students from the University of Warwick (mean age = 21.9, *SD* = 2.79), 20 in each condition (standard, positive frame and negative frame). Fourteen participants were replaced (four from the standard, six from the positive frame and four from the negative frame conditions)[6] from the original 60 because the required performance criteria were not met. The final sample contained 32 females and comprised 36 Asians, 19 Whites, 3 Mixed Race and 2 Blacks. Participants were recruited via an electronic recruitment system and were paid £4. Participants were tested individually in a small room and the experiment took approximately 40 minutes to complete.

Apparatus and Materials

Implicit Relational Assessment Procedure: (IRAP; Barnes-Holmes et al., 2010a). Each participant completed four separate IRAPs. These were a Nonword IRAP, a Social System IRAP, a Nature IRAP, and a Weight IRAP. These four different IRAPs were chosen because the category stimuli in each were expected to have varying degrees of prior associations. Both the Nature and Weight IRAP stimuli were predicted to have strong prior associations (see Greenwald et al., 1998 and O'Brien, et al., 2006). The Social System IRAP

[6] Including the original participants did not cause any significant changes to the results reported.

stimuli were expected to have weak associations[7] and the Nonword IRAP stimuli should have no prior associations. All IRAP tests were administered using an Intel Windows 7 laptop with a 15" LCD screen (IRAP software available from http://irapresearch.org/wp/downloads-and-training/).

For the Weight IRAP, on each trial one of two category labels was presented at the top of the screen ('Thin Person' or 'Fat Person') and a single positive or negative target stimulus was presented in the center of the screen (e.g., 'Healthy' or 'Ugly'). The Weight IRAP used pictures, and the other three IRAPs used words as the category labels. Two response options ('*True*' and '*False*') also appeared at the bottom left and right (respectively) of the screen (see Figure 1). The remaining IRAP tasks followed the same format but with the condition-specific stimuli presented (see Tables 1 and 2 in the supplementary materials for the full stimulus sets).

All category labels and target stimuli were matched on word length, phoneme length and word frequency using Brysbaert and New's (2009) film subtitle database. The positive and negative target stimuli for the Social System IRAP were determined by asking 50 participants (via Amazon Mechanical Turk) to generate positive and negative words that they associated strongly with Capitalism and Socialism. The most frequently reported words were selected for use in the current study. Stimuli for the nature IRAP were chosen from past implicit research (e.g., Greenwald et al., 1998). The 12 images for the weight IRAP were taken from Nolan et al. (2013) and consisted of three pictures of women and three pictures of men, before and after they had lost a significant amount of weight. These images were

---

[7] O'Shea, (2015a) found that 17 UK and 10 Chinese undergraduates had no relative comparison bias in favor or against Capitalism vs. Communism or Socialism. However, see O'Shea, (2015b) where bankers had extremely strong biases in favor of Capitalism.

controlled on a number of dimensions (e.g., picture angle, cropping, clothing and background). The most appropriate positive and negative target words were chosen from studies of implicit weight biases (Roddy et al., 2010, 2011).

Design and Procedure

The study used a mixed 4 (trial type: category 1-positive words, category 1-negative words, category 2-positive words, and category 2-negative words) X 4 (IRAP task: Nonword, Social System, Nature and Weight) X 3 (framing condition: standard, positive, negative) design, with IRAP task and IRAP trial type as within-subjects factors and framing condition as a between-subjects factor.

We illustrate the procedure using the Weight IRAP task. Participants were required to respond in a predefined way as specified before each block of trials. For example, participants might be instructed 'On this block please respond as if Thin Person is Positive and Fat Person is Negative'. This would be described as a pro-Thin block of trials (congruent responding[8]) and participants were required to respond *True* to the stimulus combination 'Thin Person – Positive Word' and 'Fat Person – Negative Word', and to respond *False* to 'Thin Person – Negative Word' and 'Fat Person – Positive Word'. After participants completed this pro-Thin block they were required to respond in the opposite manner on the next block and were thus instructed: 'On this block please respond as if Thin Person is Negative and Fat Person is Positive' (incongruent responding). On these pro-Fat trials participants had to respond *False* to 'Thin Person – Positive Word' and 'Fat Person –

---

[8] For the Non-Word IRAP and the Social System IRAP the blocks that are referred to as congruent or incongruent responding were arbitrarily selected. However, for the Nature and Weight IRAP congruent responding was defined in terms of the blocks people were expected to find easier due to prior associations in memory.

Negative Word' and to respond *True* to 'Thin Person – Negative Word' and 'Fat Person – Positive Word'.

This procedure enabled the experimenter to measure attitudes based on four separate trial types (Thin Person – Positive Word, Thin Person – Negative Word, Fat Person – Positive Word and Fat Person – Negative Word) by measuring the mean latency difference in responding *True* vs. *False* in pro-Thin and pro-Fat blocks. To respond, participants pressed either 'D' or 'K' on the keyboard. The D and K keys corresponded to the left – right positions of the response options on the screen respectively. The locations of the two response options interchanged quasi-randomly from left to right among trials, with the constraint that they could not remain in the same position three times in succession.

To complete the IRAP sequence, participants completed six test blocks which alternated between requiring pro-Thin or pro-Fat responses across blocks. Initial block (pro-Thin or pro-Fat) was counterbalanced across participants. Each block consisted of 24 trials, made up of the 6 positive and 6 negative target words presented twice in the presence of the two category stimuli (i.e., 'Thin Person' or 'Fat Person'). Trials were presented quasi-randomly, such that the same trial type could not be repeated across two successive trials. If participants pressed the correct response on a trial the screen cleared for 400 ms before the next trial was presented. If an incorrect response was given a red letter (X) appeared on screen and remained until the participant pressed the correct response key. After participants completed a block of trials their mean accuracy and median response latency scores were displayed on the screen. The first participant was assigned to the standard condition, the second to the positive framing condition and the third to the negative framing condition. This order was then repeated and maintained for the remaining participants.

Standard framing condition: For the standard condition visual instructions before each block were alternated across participants. For example, one participant would be shown,

'On this block please respond as if **Thin Person** is Positive and **Fat Person** is Negative' and 'On this block please respond as if **Thin Person** is Negative and **Fat Person** is Positive' and the next participant would always begin with **Fat Person** as the first category described in each sentence.

Positive framing condition*:* In the positive framing condition visual explanations of how to perform the task were always framed in a positive way prior to beginning each block of trials (e.g., 'On this block please respond as if **Thin Person** is Positive and **Fat Person** is Negative' or 'On this block please respond as if **Fat Person** is Positive and **Thin Person** is Negative')[9]. These instructions were dependent on participants completing a pro-Thin or pro-Fat block. Unlike in the standard framing condition, each participant in this condition was also presented with a standardized verbal explanation from the experimenter.

The script was as follows:

"A method or strategy that will help you complete this task is to keep **'Thin Person'** and 'Positive word' in your mind and base all other responses off that[10]. For example when a **'Thin Person'** and a Positive word appears, press *True* and if this does not occur press *False* (such as **'Thin Person'** and Negative word). Then use this strategy to gauge how to respond to the other category by responding in the opposite manner (such as **'Fat Person'** Positive words *False*; **'Fat Person'** Negative word *True*) To emphasize **'Thin Person'** is Positive, **'Thin Person'** is Positive"

---

[9] Since English speakers process sentences from left to right the leftmost instructions were likely be processed first.

[10] Category words in bold were replaced depending on which pro-block was presented first and which of the four versions of the IRAPs was being conducted. Sentences in brackets were only explained to participants who were confused and required more detailed instructions.

After participants completed the first practice block they were told the following:

"Now on this block put **'Fat Person'** is Positive into your mind **'Fat Person'** is Positive, '**Fat Person'** is Positive"

Participants were told the following on each successive IRAP:

"For this task use the same strategy as the last time by keeping **'XXX'** and Positive word in your mind."

Negative framing condition: The negative framing condition was identical to the positive except that the underlined words in the verbal script above were replaced with their antonyms and the visual description before each block always described the negative instructions first (e.g., 'On this block please respond as if **Fat Person** is Negative and **Thin Person** is Positive' or 'On this block please respond as if **Thin Person** is Negative and **Fat Person** is Positive').

The remaining three IRAP tasks (Nonword, Social System and Nature) followed the same procedure but with the condition-relevant stimuli substituted. The order of the four IRAPs was randomized within the standard framing condition with the same set of random orders used for the positive and negative framing conditions. For the first IRAP participants were required to complete a minimum of two practice blocks. This was to ensure that participants were accustomed to the IRAP's procedure. If a participant received first pro-Fat then pro-Thin conditions on the practice blocks this sequence was maintained for the 1st and 2nd test block, the 3rd and 4th, and the 5th and 6th test block.

To proceed to the test blocks participants had to achieve an accuracy of 79% or above and a median response latency of less than 2200 ms on two consecutive practice blocks. All participants met the practice block criteria but if these criteria were not maintained throughout each test block on the four IRAP tasks, participants' data were removed and replaced with data from new participants. This resulted in 14 participants being

replaced. When the four IRAPs were completed, participants provided demographic information, received their payment and were thanked and debriefed.

Results

The primary data obtained from the IRAP tasks are raw latency scores defined as the time in milliseconds that elapsed between the onset of the stimulus and the correct response being made by the participant. The DV was participants' mean *False* minus *True* reaction time difference for each of the four trial types (i.e. Thin Person-Positive Words, Thin Person-Negative Words, Fat Person-Positive Words, Fat Person-Negative Words) across the congruent (e.g., pro-thin) vs. incongruent (e.g., pro-fat) blocks. Analysing the trial types/categories separately provided the absolute results while averaging all the trial types provided the relative results. Following standard procedures to control for individual variation, (Barnes-Holmes et al., 2010a) each participant's response latencies were transformed using an adaption of the Greenwald, Nosek and Banaji (2003) D-algorithm. The steps involved in calculating both the absolute and relative D-IRAP scores can be found in the supplementary materials. Initially a stimulus Category X Word Valence X IRAP task within-subjects ANOVA was performed on the *False* minus *True* absolute D-IRAP scores for each of the four trial types in the standard condition. This analysis allowed us to test the first hypothesis: H1: participants will have a faster *True (*vs. *False)* response for categories presented with positive words. Inclusion of the factors stimulus Category and IRAP task also allowed us to consider the generality of the effect of Word Valence.

Following this, a 4 (trial type: category 1-positive words, category 1-negative words, category 2-positive words, category 2-negative words) X 4 (IRAP task: Nonword, Social System, Nature, Weight) X 3 (framing: standard, positive, negative) mixed ANOVA was

performed on the of absolute D-IRAP (estimate) scores[11]. This analysis was carried out to test H2 (how the task is framed will influence the estimates of a person's implicit attitudes). A further analysis combined both the positive word and negative word D-IRAP (estimate) scores for each of the eight categories (i.e., Fat Person, Thin Person, Flower, Insect etc.) in the standard condition. This method provides the average absolute D-IRAP (estimate) scores for each category and matches the standard way of calculating the IRAP's results. This analysis was used to test further the argument that the PFB has an influence on peoples' responses and to examine the extent to which this can lead to potentially inflated estimates of implicit attitudes[12]. Finally, tests were carried out to determine whether framing influenced participants' *relative* implicit attitudes (H3). In all analyses, whenever Mauchly's sphericity assumption was violated the Greenhouse-Geisser correction was applied.[13]

Test of H1: The influence of the PFB on prior associations in the standard condition

---

[11] The absolute D-IRAP (estimate) scores were found by reversing the *False* minus *True* D-IRAP scores for categories presented with negative words. Categories presented with positive words were not reversed. This was carried out so that scores above zero represent a positive attitude and those that are below the zero mark represent a negative attitude. The absolute D-IRAP (estimate) scores were used to test H2.

[12] Separate analysis of each of the four IRAPs (Nonword, Social System, Nature, and Weight) can be found in the supplementary materials.

[13] Preliminary analyses revealed the sentence sequence instructions (e.g., **Thin Person** described first or **Fat Person** described first) in the standard condition and the order of the IRAP blocks presented (e.g. beginning with either a pro-**Thin** or a pro-**Fat** block**)** did not influence the critical IRAP effect in subsequent analyses.

False minus *True* D-IRAP scores for the standard framing condition were analyzed using a 2 (Category) X 2 (Word Valence) X 4 (IRAP task) within-subjects ANOVA. This revealed a significant main effect of Valence: $F(1,19) = 67.05$, $p <. 001$, $\eta\rho^2 = .78$, as predicted by H1. There was also a significant Category x Valence two-way interaction: $F(1,19) = 12.26$, $p <. 005$, $\eta\rho^2 = .392$, and a significant IRAP x Category x Valence three-way interaction: $F(3,57) = 5.77$, $p <. 005$, $\eta\rho^2 = .233$. No other main effects or their interactions were significant (all Fs < 1.57, ps > .21). As shown in Figure 2, *True* responses were faster than *False* responses on positive word trials while for negative word trials there were no differences between pressing *True/False* supporting H1. The significant three-way interaction arises because *True* responses were faster than *False* responses for Category 1 stimuli across all four IRAPs. In contrast, for Category 2 stimuli True responses were only faster than false responses for the Nonword and Social System IRAPs (i.e., those in which the stimuli had relatively weak prior associations).

This pattern of results is assumed to reflect differences in the strength of prior associations across the four IRAP types. For the two IRAPs with the weakest prior associations (Nonword and Social System) there was little influence of whether the description was congruent or incongruent with the stimulus being presented. Thus for those IRAPs there was little difference between scores for Category 1 versus Category 2 stimuli but there was an overall effect of Word Valence. That is, *True* vs. *False* responses were faster on positive word trials. In contrast, for the IRAPs with stronger prior associations (Nature and Weight), there was an effect of stimulus congruence. That is, congruent responses (e.g., positive word – flower, negative word - insect) were faster than incongruent responses (positive word – insect, negative word – flower). As for the Nonword and Social System IRAPs, the effect of the PFB was to increase the speed of *True* responses on positive word trials. This resulted in an increased difference between positive word and negative word trials

for positive stimuli (e.g., Flowers/Thin) and a corresponding smaller difference for negative stimuli (Insects/Fat).

Test of H2: The influence of the PFB on estimates of absolute implicit attitudes

The absolute D-IRAP (estimate) scores were analyzed with a 4 (trial type: category 1-positive words, category 1-negative words, category 2-positive words, category 2-negative words) X 4 (IRAP task: Nonword, Social System, Nature, Weight) X 3 (framing: standard, positive, negative) mixed ANOVA, with trial type and IRAP task as the within-subjects factors and framing condition as the between-subjects factor. This analysis revealed a significant main effect of trial type, $F(3, 171) = 64.52$, $p < .001$, $\eta\rho^2 = .53$, but not of IRAP task, $F(2.62, 149.34) = .40$, $p = .76$, $\eta\rho^2 = .01$. Importantly, the trial type X IRAP task interaction was significant, $F(7.02, 399.99) = 10.26$, $p < .001$, $\eta\rho^2 = .15$. As shown in Figure 3 (top left), for stimuli with absent or weak priori associations (i.e., the Nonword IRAP and the Social System IRAP), elevated implicit attitudes were found (see Cat1-Positive and Cat2-Positive). For stimuli with stronger prior associations (i.e., the Nature & the Weight IRAPs) this pattern was weaker.

A significant main effect was found for framing, $F(2, 57) = 42.40$, $p < .001$, $\eta\rho^2 = .60$. As shown in Figure 3 (top right) the positive framing condition produced elevated estimates of implicit attitudes ($M = .31$) whilst negative framing resulted in reduced estimates of implicit attitudes ($M = -.09$). The standard framing condition showed estimates of implicit attitudes that were in between the estimates obtained in the positive and negative framing conditions ($M = .16$). Across the four trial types the framing manipulation had a similar effect — a framing X trial type interaction was not found, $F(6,171) = 1.02$, $p = .41$, $\eta\rho^2 = .04$. Crucially, the framing X IRAP task interaction was significant $F(6,171) = 9.50$, $p < .001$, $\eta\rho^2 = .25$. Figure 3 (bottom left) shows that framing influenced the estimates of participants' implicit attitudes and that this influence was weaker for strong prior associations. Lastly the

three way interaction was not significant $F(18, 504) = .62$, $p = .89$, $\eta\rho^2 = .02$. The overall pattern of results suggest that: i) participants had positive implicit attitudes (estimates) across all the IRAP tasks, ii) framing had a strong effect on the estimates of implicit attitudes, and iii) the effect of framing was stronger for associations with weaker rather than stronger prior associations.

Test of H2: Combining the positive and negative trials for each category in the standard condition

In order to match the standard way the IRAP results are reported in the literature we combined/averaged the absolute D-IRAP (estimate) scores for the positive and negative word associations in each category (see Figure 3; bottom right). For example, 'implicit attitudes' for 'Thin Person' were calculated by averaging the D-IRAP (estimate) scores from the 'Thin Person-Positive Words' and 'Thin Person-Negative Words' trials. One sample t-tests were conducted on the resulting values which revealed that the scores for the Nonword and Social System trials (absent/ weak prior associations) were significantly greater than (ts > 2.95, ps < .01) or marginally greater (for Capitalism; $t = 1.90$, $p = .07$) than zero. In the Nature IRAP there was an expected pro-flower bias ($t = 7.24$, $p < .001$). The apparent neutral attitude towards insects ($t = -1.07$, $p = .30$) is likely due to the tendency for participants to frame the IRAP task in a positive way, with this PFB then offsetting the negative bias people usually possess for insects. This is similar to the Weight IRAP in which a strong positive attitude was observed for thin people ($t = 4.43$, $p = <.001$) but a neutral attitude was found for fat people ($t = -1.03$, $p = .32$). These findings further underline the problem of using the IRAP's absolute results because they are likely to overestimate the positivity of a person's implicit attitudes.

Test of H3: The influence of the PFB on the relative IRAP results

The relative overall D-IRAP scores were analyzed using a similar method to that used to analyze IAT data to assess if there were differences across the standard, positive and

negative framing conditions for each of the four IRAP tasks. For all the IRAP tasks there were no significant differences across the three framing conditions (Nonword = $F(2, 57) = .35$, $p = .70$, $\eta\rho^2 = .01$; Social System = $F(2, 57) = .19$, $p = .83$, $\eta\rho^2 = .01$; Nature = $F(2, 57) = .30$, $p = .74$, $\eta\rho^2 = .01$; Weight = $F(2, 57) = .28$, $p = .78$, $\eta\rho^2 = .01$; see Figure 4).

In each of the framing conditions for the Nonword and Social System IRAP tasks (absent/ weak prior associations), the D-IRAP scores did not differ significantly from zero (all ts < 1.51, ps > .15), indicating neutral attitudes to the Nonwords as well as Capitalism relative to Socialism. In each of the framing conditions for the Nature and the Weight IRAP tasks (strong prior associations), the D-IRAP scores were a significant distance away from zero (all ts > 3.30, ps <.01), indicating the predicted pro-flower/anti-insect bias and the pro-thin/anti-fat bias.

Discussion

The main aim of the present study was to determine if default (positive) framing biases might compromise the validity of the IRAP (Barnes-Holmes et al., 2010a). According to previous research (Matlin, 2004; McKenzie & Nelson, 2003) people have a general and default bias to frame events and objects in a positive way (a Positive Framing Bias; PFB) and it is possible to manipulate this bias though framing. Based on the procedure and structure of the IRAP we reasoned that such a PFB might have the effect of biasing estimates of absolute attitudes. Indeed, such a bias might account for previous results which seem either counterintuitive (e.g., Hussey & Barnes-Holmes, 2012), or inconsistent with prior research (e.g., Roddy et al., 2011; 2012). We suggest that the effect of a positivity bias in this situation would lead to more rapid *True* responses to categories presented with positive words than *False* responses for the same stimuli. This in turn would lead to inflated estimates of positive attitudes towards the stimulus set being examined. If this is indeed the case, the validity of the IRAP as a measure of absolute implicit attitudes would be severely limited.

We set out to test three hypotheses: (H1) participants hold a default (as measured in the standard condition) Positive Framing Bias (PFB) which results in a decrease in the time required for them to press *True* to categories presented with positive words than to press *False*. (H2) framing the response task in different ways should influence the estimates of absolute implicit attitudes obtained by the IRAP, with weak prior associations being more malleable, and (H3) framing biases or direct manipulations would have no influence when a relative category analysis of the data is carried out.

The Effect of a Default Positive Framing Bias (H1)

In the standard condition the absolute IRAP scores were positively biased across all four IRAPs. In these IRAPs participants were generally faster at pressing *True* rather than *False* when associating any category with positive words. The implication of this finding is that researchers using the IRAP might incorrectly interpret their results as showing that participants have a positive (absolute) attitude towards abstract categories (e.g., social systems or atheism: O'Shea & Stewart, 2015). In contrast, the 'true' attitudes could be either neutral or even negative and so this finding calls into question the validity of the IRAP procedure. Of note, even the IRAP tasks that used stimuli with strong prior associations (Nature and Weight) were influenced by this PFB and, as predicted, a reduction in positivity was seen for categories that usually show a negative bias when assessed by other measures (e.g., insects and fat people), resulting in neutral absolute attitudes as measured by the IRAP.

These findings could account for at least some of the unusual results that have been previously published using the IRAP. For example, as detailed earlier, Roddy et al. (2011, 2012) reported a neutral bias towards fat people whereas we might expect a negative attitude. Similarly, Barnes-Holmes et al. (2010b) reported that participants evaluated black and white non-security related civilians with a gun as 'Safe' whereas we might expect that they would be viewed as dangerous. Both these results could be explained as an artefact caused by a

PFB. That is, participants are likely to be simplifying the task by picking just one of the two possible associations they have to perform on a block (usually the positive) and basing all responses on that association[14]. This bias in turn leads to an overestimation of the absolute positive attitude towards a stimulus set because, as we have shown, *True* responses to positive associations are faster than *False* responses.

The Influence of Framing on Estimates of Absolute Attitudes (H2)

We also determined the effect of directly manipulating the framing of the task on estimates of absolute implicit attitudes. If the IRAP results are influenced by framing, then directly manipulating the frame should have a related influence on the 'implicit attitude' (estimate) scores obtained when an absolute analysis of the data is calculated. As argued previously, we would expect that weak or absent prior associations would be more likely to be influenced by the framing effect than those with strong prior associations. This prediction was also confirmed. For the Nonword and Social System IRAP there was a strong effect of framing. That is, in the positive framing condition scores were elevated across all trial types

---

[14] After participants completed the current experiment those in the standard framing condition were asked to write down any strategies they used to complete the task. This question was not needed for the participants in the positive and negative framing condition because they were specifically given a strategy to follow. Of the 10 participants in the standard framing condition who made any references to focusing on mainly the associations, all mentioned the positive associations and none made reference to a focus on the negative associations. The remaining 10 could not describe any strategies they used other than remembering the two possible associations they had to perform on each block of trials. This suggests that when participants did spontaneously choose an explicit strategy they chose one based on positive associations.

whereas in the negative framing condition scores were reduced. In contrast, for the Nature and Weight IRAP the framing effects were smaller, particularly in the negative framing condition.

These outcomes suggest that the reason that using the absolute results from IRAP trial types are problematic is that responses are influenced strongly by how a person frames each block of trials. It is noteworthy that the way in which the associations are phrased before beginning an IRAP block (especially the sequence) are practically non-existent in published IRAP reports. Yet we have shown that they can have an influence on the results obtained, particularly if a researcher primes a participant in some way to focus on either the negative or positive associations. In summary, the IRAP claims to determine absolute implicit attitudes. However, as we have shown, framing effects can influence the absolute value of the scores obtained, making them unreliable at best. Such framing biases can arise naturally and by default (as in the standard condition) to frame things positively (the PFB). In addition, differences in the way the IRAP task is explained to the participants can have large influences (as in the positive and negative framing conditions).

The Influence of Framing on the Relative Results (H3)

The third hypothesis was that the framing effect would not have an influence on relative attitudes (when the data were analyzed using the standard IAT methodology). Use of stimuli with weak or no prior associations was expected to lead to neutral attitudes, while stimuli with strong prior associations were predicted to lead to findings showing the expected pro/anti-bias (e.g., pro-thin/anti-fat). Again all these predictions were confirmed across the four IRAPs. When analyzed in a relative manner we found the Nonword IRAP and the Social System IRAP indicated neutral attitudes towards the various category labels. For the Nature and Weight IRAPs, the results were consistent with previous IAT studies (i.e., pro-

flower/anti-insect; pro-thin/anti-fat). This suggests that that the IRAP can still be used when a relative analysis of the data is to be carried out.

One possible benefit of using the IRAP as a relative measure is that it appears to be difficult for participants to fake their responses (McKenna, Barnes-Holmes, Barnes-Holmes, & Stewart, 2007). However, the major disadvantages of the IRAP as a relative measure are that it is more complicated, takes longer to complete and has a higher attrition rate than the IAT (Golijani-Moghaddam et al., 2013).

Developing a New Absolute Measure of Implicit attitudes

In addition to the framing biases revealed here for the IRAP, other proposed absolute methods also suffer from problems. As discussed earlier, there are problems with both the Go/No-Go task and the Extrinsic Affective Simon Task for measuring absolute implicit attitudes (Bar-Anan & Nosek, 2013; De Houwer & De Bruycker, 2007). Initial results obtained using the single concept IAT variations (e.g., Karpinski & Steinman, 2006) have shown promise in measuring absolute implicit attitudes. However, the IAT and its derivatives are not without their limitations. For example, the sequence or order (Klauer & Mierke, 2005) in which participants carry out the critical blocks has been shown to influence the IAT results. For example beginning with the Fat Person-Positive and Thin Person-Negative block results in a weaker pro-thin/anti-fat IAT effect than when the presentation order is the other way around. However, it is possible that including more practice trials after the first critical block reduces this problem (Nosek, Greenwald, & Banaji, 2005). Nonetheless, it remains to be determined whether the PFB influences the results of the single concept IATs in a similar way to those of the IRAP.

Diversifying the tools used to measure implicit attitudes is likely to enhance our understanding of them. Consequently, the authors are currently developing a new implicit tool called the Simple Implicit Procedure (SIP) which was inspired by the present work. In

brief, this tool has parallels to the IRAP but, as the name suggests, rather than participants having to respond to two opposing associations during blocks, the SIP will only involve one association per block. For example, before each block in the IRAP participants could be told 'On this block please respond as if Thin Person is Negative and Fat Person is Positive', but during the SIP they could be told 'On this block please respond as if Thin Person is Negative' which is much simpler as it only requires one association to be kept in memory.

We anticipate that this change will reduce participants' PFB because they will be exposed to equal frequencies of instructions that are framed positively and negatively[15]. Keeping just one association in memory will reduce participants' cognitive load and hence their need to use a simplifying strategy/method to make the task easier. To ensure that participants do not focus solely on the positive and negative words and ignore the category stimuli, an inhibition/alternative response key will be used in some of the trials. For example participants will have to press the Space Bar when 'Fat Person' and any valenced word appears on a 'Thin Person' block.

The SIP will also bring the ability to measure a single concept (e.g. self-esteem) without the need for a comparison group. Additionally, this tool could act as both an absolute

---

[15] In the IRAP participants have to press "True" or "False" to positively valenced word and "True" or "False" to negative valenced words depending on which "Pro" block they are carrying out. The SIP, however, will require participants to press "True" to the positive valenced words in the positive association blocks and "True" to the negatively valenced words in the negative association blocks. The mean latency difference between pressing "True" on different block will be measured. Likewise "False" is pressed to negative words in the positive association blocks and to positive words in negative association blocks. Using this procedural set up should remove the PFB people have.

and relative measure by simply adding in a comparison group/category. This feature would enable researchers to obtain a general overview of participants' attitudes through the relative comparisons but a more complete understanding of implicit attitudes could be observed with the absolute results. The order or sequence effect apparent in the IAT (Klauer & Mierke, 2005) is not expected to have as much of an influence when using the SIP. For example in a Fat-Thin SIP participants will respond in one of four possible blocks of trials at any point (i.e. Thin Person-Positive, Thin Person-Negative, Fat Person-Positive, Fat Person-Negative). Each participant will experience a random sequence of these four blocks in the SIP (24 possible sequences) which represents an improvement over the two possible sequences available for the IATs. Pilot studies are currently being carried out to test the validity of the SIP.

Consideration of Special Populations

Although the current research implies that people in general have a PFB, it is possible that the results would not generalize to specific groups, namely those with depression or other emotional disorders. There is a wealth of evidence showing that these groups are in fact biased towards negative thought processes (e.g., Browning, Holmes, & Harmer, 2010). It would be illuminating to see if those with depression are more likely to focus on negative associations, resulting in a Negative Framing Bias (NFB) when conducting the IRAP. Indeed, Kosnes, Whelan, O'Donovan, and McHugh (2013) provided evidence for this suggestion using the IRAP with a sub-clinically depressed sample and a normative one. The authors claimed they were measuring participants' responses to future thinking (i.e., measuring the difference between pressing *True* and *False* to 'I expect' or 'I don't expect' which were presented with positive and negative words).

While the experimenters interpreted their results as the normative sample having positive future thinking and the subclinical having negative future thinking, the current findings suggest an alternative interpretation. That is, the statements 'I expect' and 'I don't

expect' are unlikely to evoke strong prior associations, similar to the Nonword IRAP and Social System IRAP above. Therefore, the cognitive heuristics or framing strategies a participant engages in (e.g., a normative sample having a PFB and a depressed sample having a NFB) may be the driving factor behind Kosnes et al.'s (2013) results.

Conclusion

The current study has provided the first empirical evidence that people have a tendency to spontaneously frame opposing associations in a positive way (Positive Framing Bias; PFB) when put under time pressure. This bias can be accentuated when participants are encouraged to focus on the positive associations and reversed when they are prompted to focus on the negative associations. These findings seriously question the validity of the IRAP as a tool for determining absolute implicit attitudes.

References

Bar-Anan, Y., & Nosek, B. A. (2013). *A comparative investigation of seven implicit measures of social cognition*. Unpublished manuscript. Retrieved from http://ssrn.com/abstract=2074556

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010a). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record, 60*, 527-542.

Barnes-Holmes, D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010b). 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.

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition, 18*, 329-353.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61*, 27-41.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 977-997.

Boroditsky, L. (2011). How language shapes thought. *Scientific American, 304*, 62-65.

Browning, M., Holmes, E. A., & Harmer, C. J. (2010). The modification of attentional bias to emotional information: A review of the techniques, mechanisms, and relevance to emotional disorders. *Cognitive, Affective, & Behavioral Neuroscience, 10*, 8-20.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word

frequency measure for American English. *Behavior Research Methods*, *41*, 977-990.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*, 3-25.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 76,* 387-404.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology, 37*, 1102-1117.

De Houwer, J., & De Bruycker, E. (2007). The implicit association test outperforms the extrinsic affective Simon task as an indirect measure of inter-individual differences in attitudes. *British Journal of Social Psychology*, *46,* 401-421.

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

Gervais, W. M., Shariff, A. F., & Norenzayan, A. (2011). Do you believe in atheists? Distrust is central to anti-atheist prejudice. *Journal of Personality and Social Psychology*, 101, 1189-1206.

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The implicit relational assessment procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2,* 105-119.

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

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*, 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.

Harrison, D. P., Stritzke, W. G. K., Fay, N., Ellison, T. M., & Hudaib, A. (2014). Probing the implicit suicidal mind: Does the Death/Suicide implicit association test reveal a desire to die, or a diminished desire to live? *Psychological Assessment*, *26*, 831-840.

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics*, *42*, 25-70.

Hussey, I., & Barnes-Holmes, D. (2012, May). *Experiential avoidance and pro-death implicit attitudes in normative participants.* Paper presented at the Northern Ireland Branch Annual Conference, Co. Fermanagh.

Karpinski, A. (2004). Measuring self-esteem using the Implicit Association Test: The role of the other. *Personality and Social Psychology Bulletin*, *30*, 22-34.

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.

Klauer, K. C., & Mierke, J. (2005). Task-set inertia, attitude accessibility, and compatibility-order effects: New evidence for a task-set switching account of the Implicit Association Test effect. *Personality and Social Psychology Bulletin*, *31*, 208-217.

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

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis.

Organizational Behavior and Human Decision Processes, 75, 23-55.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., & Nosek, B. A. (in press). Reducing implicit racial preferences: A comparative investigation of 17 interventions. *Journal of Experimental Psychology.*

Leech, G. N. (2006). *A Glossary of English Grammar*. New York: Columbia University Press.

Matlin, M. W. (2004). Pollyanna Principle. In R. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp.257-271. New York: Psychology Press.

Matthews, W. J., & Dylman, A. S. (2014). The language of magnitude comparison. *Journal of Experimental Psychology: General, 143*, 510-520.

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.

McKenzie, C. R., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review, 10*, 596-602.

Nolan, J, Murphy, C., & Barnes-Holmes, D. (2013). Implicit relational assessment procedure and body-weight bias: Influence of gender of participants and targets. *The Psychological Record, 63*, 467-488.

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. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166–180.

O'Brien, K. S., Hunter, J. A., & Banks, M. (2006). Implicit anti-fat bias in physical educators: Physical attributes, ideology and socialization. *International Journal of Obesity, 31*, 308-314.

O'Shea, B (2015a). [The IRAP fails to measure abstract concepts absolutely]. Unpublished raw data.

O'Shea, B. (2015b). Capitalism versus a new economic model: Implicit and explicit attitudes of protesters and bankers. *Social Movement Studies, 14*, 311-330.

O'Shea, B., & Stewart, I. (2015). *Implicit and explicit attitudes towards atheism and religious belief: Limitations with the IRAP.* Manuscript submitted for publication.

Peterson, C. (2000). The future of optimism. *American Psychologist, 55*, 44-55.

Randall, J. R., Rowe, B. H., Dong, K. A., Nock, M. K., & Colman, I. (2013). Assessment of self-harm risk using implicit thoughts. *Psychological Assessment, 25,* 714-721.

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.

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*, 688-694.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology, 56*, 283-294.

Sherman, M. A. (1973). Bound to be easier? The negative prefix and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 12*, 76-84.

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116*, 21-27.

Teachman, B. A., & Allen, J. P. (2007). Development of social anxiety: Social interaction

predictors of implicit and explicit fear of negative evaluation. *Journal of Abnormal Child Psychology, 35*, 63-78.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry, 48*, 59-65.

Figure Captions

Figure 1: Screen shot examples of the four IRAP trial types. The picture categories ('Thin Person' or 'Fat Person'), target word (Active, Ugly, and Healthy etc.) and response options (*True* and *False*) appeared simultaneously on each trial. Asterisks indicate the correct response option to press on either pro-Thin or pro-Fat blocks of trials.

Figure 2: Mean *False* Minus *True* D-IRAP scores as a function of Category, Word Valence and IRAP task. Values above zero indicate that responses for pressing *True* were faster than responses for pressing *False* while values below zero indicate a faster *False* vs. *True* response bias. Error bars with 95% confidence intervals have been included.

Figure 3 (Top Left): The mean absolute D-IRAP (estimate) trial type X IRAP task interaction. Points above the zero line indicate a positive attitude towards the category under investigation. This occurs when *True* responses are faster than *False* responses for categories presented with positive words and also when *False* response are faster than *True* response for categories presented with negative words. Points below the zero line indicate a negative attitude. This occurs when *False* responses are faster than *True* response for positive word associations and also when *True* response are faster than *False* response for negative word associations. Error bars that cross the zero mark indicate a statistically neutral attitude; those that do not cross zero indicate a significant positive or negative attitude.

(Top Right) The mean absolute D-IRAP (estimate) trial type X framing condition interaction.

(Bottom Left) The mean absolute D-IRAP (estimate) IRAP task X framing condition interaction. The IRAP tasks are ordered from no prior associations (Nonword IRAP; leftmost) to strong prior associations (Weight IRAP; rightmost).

(Bottom Right) The mean combined/averaged absolute D-IRAP (estimate) results for each category.

Figure 4: The mean relative D-IRAP score for the four IRAP tasks.

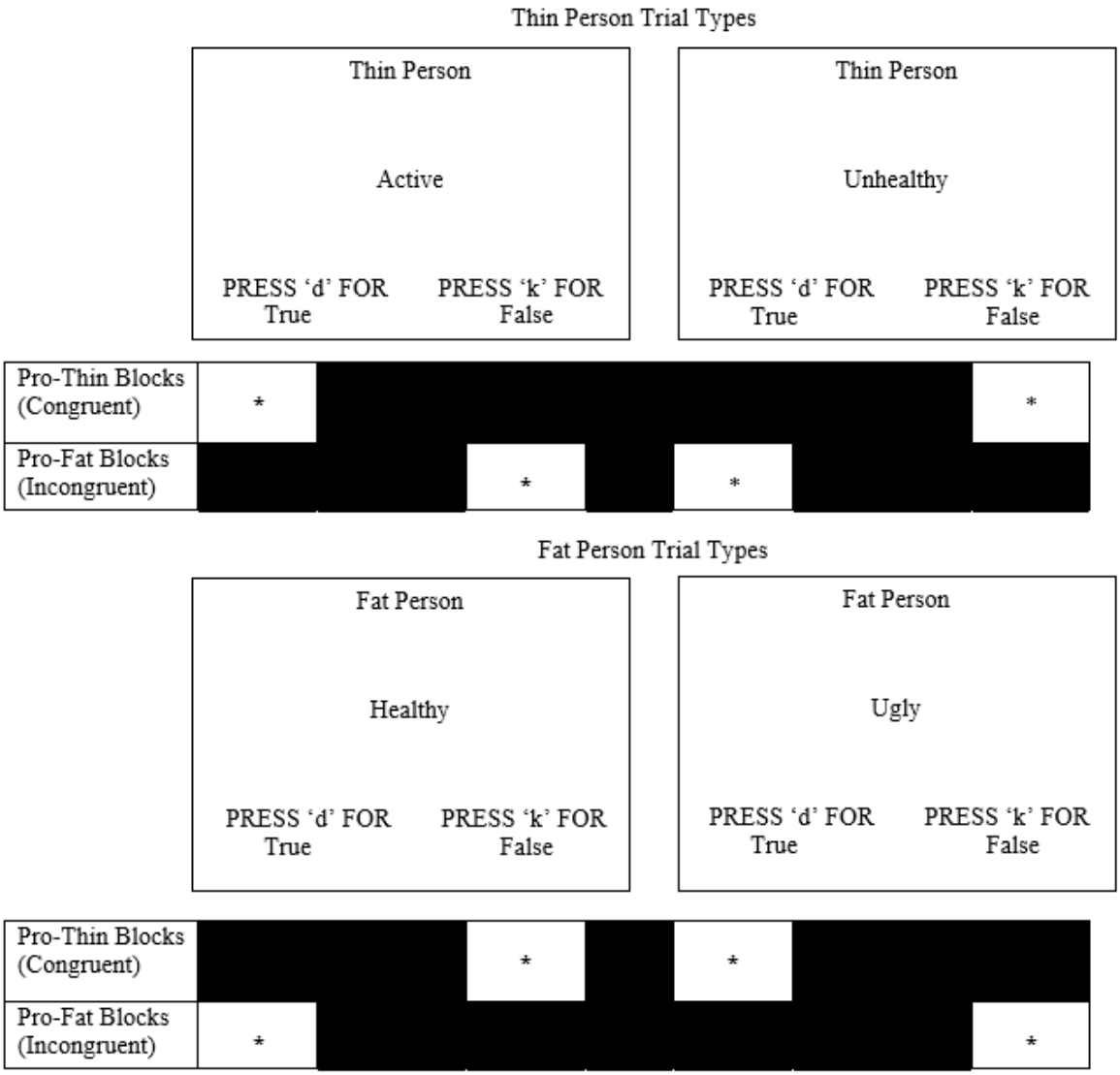

Thin Person Trial Types
Thin Person
Active
PRESS 'd' FOR True
PRESS 'k' FOR False
Thin Person
Unhealthy
PRESS 'd' FOR True
PRESS 'k' FOR False
Pro-Thin Blocks (Congruent)
Pro-Fat Blocks (Incongruent)
Fat Person Trial Types
Fat Person
Healthy
PRESS 'd' FOR True
PRESS 'k' FOR False
Fat Person
Ugly
PRESS 'd' FOR True
PRESS 'k' FOR False
Pro-Thin Blocks (Congruent)
Pro-Fat Blocks (Incongruent)

Figure 1.

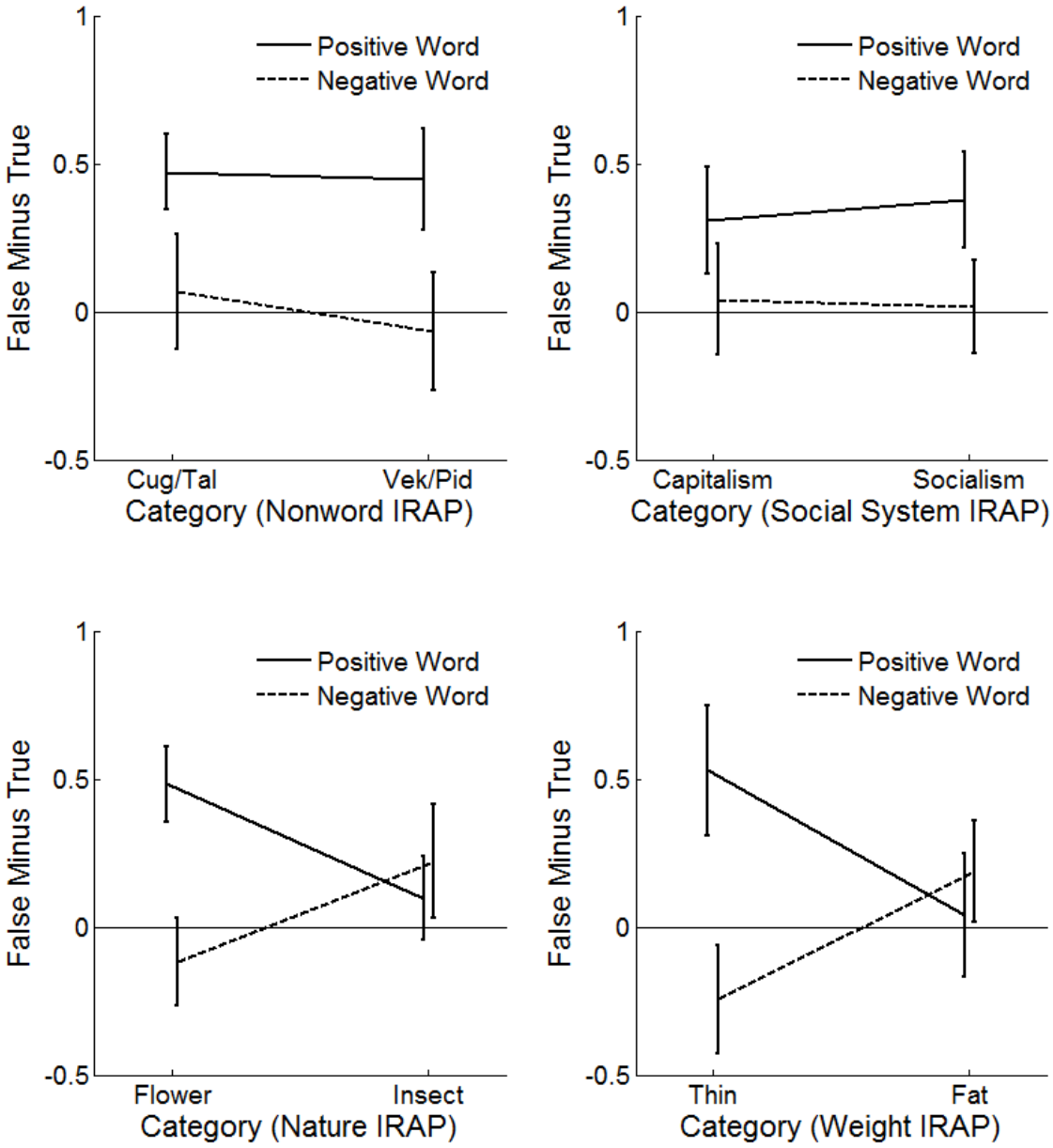
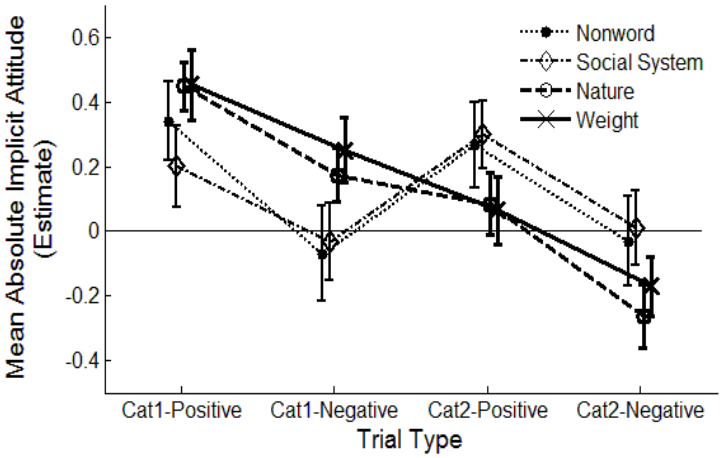
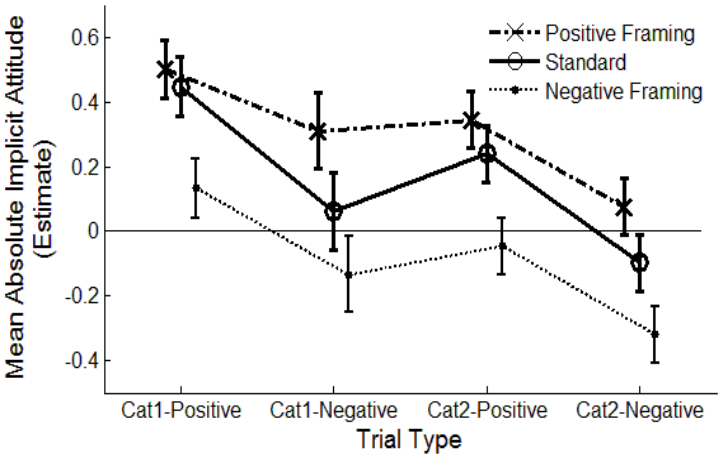
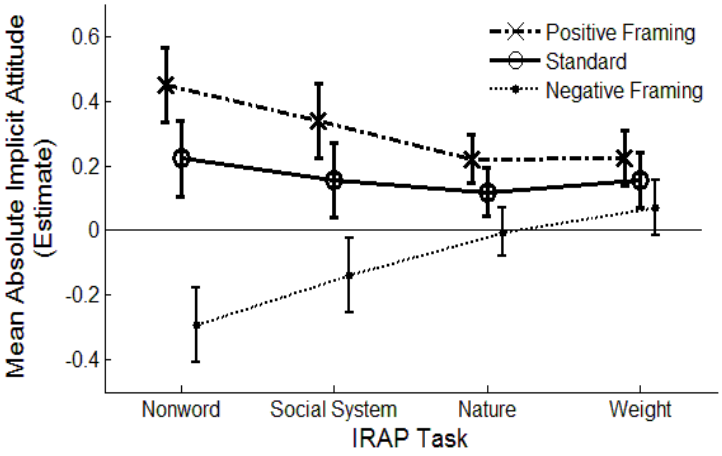

Positive Word
Negative Word
False Minus True
1
0.5
0
-0.5
Cug/Tal
Vek/Pid
Category (Nonword IRAP)
Capitalism
Socialism
Category (Social System IRAP)
Flower
Insect
Category (Nature IRAP)
Thin
Fat
Category (Weight IRAP)

Figure 2.


Mean Absolute Implicit Attitude (Estimate)
0.6
0.4
0.2
0
-0.2
-0.4
Nonword
Social System
Nature
Weight
Cat1-Positive
Cat1-Negative
Cat2-Positive
Cat2-Negative
Trial Type


Mean Absolute Implicit Attitude (Estimate)
0.6
0.4
0.2
0
-0.2
-0.4
Positive Framing
Standard
Negative Framing
Cat1-Positive
Cat1-Negative
Cat2-Positive
Cat2-Negative
Trial Type


Mean Absolute Implicit Attitude (Estimate)
0.6
0.4
0.2
0
-0.2
-0.4
Positive Framing
Standard
Negative Framing
Nonword
Social System
Nature
Weight
IRAP Task

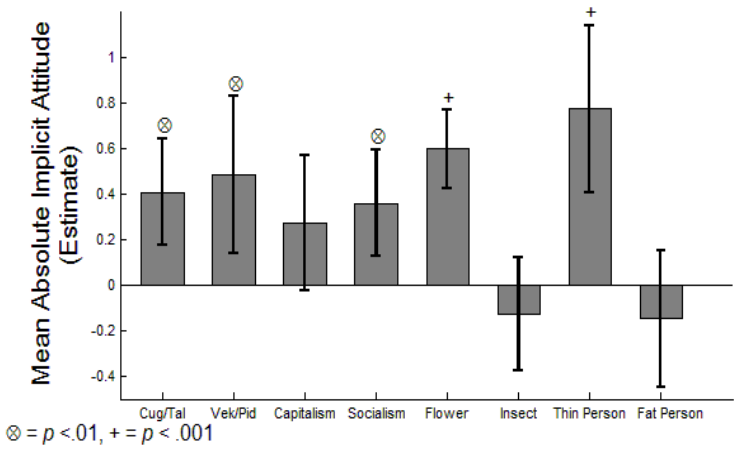

Mean Absolute Implicit Attitude (Estimate)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
⊗
⊗
⊗
+
+
Cug/Tal
Vek/Pid
Capitalism
Socialism
Flower
Insect
Thin Person
Fat Person
⊗ = p <.01, + = p < .001

Figure 3

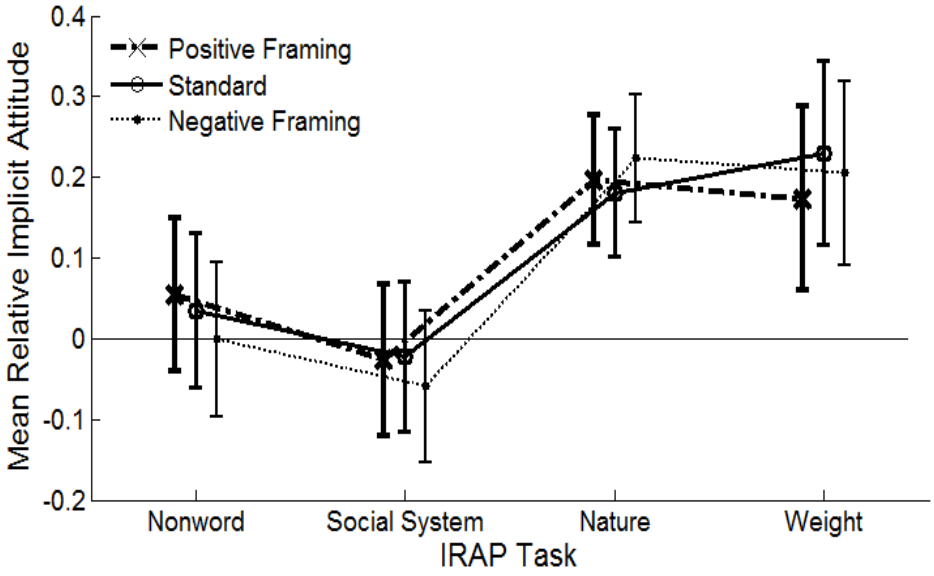

0.4
0.3
0.2
0.1
0
-0.1
-0.2
Mean Relative Implicit Attitude
Positive Framing
Standard
Negative Framing
Nonword
Social System
Nature
Weight
IRAP Task

Figure 4.

Online supplementary materials for:

Measuring Implicit Attitudes: A Positive Framing Bias Flaw in the Implicit Relational Assessment Procedure (IRAP)

Brian O'Shea

Derrick G. Watson

Gordon D. A. Brown

Table S1: The Nonword and Social System IRAP Stimuli (absent/weak associations)

Nonword IRAP		Social System IRAP	
Category 1:	Category 2:	Category 1:	Category 2:
Cug or Tal	Vek or Pid	Capitalism	Socialism
Target stimuli congruent with category 1:	Target stimuli congruent with category 2 :	Target stimuli congruent with category 1:	Target stimuli congruent with category 2:
Happy	Problems	Equality	Poverty
Alive	Died	Fair	Inequality
Positive	Negative	Community	Greed
Freedom	Hated	Wealth	Dictator
Care	Bad	Sharing	Awful
Lucky	Sick	Freedom	Corrupt
Response Option 1:	Response Option 2:	Response Option 1:	Response Option 2:
True	False	True	False

Table S2: The Nature and Weight IRAP Stimuli (strong associations)

Nature IRAP		Weight IRAP	
(Category 1: Flower)	(Category 2: Insect)	(Category 1: Thin	(Category 2: Fat
Bluebell	Mosquito	Person)	Person)
Daffodil	Spider	6 different images of	6 different images
Tulip	Wasp	thin people	of fat people
Target stimuli congruent with category 1:	Target stimuli congruent with category 2:	Target stimuli congruent with category 1:	Target stimuli congruent with category 2:
Enjoy	Unpleasant	Good	Bad
Cheer	Poison	Active	Sloppy
Happy	Evil	Attractive	Ugly
Lovely	Damage	Healthy	Mean
Friend	Ugly	Popular	Lazy
Pleasing	Hurt	Nice	Unhealthy
Response Option 1:	Response Option 2:	Response Option 1:	Response Option 2:
True	False	True	False

The steps involved in calculating the D-IRAP scores were as follows (example given for the Weight IRAP version):

(1) Only response latency data from the six test blocks were used; (2) latencies above 10,000 ms were discarded; (3) if latencies from more than 10% of a participant's trials throughout the 6 test blocks were less than 300 ms, that participant was removed from the analyses; (4) for each IRAP task, 12 standard deviations for the four trial type latencies (e.g., **Thin Person**-Congruent responses, **Thin Person**-Incongruent responses, **Fat Person**-Incongruent responses, and **Fat Person**-Congruent responses) were calculated: four for the responses latencies from Test Blocks 1 and 2, four from the latencies from Test Blocks 3 and 4, and a further four from Test Blocks 5 and 6; (5) 24 mean latencies were then calculated for the four trial types in each of the six test blocks; (6) difference scores were calculated for each of the four trial types by subtracting mean latencies of the pro-**Thin** trials from mean latencies of the pro-**Fat** trials for each test block pair; (7) each difference score was then divided by its corresponding standard deviation from step 4, yielding 12 D-IRAP scores, one score for each trial type for each pair of test blocks; (8) four overall trial type D-IRAP scores were calculated by averaging the three scores for each of the four trial types across the three pairs of test blocks. These calculations revealed the absolute/non-relative results. To compute the relative comparison, equivalent to that of the IAT, an overall D-IRAP score was calculated by averaging all the 12 trial type D-IRAP scores obtained in step 7 above.

Assessing attitudes obtained for the individual IRAPs

Each of the four IRAPs (Nonword, Social System, Nature, and Weight) was analyzed separately using the estimates of implicit attitude scores to address more directly if participants had particularly positive attitudes to categories presented with positive words in the positive and standard framing conditions compared to the negative framing condition.

Nonword IRAP

The absolute D-IRAP (estimate) data (see Figure S1; Top Left) for the Nonword IRAP were analysed with a 2 (stimulus category: category 1, category 2) X 2 (word valence: positive, negative) x 3 (framing condition: positive, standard, negative) mixed ANOVA. Stimulus category 1 consisted of the Nonwords Cug and Tal, category 2 consisted of the Nonwords Vek and Pid. Framing condition was a between-subjects factor, stimulus category and word valence were within-subjects factors.

This analysis revealed a significant main effect of IRAP word valence, $F(1, 57) = 59.82$, $p < .001$, $\eta\rho^2 = .52$; scores were higher for positive IRAP words than for negative IRAP words. There was also a significant main effect of framing condition, $F(2, 57) = 42.86$, $p < .001$, $\eta\rho^2 = .61$. The standard framing had a high mean D-IRAP (estimate) score ($M = .22$); the positive framing produced the highest mean D-IRAP (estimate) score ($M = .45$), and negative framing had the lowest ($M = -.29$). LSD pairwise comparisons showed that all the framing condition D-IRAP (estimate) scores differed significantly from each other (all ts > 2.70, ps < .01). The main effect of IRAP category was not significant, $F(1,57) = .10$, $p > .05$, $\eta\rho^2 = .00$, and nor were any of the interactions (all Fs < 2.03, $ps > .14$).

Follow-up t-tests tested whether each absolute D-IRAP (estimate) score differed from zero. A positive value indicates a positive attitude to the stimulus category and a negative value indicates a negative attitude. For the standard framing condition, scores were greater

than zero when associating the Nonword stimuli with positive words (all *t*s > 5.19, *p*s < .001). Scores did not differ from zero for associating Nonword stimuli with negative words (all *t*s < .69, *p*s > .50). For the positive framing condition, all D-IRAP (estimate) scores were significantly greater than zero (all *t*s > 3.12, *p*s < .01). For the negative framing condition scores were significantly below zero for associating Nonwords with negative words (all *t*s > 3.44, *p*s < .01) but did not differ for associating Nonwords with positive words (all *t*s < 1.58, *p*s > .13) [16].

Social System IRAP

[16] Two pilot studies were conducted that used various response options to ensure that the word frequency or length of these response options was not causing the PFB. Pilot Study 1 tested 20 participants and was similar in structure to the standard condition. Each participant completed three IRAPs which incorporated various combinations of Nonwords and the positive and negative target words were taken from Barnes-Holmes, Barnes-Holmes, Power, Hayden, Milne, & Stewart (2006). The response options used in each IRAP were: 'Similar' and 'Opposite'; 'Similar' and 'Different'; 'True' and 'False'. Pilot study 2 had two conditions, standard and negative framing, similarly to the current experiment, with fifteen different participants in each condition. Various Nonwords were used and the same positive and negative words used in the Nonword IRAP above were presented. Each participant completed 5 IRAPs that had different response options (i.e. 'Confirmation' and 'No'; Symbol of a Thumb Up and a Thumb Down; Picture of a Happy Face and a Sad Face; 'Similar' and 'Opposite'; 'Not Different' and 'Different'. These pilot studies produced comparable results to the Nonword IRAP.

The absolute D-IRAP (estimate) Social System results showed a similar pattern to those of the Nonword IRAP and are shown in Figure S1 (Top Right). A 2 (stimulus category: Capitalism, Socialism) x 2 (word valence: positive, negative) x 3 (framing condition: standard, positive, negative) mixed ANOVA revealed a significant main effect of IRAP word valence, $F(1, 57) = 27.51$, $p < .001$, $\eta\rho^2 = .33$; positive IRAP words had higher scores than did negative IRAP words. There was also a significant main effect of framing condition, $F(2, 57) = 17.46$, $p < .001$, $\eta\rho^2 = .38$. A high mean D-IRAP (estimate) score was found for standard framing ($M = .16$), with the positive framing producing the highest ($M = .34$), and negative framing producing the lowest ($M = -.14$) values. LSD pairwise comparisons showed that all the framing condition D-IRAP (estimate) scores differed significantly from each other (all ts > 2.24, ps < .05). The main effect of IRAP category was not significant, $F(1,57) = 1.76$, $p > .05$, $\eta\rho^2 = .03$, and all the interactions were non-significant (all Fs < 2.46, ps > .09).

Follow-up t-tests showed that in the standard condition scores were greater than zero for associating Capitalism and Socialism with positive words (all ts > 3.45, ps < .01), but scores did not differ from zero for associating Capitalism and Socialism with negative words (all ts < .44, ps > .67). In the positive framing condition, all but the Capitalism negative word association D-IRAP (estimate) scores (t =1.81, p =.09) were significantly greater than zero (all ts > 3.91, ps < .01). In the negative framing condition scores were significantly below zero for associating Capitalism and Socialism with negative words (all ts > 2.64 ps < .05) but did not differ for associating Capitalism and Socialism with positive words (all ts < 1.61, all ps > .13).

Nature IRAP

A 2 (stimulus category: Flower, Insect) x 2 (word valence: positive, negative) x 3 (framing condition: standard, positive, negative) mixed ANOVA showed a significant main

effect of IRAP word valence, $F(1,57) = 49.98$, $p < .001$, $\eta\rho^2 = .47$: positive IRAP words had higher scores than did negative IRAP words (see Figure S1; Bottom Left). This time a significant main effect was found for stimulus category, $F(1,57) = 76.67$, $p < .001$, $\eta\rho^2 = .54$: more positive attitudes were found for flowers than for insects. A significant main effect of framing condition was also found, $F(2, 57) = 9.01$, $p < .001$, $\eta\rho^2 = .24$. A mean D-IRAP (estimate) score of .12 was found in the standard framing conditions. This score was elevated in the positive framing condition ($M = .22$), and reduced in the negative framing condition ($M = -.01$). LSD pairwise comparisons showed that the standard framing D-IRAP (estimate) score was significantly different from the negative framing score ($t = 2.30$, $p < .05$), and marginally different from the positive framing score ($t = 1.94$, $p < .06$). The positive and negative framing D-IRAP (estimate) scores also differed from each other ($t = 4.25$, $p < .001$). None of the interactions were found to be significant (all Fs < 1.40, ps > .24).

Follow-up t-tests showed that in the standard condition, scores were significantly greater than zero for associating flowers with positive words ($t = 7.69$, $p < .001$), and below zero for associating insects with negative words ($t = 2.28$, $p < .05$), but scores did not differ from zero when associating flowers with negative words or insects with positive words (all ts < 1.54, ps > .14). In the positive framing condition, scores were greater than zero for the flower and insect categories presented with positive words as well as the flower category presented with negative words (all ts > 2.75, ps < .05). The score for the insect category presented with negative words was significantly below zero ($t = 2.36$, $p < .05$). For the negative framing condition, scores were significantly above zero for the flower category presented with positive words ($t = 5.87$, $p < .001$) and below zero for the insect category presented with negative words ($t = 5.02$, $p < .001$). The other two D-IRAP (estimate) scores were not significantly different from zero (all ts <1.10, ps > .29).

Weight IRAP

A 2 (stimulus category: Thin Person, Fat Person) x 2 (word valence: positive, negative) x 3 (framing condition: standard, positive, negative) mixed ANOVA showed a significant main effect of IRAP word valence $F(1, 57) = 26.86$, $p < .001$, $\eta\rho^2 = .32$: positive IRAP words produced higher scores than negative IRAP words (see Figure S1; Bottom Right). A significant main effect was found for stimulus category, $F(1, 57) = 37.79$, $p < .001$, $\eta\rho^2 = .39$: more positive attitudes were found for the Thin Person category than for the Fat Person category, and a significant main effect was also found for framing condition, $F(1, 57) = 3.28$, $p < .05$, $\eta\rho^2 = .10$. A positive mean D-IRAP (estimate) score was found in the standard framing condition ($M = .16$); this score was higher in the positive framing condition ($M = .22$) and lower in the negative framing condition ($M = .07$). LSD pairwise comparisons revealed that only the positive and the negative framing condition D-IRAP (estimate) scores differed from each other ($t = 2.56$, $p < .05$). The remaining interactions were non-significant (all Fs < .83, $ps > .44$).

Follow-up t-tests showed that in the standard framing condition the absolute D-IRAP (estimate) scores for the Thin Person category presented with positive words and Thin Person category presented with negative words were significantly greater than zero (all ts > 2.68, ps < .05). The scores for the Fat Person category presented with negative words was significantly below zero ($t = 2.17$, $p < .05$) and the scores for the Fat Person category presented with positive words did not differ from zero ($t = .371$, $p > .05$). In the positive framing condition the scores for Thin Person positive words, Thin Person negative words and Fat person positive words were all significantly greater than zero (all ts > 2.31, ps < .05). Score for Fat Person negative words did not differ from zero ($t = 1.24$, $p > .05$). In the negative framing condition the scores for the Thin Person category presented with positive

word was greater than zero (t = 3.99, p < .01) and the scores for the Fat Person category presented with negative words was below zero (t = 3.08, ps < .01). Both Thin Person negative word and Fat Person positive word scores did not differ from zero (all ts < 2.05, ps > .05).

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*, 169-177.

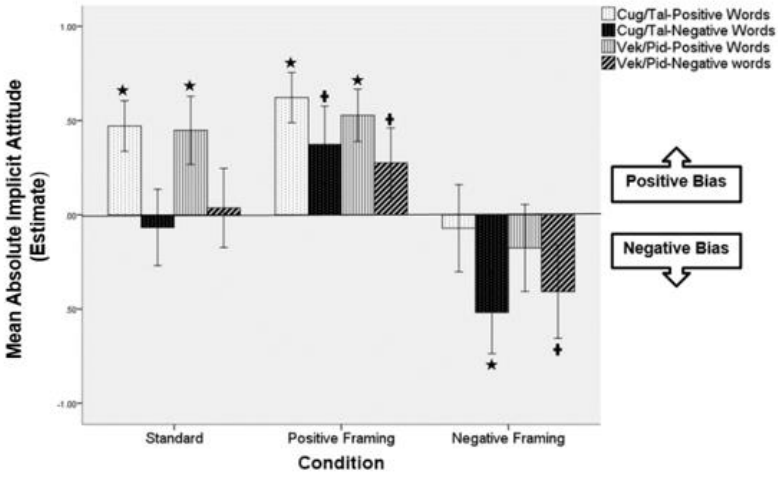
Figure Caption

Figure S1 (Top Left): The four mean absolute Nonword D-IRAP (estimate) scores for each of the three framing conditions. Bars above the zero line indicate a positive attitude towards the category under investigation. This occurs when *True* responses are faster than *False* responses for categories presented with positive words and also when *False* response are faster than *True* response for categories presented with negative words. Points below the zero line indicate a negative attitude. This occurs when *False* responses are faster than *True* response for positive word associations and also when *True* response are faster than *False* response for negative word associations. Error bars with 95% confidence intervals have been included. Error bars that cross the zero mark indicate a statistically neutral attitude; those that do not cross zero indicate a significant positive or negative attitude (see stars).

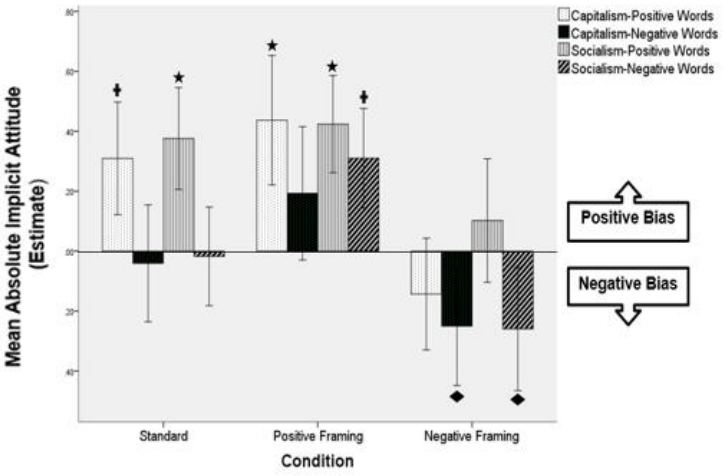
(Top Right): The four mean absolute Social System D-IRAP (estimate) scores for each of the framing conditions.

(Bottom Left): The four mean absolute Nature D-IRAP (estimate) scores for each of the framing conditions.

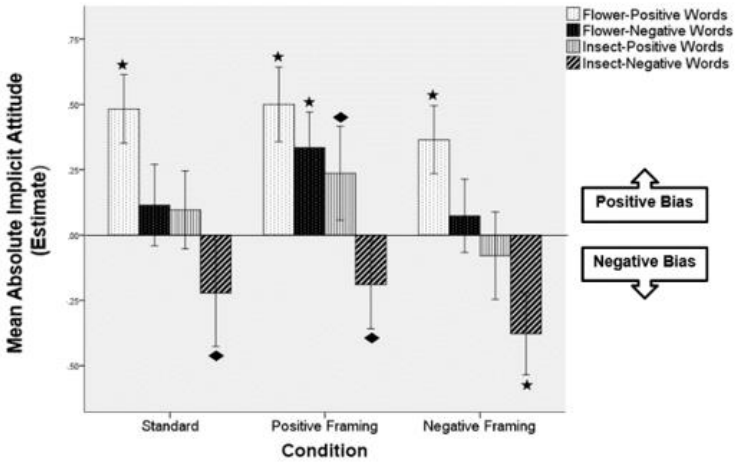
(Lower Right): The four mean absolute Weight D-IRAP estimate scores for each of the framing conditions.



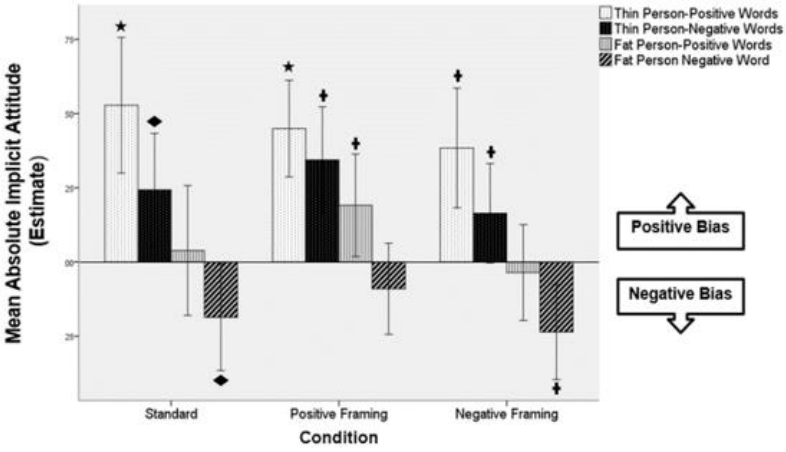
★ = $p < .001$, ✝ = $p < .01$



★ = $p < .001$, ✝ = $p < .01$, ◆ = $p < .05$



★ = $p < .001$, ◆ = $p < .05$



★ = $p < .001$, ✝ = $p < .01$, ◆ = $p < .05$

Figure S1