Are relational implicit measures sensitive to relational information?

Jamie Cummins & Jan De Houwer

Ghent University, Belgium

Author note

JC & JDH, Department of Experimental Clinical and Health Psychology, Ghent University. This research was conducted with the support of Grant BOF16/MET_V/002 to JDH. Correspondence should be sent to jamie.cummins@UGent.be.
SENSITIVITY TO RELATIONAL INFORMATION

Abstract

Evidence increasingly suggests that information about the relation between stimuli can impact responses on implicit measures, but those measures were not designed to capture this information in a controlled manner. Relational implicit measures have therefore been developed to assess such relational information in a more specific manner. Such measures have been used to assess relational beliefs in a variety of contexts. However, despite numerous theoretical inferences having been made under the premise that these measures validly measure relational content, this assumption has not been directly tested. Across four preregistered experiments (N = 747) participants learned information about different relations between two fictitious social groups and traits (e.g., Niffites are kind, Luupites should be kind). We then tested whether relational implicit measures (the RRT, aIAT, and PEP) reflected this information. Overall, only a mousetracking variant of the PEP effectively produced the expected effects for both are- and should-based relational information. Our results suggest that many relational implicit measures (as currently used) are not sensitive to relational information in ways previously assumed, although the use of alternative scoring methods that incorporate accuracy information may represent a step towards improving this sensitivity.

Keywords: Implicit measures, relational information, Relational Responding Task, Autobiographical IAT, Propositional Evaluation Paradigm
Are relational implicit measures sensitive to relational information?

Assessing people’s automatic thoughts and evaluations is critical to social psychologists (Petty et al., 2008). Typically, a class of measurement procedures whose outcomes are referred to as implicit measures are used by psychologists in attempts to measure such phenomena (Gawronski & De Houwer, 2014). These measures have seen use in contexts as varied as racial discrimination (Calanchini et al., 2020), gender stereotypes (Ye & Gawronski, 2018), intergroup attitudes more generally (Kurdi et al., 2019), and self-esteem (Greenwald & Farnham, 2000), to name but a few. Most implicit measures have been developed from the perspective that the psychological constructs claimed to be measured by these tasks are purely associative in nature (Corneille & Hütter, 2020; Hughes et al., 2011). For example, these measures have been claimed to capture “implicit stereotypes”, which have been conceptualized as associations between social groups and traits (Greenwald et al., 2002). However, some recent findings do not fit well with this associative perspective. For instance, evaluative conditioning effects on implicit measures vary in magnitude depending on whether relations between stimuli are presented as causal, predictive, or incidental (Hughes et al., 2019; see also Moran et al., 2021). Likewise, implicit measures may show dissociations when assessing different types of relations, and subsequently predict different behaviors (e.g., implicit liking vs. implicit wanting; Grigutsch et al., 2019; although see Tibboel et al., 2015 for contradictory results). Due to these findings, an alternative “propositional” perspective has emerged: namely, the perspective that the purported construct measured by these tasks can also by relational in nature (for a review see De Houwer et al., 2020).

From this propositional perspective, psychologists have recently started developing so-called relational implicit measures: implicit measures whose procedures are claimed to capture relational information. Although only emerging, these measures have already
SENSITIVITY TO RELATIONAL INFORMATION

produced several interesting findings; for instance, that descriptive and prescriptive gender stereotypes are dissociable (Cummins & De Houwer, 2019), and that body dissatisfaction is predicted by dissociations between actual vs. ideal body image assessed in these tasks (Heider et al., 2018). Likewise, these measures have successfully predicted racial- and gender-based prejudice (Müller & Rothermund, 2019; Cummins & De Houwer, 2021).

However, not all uses of relational implicit measures have provided utility. Cummins and colleagues (2021) found that implicit endorsements of descriptive (“I am a drinker”) vs. prescriptive (“I should be a drinker”) drinking self-identity statements provided little unique utility in predicting drinking-related outcomes. Other studies have found a consistent pattern wherein descriptive relations (i.e., are-relations) exhibit effects, while any other type of relation exhibits null effects (e.g., want-relations; should-relations; Heider et al., 2018; Glashouwer et al., 2018).

Considering these mixed findings, it is important to further test the validity of relational implicit measures. The claim that a measure qualifies as a relational implicit measure is valid only if it can be shown that the measures are “implicit” (typically defined as being produced under one or more automaticity conditions, Moors & De Houwer, 2006, but see Corneille & Hütter, 2020 for discussion on the meaningfulness of the “implicit” term) and that they are relational (i.e., that they are sensitive to how concepts are related). In this paper, we focus on testing the second assumption. Only to the extent that it is possible to find evidence for the relational nature of relational implicit measures should one take serious theoretical claims that are made on the basis of these measures. Consider, for instance, the null effects that have been found in past research for want- and should- relations (e.g., Heider et al., 2018). Until there is sound evidence that relational implicit measures are sensitive to those relations, it is unclear whether those null effects reflect the absence of relational beliefs or the incapacity of the measures to pick up those beliefs.
SENSITIVITY TO RELATIONAL INFORMATION

As noted by De Houwer et al. (2009), experimental studies can provide a crucial contribution to research on the validity of implicit measures. In line with this proposal, we tested whether relational implicit measures reflect experimentally-induced relations between stimuli. Indeed, novel implicit evaluations and stereotypes have already been extensively induced and manipulated in similar studies (e.g., Charlesworth et al., 2020; De Houwer et al., 1998; Cummins et al., 2018; Cummins & Roche, 2020). In this paper, we used this “training-and-testing” strategy with three relational implicit measures (the RRT, the aIAT, and the PEP). We provided a learning intervention about two fictitious social groups: Niffites and Luupites. Participants learned two different stereotypes between each social group and different traits (e.g., Niffites are direct, Luupites are kind, Niffites should be kind, Luupites should be direct). We then tested whether the relational implicit measures produced learning-consistent effects for each stereotype relation.

Method

All materials, data, processing and analysis scripts, and preregistrations can be found on the Open Science Framework

(https://osf.io/qw6jr/?view_only=cafbeff4bc234fbc930c309819a70229 ;
https://osf.io/4v8tg/?view_only=d2da873eaebe49e0ad1c3085260f99b9 ;
https://osf.io/5kf9z/?view_only=59861bdad4fc4b1ca41a46967e7535ae ; and
https://osf.io/ydf27/?view_only=ded4644c172040cb8b033690541e2cd0). Deviations from our preregistration are detailed in the Deviations from Preregistration document. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

Data were collected online via prolific.ac. Participants were paid at a rate of £5 per hour. We used a priori power analyses to determine sample sizes for all experiments (220
SENSITIVITY TO RELATIONAL INFORMATION

participants in Experiments 1 and 3 to achieve 95% power to detect a small-to-medium Cohen’s d interaction effect (0.35), 150 participants in Experiments 2a and 2b to achieve 95% power to detect a small Cohen’s d main effect (0.3); see preregistrations for more information). Participants were included in analyses only if they (1) learned all information within 3 learning-phase cycles; (2) recalled these relations after the implicit measure; (3) met screening criteria for the respective implicit measure (see data preparation section); and (4) indicated that their data should be included in our analyses, or did not provide a valid reason to exclude their data. Experiment 1 consisted of a final sample of 227 participants (112 male, 111 female, 4 non-binary; 114 completed the standard RRT, 113 completed the modified RRT; Age M = 26.02, SD = 7.29). The final sample of Experiment 2a consisted of 154 participants (100 male, 53 female, 1 non-binary; Age M = 26.71, SD = 7.88). Experiment 2b’s final sample consisted of 153 participants (83 male, 68 female, 2 no gender given; Age M = 26.34, SD = 7.93). The final sample of Experiment 3 consisted of 230 participants (121 male, 109 female; 120 completed the MT-PEP, 110 completed the RT-PEP; Age M = 28.83, SD = 11.09).

Procedure.

All materials were programmed and administered in Inquisit 5, except for Experiment 3 (in lab.js; Henninger et al., 2019). Participants firstly provided demographic information, then completed the learning phase. Information was presented about two social groups (Niffites and Luupites) and their relationship with two traits (direct and kind). Participants learned via vignettes that one group is direct and not kind and should be kind and not direct, and that the other group is kind and not direct and should be direct and not kind. The trait which one group had was always the trait which they should not have, and the trait which they should have was always the trait they did not have. Information was counterbalanced across participants. Participants were then presented with these eight statements about the
SENSITIVITY TO RELATIONAL INFORMATION

relationships between the groups and traits for 30 seconds. Participants then answered eight multiple choice questions probing for this information. If participants failed to answer all eight questions correctly, then they were presented the eight sentences again for another thirty seconds and required to answer the questions again. This cycle continued three times or until participants answered all questions correctly. Participants then completed an implicit measure (varied between experiments) assessing both “are” and “should” relations. Since the RRT and aIAT can only assess one relation at a time, we administered are- and should-variants of each measure in a counterbalanced order.

Blocks/trials in implicit measures are commonly referred to as either “consistent” or “inconsistent” to clarify different response requirements. We adopt this strategy here. Blocks/trials with sentences using the “are” relation are referred to as “consistent” if they are consistent with the trained are-relations, and as “inconsistent” if they are inconsistent with are-relations. Likewise, blocks/trials with sentences using the “should” relation are referred to as “consistent” if they are consistent with the trained should-relations, and “inconsistent” if they are inconsistent with the trained should-relations.

In Experiment 1, participants completed either a standard or modified version of the RRT (De Houwer et al., 2015). Each RRT consisted of 160 total critical trials. In the standard RRT (see Figure 1), participants use button presses to respond ‘true’ (using the ‘E’ key) and ‘false’ (using the ‘I’ key) to stimuli. Two types of stimuli are presented: synonyms of the words true/false (e.g., correct, right, wrong, incorrect), and sentences relating to Niffites and Luupites (e.g., “Niffites are direct”; “Luupites should be kind”). If participants see a synonym, then they should press the key which corresponds to that word. For instance, if participants saw “true”, then they should respond “E” for true. For the sentence stimuli, participants are required to respond as if certain statements are true/false, which differed across the two critical RRT blocks. For example, in the should-RRT, the first block might
SENSITIVITY TO RELATIONAL INFORMATION

instruct participants to respond as if Niffites should be kind/Luupites should be direct. In this case, if participants see the sentence “Niffites should be kind”, then they would be required to press “E” (i.e., for “true”). The second block would reverse this instruction (i.e., respond as if Niffites should be direct/Luupites should be kind).

| True   | False | True | False |
|--------|-------|------|-------|
|        |       |      |       |
| **CORRECT** |       |      |       |
| **NIFFITES ARE DIRECT** |       |      |       |

Figure 1. Trial screens from the standard RRT. On the left, a trial displaying a synonym of the word “true”. On the right, a trial displaying a sentence relating to Niffites based on an “are” relation. In the latter case, the required response from participants (i.e., true or false) varies between blocks, based on instructions provided at the beginning of the block.

The modified RRT was very similar to the standard RRT but increased the likelihood that participants would attend to the relational information when responding (see Figure 2). Synonyms of the words true/false were instead replaced with synonyms of the relation used within the task (i.e., synonyms of are/are not, or synonyms of should be/should not be). Sentences relating to the social groups were presented in the form “Niffites/Luupites -> direct.kind”, and participants were required to respond ‘are’ or ‘are not’ in the are-RRT, and ‘should be’ or ‘should not be’ in the should-RRT. For both versions of the RRT (standard and modified), there was one RRT that assessed are relations (are-RRT) and one RRT that assessed should relations. Each RRT consisted of two blocks, one of which was consistent with training and one which was inconsistent with training. Participants completed both relational variants (are and should) of either the standard or modified RRT.
SENSITIVITY TO RELATIONAL INFORMATION

| Are         | Are not | Are         | Are not |
|-------------|---------|-------------|---------|
| IS NOT THE CASE | NIFFITES → KIND |

**Figure 2.** Trial screens from the modified RRT (variant assessing the “are” relation). On the left, a trial with a stimulus synonymous with an “are not” relation. On the right, a trial with a stimulus probing the relation between “Niffites” and “kind”. As in the standard RRT, the required response for this latter trial varies between RRT blocks.

In Experiment 2a, participants completed the aIAT (Sartori et al., 2008). One aIAT assessed are-relations, the second aIAT assessed should-relations. Each aIAT consisted of 120 total critical trials. The aIAT required participants to simultaneously categorise normatively true and false sentences as true or false (e.g., “I am completing a psychology experiment”; “I am eating at a downtown restaurant”) and sentences about Niffites as either “Niffites are [should be] kind” or “Niffites are [should be] direct”. Participants also completed a standard IAT (Greenwald et al., 1998) where they categorised stimuli as synonyms of direct/kind, and other stimuli as members of the Niffites/Luupites1. Whereas the RRT presented sentences about both Niffites and Luupites, the standard aIAT presented sentences only about Niffites. We did this because the use of the aIAT typically involves all sentences sharing one common feature. In an aIAT assessing autobiographical knowledge of travelling to a city (e.g., Paris), for example, participants might need to respond based on the categories “I travelled to Paris” (a true event) and “I travelled to Dubai” (a false event; Sartori et al., 2008). Notably, the categories vary in terms of their destination, but the subject of the sentence remains consistent (i.e., “I”). To maintain consistency with how the aIAT is

---

1 Analyses based on the standard IAT can be found in the Supplementary Materials.
SENSITIVITY TO RELATIONAL INFORMATION

typically used, we therefore examined a consistent subject in Experiment 2a (i.e., Niffites; see Figure 3).

| True OR Niffites are direct | False OR Niffites are kind | True OR Niffites are direct | False OR Niffites are kind |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| I AM DOING A PSYCHOLOGY EXPERIMENT | NIFFITES ARE DIRECT |

**Figure 3.** Trial screens from the standard aIAT (variant assessing the “are” relation). On the left, a trial presenting factual information to which participants should respond based on their truth value (i.e., in the presented example, participants should respond “true”). On the right, a trial presenting a statement about Niffites. The mappings of responses required for Niffite stimuli switch between aIAT blocks.

An issue with the aIAT in Experiment 2a was that participants could focus solely on the final word of each sentence stimuli, given that both the subject (i.e., Niffites) and relation (i.e., “are” in one aIAT, “should” in the other) remained constant. If participants avoid reading the entire sentence, they can easily ignore the relational information. In Experiment 2b, we modified the aIAT so that participants would not be able to simply focus on the final word of each sentence (see Figure 4). Specifically, we required participants to categorise sentences about both Niffites and Luupites. Because both the subject (Niffites/Luupites) and the object (direct/kind) of the sentences now varies, participants would now be required to attend to the sentence in more detail. This in principle should increase the likelihood that relational information would influence responding.
Figure 4. Trial screens from the modified aIAT (variant assessing the “are” relation). On the left, a trial presenting factual information. On the right, a trial presenting a statement about Niffites. In this modified aIAT, statements about both Niffites and Luupites were presented, compared to only statements about Luupites in the standard aIAT. The response mappings for Niffite/Luupite responses switch between aIAT blocks.

In Experiment 3, participants completed one of two versions of the Propositional Evaluation Paradigm (PEP; Müller & Rothermund, 2019): either a response-time PEP (RT-PEP) or a mousetracking PEP (MT-PEP; Cummins & De Houwer, 2021). Each PEP consisted of 120 critical trials. In each trial, participants were presented with a sentence relating to Niffites/Luupites and their relationship to direct/kind. After this, they were presented with one of two types of prompts. On catch trials, participants saw the prompt “??true or false??” and were required to truth-evaluate the presented sentence. On probe trials (the to-be-analysed trials), either the word ‘true’ or the word ‘false’ was presented. Participants were required to respond to the identity of this word (e.g., respond true for ‘true’) and to ignore the previous sentence. In the MT-PEP, participants were required to respond by moving the mouse and clicking boxes in the top-right or top-left corners; in the RT-PEP, participants responded by pressing the ‘A’ (true) and ‘L’ (false) computer keys (see Figure 5a and 5b for illustrations of the task). In the MT-PEP, we measured (separately for each relation) the degree of deviation of mouse movements on probe trials where the truth values

---

2 This bears strong resemblance to evaluative priming procedures, but instead of focusing on evaluative content, the task instead focuses on truth value.
SENSITIVITY TO RELATIONAL INFORMATION

of the previously-presented sentence and the probe word were consistent with one another compared to when they were inconsistent with one another. In the RT-PEP, we measured this same discrepancy, but based on response times.

Figure 5a. The sequence of a probe trial within the PEP task. Each word of the sentence was presented sequentially, followed by either the word “true” (as in the Figure) or “false”. Participants were required to respond based on the identity of this probe word and to ignore the truth value of the preceding sentence. Participants in the MT-PEP responded using the mouse. Participants in the RT-PEP responded using computer key presses.
Figure 5b. The sequence of a catch trial within the PEP task. Each word of the sentence was presented sequentially, followed by the phrase “??true or false??”. In these trials, participants were required to respond based on the truth value of the presented sentence.

After completing the implicit measure, participants completed a recall phase, where they answered the same eight questions as during the learning phase. Finally, participants completed a self-exclusion measure where they could report potential reasons for why their data should not be analysed (e.g., due to a disruption).

Data preparation.

To ensure scoring uniformity, we scored all measures using a Probabilistic Index (PI; De Schryver & De Neve, 2019). The PI is less sensitive to outliers than other commonly-used scores, such as the D score or raw RTs (De Schryver et al., 2018). A PI represents the probability that a randomly-selected value (e.g., response time) from one block/trial type will be larger than a randomly selected value from another block/trial type. For example, in the are-RRT, a PI of .75 would suggest that, in 75% of cases, a randomly-selected trial from one block (e.g., the inconsistent block) would have a longer RT than a randomly-selected trial
SENSITIVITY TO RELATIONAL INFORMATION

from the other block (e.g., the consistent block). In each measure, a PI greater than 0.5 indicates responding consistent with training, while a PI less than 0.5 indicates responding inconsistent with training (see Ruscio & Mullen, 2012, for more detail).

We computed PI scores for all measures. For the MT-PEP, we computed the PI based on AUC scores. The AUC refers to the area between the mouse trajectory of a participant and the idealized trajectory (i.e., the shortest distance possible to the response option; Calcagnì et al., 2017). A larger AUC score indicates greater attraction to the alternative response option (analogous to the interpretation of longer RTs as indicated more difficulty in responding). For all other measures, we computed PIs based on RTs.

For RRT and aIAT data, we excluded participants if they failed to meet any of the following criteria on any of the measures: (i) >= 35% of responses < 300ms in any practice block (ii) >= 25% of responses <300ms in any critical block, (iii) >10% of responses <300ms across all critical blocks, (iv) > 50% error rate in any given practice block, (v) >40% error rate across all practice blocks, (vi) >40% error rate in any given critical block, (vii) >30% error rate across all critical blocks, and (viii) >=10% of responses >10000ms in any given critical block (Greenwald et al., 2003; Hussey et al., 2020). We also applied these exclusions to the PEPs where possible (i.e., >= 25% of overall responses <300ms, >30% error rate of overall responses, and >=10% of overall responses >10000ms).

Analytic strategy.

For all confirmatory analyses (except when comparing between variants of measures), we computed both a mixed-effects model, and a separate t-test. All confirmatory mixed-effects models used raw RTs/AUCs as the dependent variable, with trial/block type (consistent vs. inconsistent) and relation type (are vs. should) as predictors (and their interaction). Participant ID was modelled as a random intercept, and trial/block type and
SENSITIVITY TO RELATIONAL INFORMATION

relation type were modelled as random slopes. The Wilkinson notation for these models is denoted as:

\[ \text{RT/AUC} \sim \text{block/trial type} \times \text{relation type} + (1 + \text{block/trial type} + \text{relation type} | \text{participant}) \]

In addition, t-tests were applied to scored data (i.e., comparing PIs across relation types). We report these t-tests for familiarity and utility to readers; however, our preregistered criteria for all experiments use the mixed-effects analyses as our primary decision point.

For all analyses, we firstly compare RTs (or AUCs for the MT-PEP) between consistent and inconsistent blocks/trials for each relation in each procedure. A main effect of block/trial type on RTs/AUCs would imply that, in general, there were differences between the blocks on the implicit measures. We therefore expected to find main effects of block/trial type on all measures (with shorter/smaller RTs/AUCs on learning-consistent blocks/trials). In addition, a significant interaction between block/trial type and relation type would imply that the magnitude of these effects differed as a function of the type of relation assessed. If the measures are equally sensitive to both types of relational information, this would not be the case; therefore, we expected no significant interaction effect. We also expected that there would be no statistically significant difference between PI scores for “are” vs. “should” relations for a similar reason.

Results

Experiment 1 (RRT). In the standard RRT, we found the expected main effect of block on response times, with shorter RTs for consistent blocks than inconsistent blocks, \( \beta = 0.10, 95\% \text{ CI} [0.06, 0.13], p < .001 \). However, counter to our expectations, we found a significant interaction between block and relation type, \( \beta = -0.09, 95\% \text{ CI} [-0.13, -0.05], p < .001 \). For the standard are-RRT, reaction times in the consistent block were significantly shorter than response times in the inconsistent block as expected, \( \beta = 0.08, 95\% \text{ CI} [0.05, 0.11], p < .001 \). However, for the standard should-RRT, we did not find the expected
difference between the blocks, $\beta = 0.01$, 95% CI [-0.02, 0.04], $p = .438$ (see Figure 6). We also found a significant main effect of block on response times in the modified RRT, $\beta = 0.17$, 95% CI [0.13, 0.21], $p < .001$, as well as an unexpected interaction between block and trial type, $\beta = -0.22$, 95% CI [-0.26, -0.18], $p < .001$. In the modified are-RRT, response times in the consistent block were again significantly shorter than in the inconsistent block, $\beta = 0.18$, 95% CI [0.15, 0.21], $p < .001$. In the modified should-RRT, response times in the consistent block were unexpectedly *longer* than in the inconsistent block, $\beta = -0.05$, 95% CI [-0.08, -0.02], $p < .001$ (see Figure 7).

PI scores were unexpectedly larger in the standard are-RRT ($M = .53$, $SD = .06$) compared to the standard should-RRT ($M = .50$, $SD = .07$), $t(113) = 5.04$, $p < .001$, Cohen’s $dz = 0.47$, 95% CI [0.28, 0.67]. This indicates that responding was more in line with training in the standard are-RRT than the standard should-RRT. We also unexpectedly found larger PI scores for the modified are-RRT ($M = .55$, $SD = .07$) than the modified should-RRT ($M = .47$, $SD = .06$), $t(112) = 8.45$, $p < .001$, Cohen’s $dz = 0.79$, 95% CI [0.58, 1.01].

To compare PI scores between the variants, we conduct a linear regression with PI score as the DV, RRT variant (standard vs. modified) as one IV, and relation type (are vs. should) as the other IV. Although we expected that the modified RRT may lead to larger effects overall, we did not find a significant main effect of RRT variant, $\beta = -0.16$, 95% CI [-0.40, 0.08], $p = .184$. This was qualified by a significant variant * relation interaction in predicting PI scores, $\beta = 0.49$, 95% CI [0.15, 0.84], $p = .005$. PIs did not significantly differ between the variants for are-relations, $\beta = -0.18$, 95% CI [-0.44, 0.08], $p = .168$, but differed significantly for should-relations, $\beta = 0.34$, 95% CI [0.08, 0.60], $p = .010$ (unexpectedly, with larger PI scores for the standard should-RRT).
Figure 6. Differences in RTs in the standard RRT between blocks consistent and inconsistent with learning for each relation type. Error bars in this, and all subsequent plots, indicate 95% CIs.
SENSITIVITY TO RELATIONAL INFORMATION

Figure 7. Differences in RTs in the modified RRT between blocks consistent and inconsistent with learning for each relation type.

Experiment 2a (standard aIAT). We found the expected main effect of block on response times across the aIATs, $\beta = 0.05$, 95% CI [0.01, 0.08], $p = .005$. However, like the RRTs, this was qualified by an unexpected interaction between block and relation, $\beta = -0.08$, 95% CI [-0.12, -0.04], $p < .001$. In the are-aIAT, response times in the consistent block were significantly shorter than in the inconsistent block, as expected, $\beta = 0.04$, 95% CI [0.01, 0.07], $p = .006$. In the should-aIAT, however, there was no difference between response times on the two blocks, $\beta = 0.00$, 95% CI [-0.03, 0.02], $p = .769$ (see Figure 8). PI scores were unexpectedly significantly larger for the are-aIAT ($M = .53, SD = .11$) than the should-aIAT ($M = .47, SD = .11$), $t(153) = 4.51, p < .001$, Cohen’s dz = 0.36, 95% CI [0.20, 0.53].

Figure 8. Differences in RTs in the aIAT between blocks consistent and inconsistent with learning for each relation type.
SENSITIVITY TO RELATIONAL INFORMATION

**Experiment 2b (Modified aIAT).** In the modified aIAT, we found the expected main effect of block on response times, $\beta = 0.11$, 95% CI [0.06, 0.15], $p < .001$. Yet again, however, this was qualified by an unexpected interaction effect between block and relation type, $\beta = -0.09$, 95% CI [-0.13, -0.05], $p < .001$. In the are-aIAT, response times in consistent block were again shorter than in the inconsistent block, $\beta = 0.11$, 95% CI [0.09, 0.14], $p < .001$. In the modified should-aIAT, however, there was no significant difference between the two blocks, $\beta = 0.02$, 95% CI [-0.01, 0.04], $p = .276$ (see Figure 9). PI scores for the modified are-aIAT ($M = .54, SD = .10$) were significantly larger than the modified should-aIAT ($M = .51, SD = .08$), $t(152) = 3.78$, $p < .001$, Cohen’s $d_z = 0.31$, 95% CI [0.14, 0.47].

![Effects in the Modified aIAT](chart.png)

**Figure 9.** Differences in RTs in the modified aIAT between blocks consistent and inconsistent with learning for each relation type.

**Experiment 3 (PEP).** For the RT-PEP, we did not find a main effect of trial type on response times, $\beta = 0.03$, 95% CI [-0.01, 0.05], $p = .196$, nor an interaction effect between trial type and relation type, $\beta = -0.02$, 95% CI [-0.06, 0.02], $p = .347$). For are-based
SENSITIVITY TO RELATIONAL INFORMATION

relations, response times for consistent trials were significantly shorter than for inconsistent trials, $\beta = 0.03$, 95% CI [0.00, 0.06], $p = .041$). However, no significant difference between trial types was found on should-based relations ($\beta = 0.01$, 95% CI [-0.02, 0.04], $p = .518$; see Figure 10). For the MT-PEP, we found a significant main effect of trial type, $\beta = 0.50$, 95% CI [0.40, 0.60], $p < .001$. Like the RRTs and aIATs, this was qualified by an unexpected interaction effect, $\beta = -0.06$, 95% CI [-0.10, -0.02], $p = .003$. For are-based relations in the MT-PEP, AUCs were smaller for consistent trials than inconsistent trials ($\beta = 0.50$, 95% CI [0.48, 0.53], $p < .001$. However, unlike the previous measures, for should-based relations AUCs were also smaller for consistent trials compared to inconsistent trials ($\beta = 0.44$, 95% CI [0.42, 0.47], $p < .001$; see Figure 11).

In terms of PI scores, mean PIs in the RT-PEP differed significantly between are relations ($M = .54$, $SD = .09$) and should relations ($M = .50$, $SD = .08$), $t(109) = 3.00$, $p = .003$, Cohen’s $d_z = 0.29$, 95% CI [0.10, 0.48]. By contrast, PIs in the MT-PEP did not differ significantly between are ($M = .64$, $SD = .15$) and should relations ($M = .62$, $SD = .15$), $t(115) = 1.61$, $p = .111$, Cohen’s $d_z = 0.15$, 95% CI [-0.03, 0.33].
Figure 10. Differences in RTs in the RT-PEP between trials consistent and inconsistent with learning for each relation type.

Figure 11. Differences in AUCs in the MT-PEP between trials consistent and inconsistent with learning for each relation type.
In comparing directly between the procedures, we found a significant main effect of variant on PI scores, $\beta = -0.76$, 95% CI [-0.99, -0.56], $p < .001$. PIs were significantly larger in the MT-PEP compared to the RT-PEP for both are-relations ($\beta = -0.78$, 95% CI [-1.02, -0.53], $p < .001$) and should-relations ($\beta = -0.88$, 95% CI [-1.12, -0.64], $p < .001$).

**Exploratory analyses.** In addition to the preregistered analyses detailed above, we also conducted two exploratory analyses. First, we investigated whether error rates (rather than RTs/AUCs) within each procedure exhibited the expected main effect of block/trial consistency, as well as the unexpected interaction between block/trial consistency and type of relation. These results are presented in Table 1.

**Table 1.** Results (standardized beta and p values) of the main effect of trial/block consistency, and the interaction of trial/block consistency and relation type, on the proportion of correct responses within each of the measures. Significant effects are emboldened.

| Measure    | Main effect of consistency | Relation type * consistency interaction |
|------------|----------------------------|-----------------------------------------|
| Standard RRT | $\beta = 0.43,$ $p = .001$ | $\beta = 0.19,$ $p = .301$ |
| Modified RRT | $\beta = 0.54,$ $p < .001$ | $\beta = 0.47,$ $p = .012$ |
| Standard aIAT | $\beta = 0.04,$ $p = .669$ | $\beta = 0.03,$ $p = .872$ |
| Modified aIAT | $\beta = 0.25,$ $p = .029$ | $\beta = 0.14,$ $p = .384$ |
| RT-PEP     | $\beta = 0.40,$ $p = .003$ | $\beta = 0.01,$ $p = .957$ |
We also investigated whether RTs (rather than AUCs) in the MT-PEP exhibited the expected main effect on trial consistency. Similar to the analyses for RTs, we found a significant main effect of trial type, $\beta = 0.14, 95\% \text{ CI} [0.11, 0.16], p < .001$. Like the AUC analyses, RTs for both are- and should-based relations were quicker on consistent trials than inconsistent trials ($\beta = 0.15, 95\% \text{ CI} [0.12, 0.17], p < .001$ and $\beta = 0.11, 95\% \text{ CI} [0.08, 0.13], p < .001$ respectively; see Figure 12).

**Figure 12.** Differences in RTs in the MT-PEP between trials consistent and inconsistent with learning for each relation type.

**General Discussion**

Our results suggest that many implicit measures are not as sensitive to relational information as has previously been considered in the literature. Although all measures produced are-based effects, only the MT-PEP produced effects for both types of relations.
SENSITIVITY TO RELATIONAL INFORMATION

The other measures produced one of two results for should-based relations: either there was no significant effect (the standard RRT, standard and modified aIAT, and RT-PEP) or the effect was opposite to what was expected (modified RRT).

One might argue that our results could be due to the limited amount of training provided. In other words, participants simply did not receive enough exposure to the relations, and this combined with their complexity meant that participants simply did not learn the required information. However, this explanation is insufficient for three reasons. First, as discussed in the introduction, very similar amounts of training can both produce and change effects in non-relational implicit measures. Second, all participants learned effectively enough to recall the information after completing the implicit measure (i.e., those who did not were excluded from analyses). Third, participants demonstrated the expected effects in the MT-PEP, suggesting that relational implicit measures can indeed be sensitive to this information in principle. What features of the MT-PEP, then, make it (relatively) sensitive to relational information? Using the RT-PEP as a close-but-less sensitive comparison can help us narrow down potential explanations. The obvious first difference is that the MT-PEP assessed AUCs rather than RTs. However, this cannot fully explain the MT-PEP’s sensitivity; the MT-PEP also produced learning-consistent effects when using RTs as the dependent variable. The key difference is therefore more likely procedural.

The largest procedural distinction between the measures is the response modality involved; button presses in the RT-PEP compared to mouse movements/clicks in the MT-PEP. These methods handle response deviation in a trial very differently, and this could contribute to the MT-PEP’s effectiveness in two ways. First, it has been demonstrated that on trials where participants ultimately emit a correct response, mouse measures can better quantify response competition than RTs (Schneider et al., 2015). Second, compared to button-pressing, mouse movements provide participants with a greater opportunity to avoid
SENSITIVE TO RELATIONAL INFORMATION

incorrect responses. With button pressing, a single slip of the finger results in an incorrect response. With mouse movements, participants have time to correct an initial movement in the wrong direction. Indeed, 64% of participants in the MT-PEP exhibited errorless responding, compared to only 36% in the RT-PEP (see Supplementary Materials). Given that these tasks only analyze correct responses, mousetracking better captures “slips of mind” which are unlikely to be reflected by RTs (because they will simply produce incorrect, and therefore unanalyzed, responses). Indeed, when using proportion of correct responses as a dependent variable in exploratory analyses, we found expected effects for both are- and should-relations in all the implicit measures (except the modified RRT and standard aIAT; see Supplementary Materials).

One way to harness this information and to improve the relational sensitivity of these measures may be to consider alternative scoring methods which directly incorporate accuracy information. Indeed, despite the relative sensitivity of accuracy here, it may be unwise to use this solely as the dependent variable for these tasks, given that accuracy may suffer from constrained variance which may inhibit the discriminability of these tasks. One promising strategy may be, however, to incorporate information regarding both RTs and accuracy in scoring these tasks. This can be done in several ways, for example with the use of drift-diffusion modelling (Klauer et al., 2007), the modelling of response time as a random effect in mixed-effects modelling (Davidson & Martin, 2013), CAVEAT modelling (Kvam et al., 2022) or multinomial processing tree models (Heck & Erdfelder, 2016). Further investigation would be of use to determine which (if any) of these approaches best tracks with relational information.

The fact that the use of an alternative DV in these measures revealed effects more consistent with the relational information serves as an important reminder of the fact that we cannot speak about the properties of a specific implicit measure in general. Rather, the
SENSITIVITY TO RELATIONAL INFORMATION

decisions we make with these measures – be it in terms of the dependent variable to-be-analyzed, the specific subject matter within the task, or varied features of the task itself – have a critical impact on the inferences we make about the properties of these tasks (cf. Cummins et al., 2022). As such, the focus of the field should not be on identifying which measures are good or bad at picking up relational information; rather, the focus should be on identifying which procedural/quantitative features of measurement procedures best enable us to detect relational information, and to utilize tasks which possess these features, or modify existing tasks to include those features, in instances where investigating such relational information is required.

Importantly, our results help to contextualize many previously puzzling findings in the relational implicit measures literature. Studies which have used the RRT have consistently found that are-relations tend to produce effects in the expected direction, whereas other relations tend to produce null effects (Heider et al., 2018; Dewitte et al., 2017). Our results suggest that these previous findings may well be due to a failure of the measure, rather than the absence of the construct. However, this finding only serves to highlight the need for more valid relational implicit measures. If relational information had no impact on performances in the measures, we would expect that are- and should-based variants of each measure in the current research would demonstrate very similar effects. However, this was not the case. If relational information differentially impacts responding, then it remains important to know which relations are impacting responding, as responses to different relations may predict very different patterns of behavior (e.g., prescriptive vs. descriptive beliefs in predicting gender-based discrimination; Gill, 2004; Heilman, 2012).

Although our experiments here both contextualise previously research findings and highlight critical limitations of relational implicit measures as they are current conceived, our work is not without limitations. Firstly, our experiments focused only on two specific
SENSITIVITY TO RELATIONAL INFORMATION

relations (are and should) and two attributes (kind and direct). Although these relations were chosen precisely because they are the relations which were most examined in previous relational implicit measure studies, it is important to note that the generalizability of these results is somewhat limited. Notably, there has been no previous suggestion in the literature that these measures should be differentially sensitive to other types of relations (e.g., more sensitive to “want” relations than “should” relations), and there appears to be little a priori reason to predict such a difference. However, we did note such differential sensitivity here in our data, with “are” relations exhibiting much more robust effects than “should” relations. Indeed, as one reviewer noted, it may have been the case that participants interpreted “should” relations simply in opposition to “are”-based relations (i.e., reading “should” as meaning “are not”), which may have impacted the dynamics of how this relation exerted control over responding. Future research would certainly benefit from also considering additional relational terms when examining the validity of these relational implicit measures.

Additionally, future research would also benefit from the use of a more elaborate and elongated learning phase. It may well be the case that providing participants with multiple learning sessions could facilitate more fluent learning, which in turn may be reflected in improved relational sensitivity across the measures. However, it is important to reiterate that (i) included participants in these experiments consisted only of those who were able to effectively learn and later recall the trained relations, and (ii) relational sensitivity in the implicit measures was broadly demonstrated upon examining an alternative dependent variable (namely, the proportion of accurate responses).

Importantly, these measures in general face many substantial limitations even if this issue of relational sensitivity can be overcome. There are outstanding questions relating to the nature and meaningfulness of the “implicit” term (Corneille & Hütter, 2020), whether this class of measurement procedure capture a distinct construct to “explicit” measures
SENSITIVITY TO RELATIONAL INFORMATION

(Schimmack, 2021), whether these tasks have any incremental predictive utility beyond other procedures (Cummins et al., 2021), and whether the precision of measurement of these tasks allows for interpretable inferences relating to inter-individual differences (Klein, 2020; Hussey, 2020), to name but a few. It is important that researchers bear all of these limitations in mind when deliberating on whether or not to use these measures, and to qualify any inferences they make based on results from these measures accordingly. The issue of “implicitness” is indeed the most critical in this regard: as mentioned earlier, relational implicit measures are premised on two assumptions: that they are relational and that they are implicit. Our work here has focused on the first assumption, but the issue of the second assumption remains an open question which is critical to address.

It appears that many relational implicit measures are not as ‘relational’ in nature as has previously been assumed. However, our results also demonstrate that relational information in general has an impact on responding in implicit measures, and can be measured effectively (i.e., in the MT-PEP, or when considering alternative DVs in most of the other measures). Most critically, our findings demonstrate that the basic assumptions we make about the procedures we use can easily be mistaken, and the use of using basic experimental control to test these assumptions is critical to ensure the validity of the inferences we make.
SENSITIVITY TO RELATIONAL INFORMATION

Open Practices

All materials, data, processing and analysis scripts, and preregistrations can be found on the Open Science Framework

(https://osf.io/qw6jr/?view_only=cafbeff4bc234fbc930c309819a70229; https://osf.io/4v8tg/?view_only=d2da873eaebe49e0ad1c3085260f99b9; https://osf.io/5kf9z/?view_only=59861bdad4fc4b1ca41a46967e7535ae; and https://osf.io/ydf27/?view_only=ded4644c172040cb8b033690541e2cd0). Deviations from our preregistration are detailed in the Deviations from Preregistration document.

Conflict of Interest

The authors have no conflict of interest to declare.
References

Calanchini, J., Lai, C. K., & Klauer, K. C. (2020). Reducing implicit racial preferences: III. A process-level examination of changes in implicit preferences. *Journal of Personality and Social Psychology*. [https://doi.org/10.1037/pspi0000339](https://doi.org/10.1037/pspi0000339)

Calcagnì, A., Lombardi, L., & Sulpizio, S. (2017). Analyzing spatial data from mouse tracker methodology: An entropic approach. *Behavior Research Methods, 49*(6), 2012–2030. [https://doi.org/10.3758/s13428-016-0839-5](https://doi.org/10.3758/s13428-016-0839-5)

Charlesworth, T. E. S., Kurdi, B., & Banaji, M. R. (2020). Children’s implicit attitude acquisition: Evaluative statements succeed, repeated pairings fail. *Developmental Science, 23*(3), e12911. [https://doi.org/10.1111/desc.12911](https://doi.org/10.1111/desc.12911)

Corneille, O., & Hütter, M. (2020). Implicit? What Do You Mean? A Comprehensive Review of the Delusive Implicitness Construct in Attitude Research. *Personality and Social Psychology Review, 24*(3), 212–232. [https://doi.org/10.1177/1088868320911325](https://doi.org/10.1177/1088868320911325)

Cummins, J., & De Houwer, J. (2019). An inkblot for beliefs: The Truth Misattribution Procedure. *PloS One, 14*(6), e0218661. [https://doi.org/10.1371/journal.pone.0218661](https://doi.org/10.1371/journal.pone.0218661)

Cummins, J., & De Houwer, J. (2021). The shape of belief: Developing a mousetracking-based relational implicit measure. *Social Psychological and Personality Science, 12*(8), 1517-1526. [https://doi.org/10.1177/1948550620978019](https://doi.org/10.1177/1948550620978019)

Cummins, J., Hussey, I., & Spruyt, A. (2022). The role of attitude features in the reliability of IAT scores. *Journal of Experimental Social Psychology*. In press.

Cummins, J., Lindgren, K., & De Houwer, J. (2021). On the role of (implicit) drinking self-identity in alcohol use and problematic drinking: A comparison of five measures. *Psychology of Addictive Behaviors, 35*(4), 458-471. [https://doi.org/10.1037/adb0000643](https://doi.org/10.1037/adb0000643)

Cummins, J., & Roche, B. (2020). Measuring differential nodal distance using the function acquisition speed test. *Behavioural Processes, 178*, 104179. [https://doi.org/10.1016/j.beproc.2020.104179](https://doi.org/10.1016/j.beproc.2020.104179)
SENSITIVITY TO RELATIONAL INFORMATION

Cummins, J., Roche, B., Tyndall, I., & Cartwright, A. (2018). The relationship between differential stimulus relatedness and implicit measure effect sizes. *Journal of the Experimental Analysis of Behavior, 110*(1), 24–38. [https://doi.org/10.1002/jeab.437](https://doi.org/10.1002/jeab.437)

Davidson, D. J., & Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta Psychologica, 144*(1), 83-96.

De Houwer, J., Heider, N., Spruyt, A., Roets, A., & Hughes, S. (2015). The relational responding task: Toward a new implicit measure of beliefs. *Frontiers in Psychology, 6*. [https://doi.org/10.3389/fpsyg.2015.00319](https://doi.org/10.3389/fpsyg.2015.00319)

De Houwer, J., Hermans, D., & Eelen, P. (1998). Affective and Identity Priming with Episodically Associated Stimuli. *Cognition and Emotion, 12*(2), 145–169. [https://doi.org/10.1080/026999398379691](https://doi.org/10.1080/026999398379691)

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

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 61, pp. 127–183). Academic Press. [https://doi.org/10.1016/bs.aesp.2019.09.004](https://doi.org/10.1016/bs.aesp.2019.09.004)

De Schryver, M., & De Neve, J. (2019). A tutorial on probabilistic index models: Regression models for the effect size P(Y1 < Y2). *Psychological Methods, 24*(4), 403–418. [https://doi.org/10.1037/met0000194](https://doi.org/10.1037/met0000194)

De Schryver, M., Hussey, I., De Neve, J., Cartwright, A., & Barnes-Holmes, P. M. D. (2018). The PI IRAP: An alternative scoring algorithm for the IRAP using a probabilistic semiparametric effect size measure. *Journal of Contextual Behavioral Science, 7*, 97–103. [https://doi.org/10.1016/j.jcbs.2018.01.001](https://doi.org/10.1016/j.jcbs.2018.01.001)

Dewitte, M., De Schryver, M., Heider, N., & De Houwer, J. (2017). The actual and ideal sexual self concept in the context of genital pain using implicit and explicit measures. *The Journal of Sexual Medicine, 14*(5), 702-714.
SENSITIVITY TO RELATIONAL INFORMATION

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In Handbook of research methods in social and personality psychology, 2nd ed (pp. 283–310). Cambridge University Press.

Gill, M. J. (2004). When information does not deter stereotyping: Prescriptive stereotyping can foster bias under conditions that deter descriptive stereotyping. Journal of Experimental Social Psychology, 40(5), 619-632.

Glashouwer, K. A., Bennik, E. C., de Jong, P. J., & Spruyt, A. (2018). Implicit measures of actual versus ideal body image: Relations with self-reported body dissatisfaction and dieting behaviors. Cognitive Therapy and Research, 42(5), 622–635. https://doi.org/10.1007/s10608-018-9917-6

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. Psychological Review, 109(1), 3–25. https://doi.org/10.1037/0033-295x.109.1.3

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. Journal of Personality and Social Psychology, 79(6), 1022–1038. https://doi.org/10.1037/0022-3514.79.6.1022

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology, 74(6), 1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. Journal of Personality and Social Psychology, 85(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Grigutsch, L. A., Lewe, G., Rothermund, K., & Koranyi, N. (2019). Implicit ‘wanting’ without implicit ‘liking’: A test of incentive-sensitization-theory in the context of smoking addiction using the wanting-implicit-association-test (W-IAT). Journal of behavior therapy and experimental psychiatry, 64, 9-14.
Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review, 23*(5), 1440-1465.

Heider, N., Spruyt, A., & De Houwer, J. (2018). Body Dissatisfaction Revisited: On the Importance of Implicit Beliefs about Actual and Ideal Body Image. *Psychologica Belgica, 57*(4), 158–173. [https://doi.org/10.5334/pb.362](https://doi.org/10.5334/pb.362)

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior, 32*, 113-135.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). *lab.js: A free, open, online study builder*. PsyArXiv. [https://doi.org/10.31234/osf.io/fqr49](https://doi.org/10.31234/osf.io/fqr49)

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives. *The Psychological Record, 61*(3), 465–496. [https://doi.org/10.1007/BF03395772](https://doi.org/10.1007/BF03395772)

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2019). When People Co-occur With Good or Bad Events: Graded Effects of Relational Qualifiers on Evaluative Conditioning. *Personality and Social Psychology Bulletin, 45*(2), 196–208. [https://doi.org/10.1177/0146167218781340](https://doi.org/10.1177/0146167218781340)

Hussey, I. (2020). The IRAP is not suitable for individual use due to very wide confidence intervals around D scores. *Preprint.*

Hussey, I., Hughes, S., & Nosek, B. A. (2020). Attitudes, Identities and Individual Differences: A large dataset for investigating relations among implicit and explicit attitudes and identity. *Unpublished manuscript.* [https://osf.io/pcjwf](https://osf.io/pcjwf)

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: a diffusion-model analysis. *Journal of Personality and Social Psychology, 93*(3), 353.

Klein, C. (2020). Confidence intervals on implicit association test scores are really rather large. *Preprint.*

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomeszko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit
SENSITIVITY TO RELATIONAL INFORMATION

Association Test and intergroup behavior: A meta-analysis. American Psychologist, 74(5), 569–586. https://doi.org/10.1037/amp0000364

Kvam, P. D., Smith, C., Irving, L. H., & Sokratous, K. (2022). Improving the reliability and validity of the IAT with a dynamic model driven by associations. Preprint.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. Psychological Bulletin, 132(2), 297.

Moran, T., Cummins, J., & De Houwer, J. (2021). Examining automatic stereotyping from a propositional perspective: Is automatic stereotyping sensitive to relational and validity information? Personality and Social Psychology Bulletin. https://osf.io/yq34n/

Müller, F., & Rothermund, K. (2019). The Propositional Evaluation Paradigm: Indirect Assessment of Personal Beliefs and Attitudes. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.02385

Petty, R. E., Fazio, R. H., & Brinol, P. (2008). Attitudes: Insights from the New Implicit Measures. Psychology Press.

Ruscio, J., & Mullen, T. (2012). Confidence Intervals for the Probability of Superiority Effect Size Measure and the Area Under a Receiver Operating Characteristic Curve. Multivariate Behavioral Research, 47(2), 201–223. https://doi.org/10.1080/00273171.2012.658329

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to Accurately Detect Autobiographical Events. Psychological Science, 19(8), 772–780. https://doi.org/10.1111/j.1467-9280.2008.02156.x

Schimmack, U. (2021). The Implicit Association Test: A method in search of a construct. Perspectives on Psychological Science, 16(2), 396–414.

Schneider, I. K., van Harreveld, F., Rotteveel, M., Topolinski, S., van der Pligt, J., Schwarz, N., & Koole, S. L. (2015). The path of ambivalence: Tracing the pull of opposing evaluations using mouse trajectories. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.00996
SENSITIVITY TO RELATIONAL INFORMATION

Tibboel, H., De Houwer, J., & Van Bockstaele, B. (2015). Implicit measures of “wanting” and “liking” in humans. *Neuroscience & Biobehavioral Reviews, 57*, 350–364. [https://doi.org/10.1016/j.neubiorev.2015.09.015](https://doi.org/10.1016/j.neubiorev.2015.09.015)

Ye, Y., & Gawronski, B. (2018). Validating the semantic misattribution procedure as an implicit measure of gender stereotyping. *European Journal of Social Psychology, 48*(3), 348–364. [https://doi.org/10.1002/ejsp.2337](https://doi.org/10.1002/ejsp.2337)
Deviations from preregistration

Alternate labelling of blocks/trial types.

In our preregistrations, we had originally conceptualised a block/trial labelling strategy which would not treat blocks as consistent/inconsistent with training, but rather consistent/inconsistent with specific trained relations. For instance, if a participant learned Niffites are direct and Niffites should be kind, and then completed an are-RRT, the block where participants would need to respond as if “Niffites are direct” would be consistent-with-trained-are-relations, whereas the block where participants would need to respond as if “Niffites are kind” would be consistent-with-trained-should-relations, because Niffites have been coordinated with kind in the context of a should-relation during training. The intention of this labelling strategy was to infer the presence of learning of both relations by the presence of an interaction between block/trial type and relation type. Specifically, we would have expected that RTs would be longer in the consistent-with-should compared to the consistent-with-are block when the relation-type used was “are”, but would expect the opposite pattern when the relation-type used was “should”.

As readers can likely notice from the description above, our original plan for labelling blocks/trial types was relatively difficult to understand for readers unfamiliar with the procedure. As a consequence, we adopted the more conventional approach of simply labelling blocks/trial types as either consistent or inconsistent with learning. However, as a result, our analytic strategies differ somewhat to our preregistration. Specifically, in instances where we initially preregistered a significant interaction effect as the key effect in our analyses, we now would expect to not find a significant interaction effect, coupled with a significant main effect of block/trial type. However, even this is not fully informative, because the presence of an interaction can still indicate that both are- and should-based learning has taken place, but simply to varying degrees. As a result, we also supplement this...
SENSITIVITY TO RELATIONAL INFORMATION

analysis by investigating also two main effects to qualify the specific pattern of any unexpected interaction which is found (i.e., one for are-based relations, and one for should-based relations) for each implicit measure. Notably different analytic approaches do not significantly alter the general conclusions drawn for any of the measures.

We originally preregistered that we would compare the two RRT variants using both a mixed-effects model and a fixed-effects model. However, given the changes above (and the fact that we now expected no two-way interaction between block and relation type), interpreting the preregistered three-way interaction of interest for the mixed-effects model would be exceedingly difficult. We therefore instead only conduct the fixed-effects analysis (which is identical to the analysis conducted between the two PEP variants).

More stringent screening criteria.

Following completion of our second experiment (using the aIAT) we realised that we had preregistered screening criteria which were less stringent than those typically used for both the IAT and the RRT. As such, we changed our approach to the RRT data, instead employing the more stringent processing/exclusion strategy used for the the aIATs to ensure a more rigorous treatment of the data, as well as keeping data processing relatively consistent across studies. Treating the data in this way did not significantly alter the results compared to using the preregistered screening criteria. However, we did need to resample participants, as the more stringent criteria excluded a substantially larger number of participants than were originally excluded using the less stringent criteria.