Cognition 214 (2021) 104792

Contents lists available at ScienceDirect

Cognition

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

Sensitivity of implicit evaluations to accurate and erroneous propositional inferences

Benedek Kurdi *, Yarrow Dunham

Department of Psychology, Yale University, New Haven, CT, United States of America

ARTICLE INFO

Keywords:
Affect Misattribution Procedure
Associative theories
Implicit Association Test
Implicit evaluations
Inferential reasoning
Propositional theories

ABSTRACT

Explicit (directly measured) evaluations are widely assumed to be sensitive to logical structure. However, whether implicit (indirectly measured) evaluations are uniquely sensitive to co-occurrence information or can also reflect logical structure has been a matter of theoretical debate. To test these competing ideas, participants ($N = 3928$) completed a learning phase consisting of a series of two-step trials. In step 1, one or more conditional statements (A → B) containing novel targets co-occurring with valenced adjectives (e.g., "if you see a blue square, Ibbonif is sincere") were presented. In step 2, a disambiguating stimulus, e.g., blue square (A) or gray blob (¬A) was revealed. Co-occurrence information, disambiguating stimuli, or both were varied between conditions to enable investigating the unique and joint effects of each. Across studies, the combination of conditional statements and disambiguating stimuli licensed different normatively accurate inferences. In Study 1, participants were prompted to use *modus ponens* (inferring B from A → B and A). In Studies 2–4, the information did not license accurate inferences, but some participants made inferential errors: affirming the consequent (inferring A from A → B and B; Study 2) or denying the antecedent (inferring ¬B from A → B and ¬A; Studies 3A, 3B, and 4). Bayesian modeling using ordinal constraints on condition means yielded consistent evidence for the sensitivity of both explicit (self-report) and implicit (IAT and AMP) evaluations to the (correctly or erroneously) inferred truth value of propositions. Together, these data suggest that implicit evaluations, similar to their explicit counterparts, can reflect logical structure.

1. Introduction

To obtain rewards and avoid punishments, organisms must maintain accurate representations of their environment. For humans, who live in societies of unprecedented scale and complexity, interacting with social partners carries paramount importance for success and even survival. As such, to adequately navigate the social world, people require up-to-date knowledge of whom they can trust and whom they ought to mistrust. Accordingly, investigating representations of the goodness or badness of social entities, as well as the origins of such representations, has had a pride of place in scientific psychology since the very inception of the field (e.g., Allport, 1935; Eagly & Chaiken, 1993; McGuire, 1985; Wood, 2000).

In the present work we explore how social evaluations are acquired and updated, with special focus on implicit evaluations (Devine, 1989; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Greenwald & Banaji, 1995), i.e., evaluations whose presence is inferred without asking participants to intentionally reflect on the to-be-measured mental content. Under this definition, implicit evaluations differ from their explicit counterparts in features of the measurement context, with the former relying on indirect indices of underlying knowledge (such as response latencies) and the latter on direct indices of underlying knowledge (such as different forms of self-report). Whether implicit and explicit evaluations also differ from each other more deeply, especially in the types of learning and information to which they are sensitive, has been a matter of debate both in psychology (e.g., De Houwer, 2014; DeCoster, Banner, Smith, & Semin, 2006; Gawronski & Bodenhausen, 2006; Hughes, Barnes-Holmes, & De Houwer, 2011; Kruglanski & Gigerenzer, 2011; Kurdi & Dunham, 2020; Mitchell, De Houwer, & Lovibond, 2009; Smith & DeCoster, 2000; Strack & Deutsch, 2004) and philosophy (e.g., Gendler, 2008; Levy, 2014; Madva, 2016; Mandelbaum, 2016).

Specifically, among many other modalities of learning, both explicit and implicit evaluations have been shown to respond to verbal statements about social targets (e.g., Cone & Ferguson, 2015; Kurdi & Banaji, 2017, 2019; Peters & Gawronski, 2011; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2006). For instance, exposure to a statement

* Corresponding author at: Yale University, 2 Hillhouse Ave, New Haven, CT 06511, United States of America.
E-mail address: benedek.kurdi@yale.edu (B. Kurdi).

https://doi.org/10.1016/j.cognition.2021.104792
Received 17 June 2020; Received in revised form 13 May 2021; Accepted 27 May 2021

such as "All Niffians are good and all Laapians are bad" (Kurdi & Banaji, 2017; Study 6C) can result in the updating of both explicit and implicit evaluations of these novel targets in the corresponding directions. However, results of this kind leave open the question of what specific features of verbal material social evaluations are responsive to and whether such features are the same or different depending on whether evaluations are measured explicitly or implicitly. This is the question that we take up in the present work.

On the one hand, verbal statements such as "All Laapians are bad" are characterized by a physical co-occurrence structure. Specifically, a target ("Laapians") occurs close in space and time to a negatively valenced adjective ("bad"). On the other hand, the same statement also has logical structure. In the example above, the speaker asserts something about the target's character (e.g., "Laapians are bad people"). Crucially, the propositional content implied by such statements can participate in inferential reasoning in ways that go beyond mere co-occurrence information.

To elaborate, in the example above, as in many other cases, the co-occurrence information embedded in and the logical structure entailed by language have identical evaluative implications. However, in some cases, such evaluative implications can diverge. For instance, a speaker could say, "If I'm not mistaken, all Laapians are bad." In this case, the co-occurrence information embedded in the statement is the same as above; however, in order to know what the statement logically entails, one must ascertain the truth of the premise (i.e., whether the speaker is or is not mistaken).

In the five experiments reported below we investigate patterns of updating in explicit and implicit evaluations in cases where the evaluative implications of co-occurrence information and of logical structure diverge. Existing theoretical perspectives are aligned in their prediction that, in such cases, explicit evaluations should be sensitive to logical structure above and beyond mere co-occurrence information (De Houwer, 2014; DeCoster et al., 2006; Gendler, 2008; Hughes et al., 2011; Kruglanski & Gigerenzer, 2011; Kurdi & Dunham, 2020; Levy, 2014; Madva, 2016; Mandelbaum, 2016; Mitchell et al., 2009; Smith & DeCoster, 2000; Strack & Deutsch, 2004). Specifically, in the example above, if it turns out that the speaker is mistaken, explicit evaluations of Laapians should remain unchanged assuming that the observer reasons in line with the rules of propositional logic under which A → B and ¬A do not jointly license any accurate inference. If the observer makes a mistake in inferential reasoning and concludes ¬B from A → B and ¬A (an error known as denying the antecedent), then contrary to the valence of the co-occurrence information embedded in the statement, she should update evaluations of Laapians in a positive direction. Crucially, in neither case are explicit evaluations expected to merely reflect co-occurrence information. Rather, they are thought to be modulated by the nature of the inferences that the observer has made.

By contrast, predictions about the updating of implicit evaluations are less straightforward. Under some theoretical perspectives and in light of some empirical evidence, implicit evaluations could be expected to respond solely to the co-occurrence structure of language. That is, according to a co-occurrence hypothesis, in the case described above, implicit evaluations of Laapians should be updated toward negativity simply by virtue of Laapians becoming linked to a negatively valenced description irrespective of the propositional inferences that the observer makes on the basis of that description. Under competing theoretical perspectives and in light of a different body of empirical evidence, implicit evaluations, similar to their explicit counterparts, should reflect logical structure. Specifically, according to an inferential hypothesis, implicit evaluations of Laapians should be updated if the observer makes the (erroneous) propositional inference that Laapians are not bad people and should remain unchanged if she makes the (accurate) propositional inference that the information is inconclusive with regard to the Laapians' character.

1.1. Against the sensitivity of implicit cognition to logical structure

Several dual-process theories of social cognition have endorsed some version of the co-occurrence hypothesis described above (Evans, 2003; Gawronski & Strack, 2004; Lieberman, Gaunt, Gilbert, & Trope, 2002; Sloman, 1996; Smith & DeCoster, 2000; Strack & Deutsch, 2004). That is, these theories posit that implicit evaluations should be uniquely responsive to the co-occurrence structure embedded in language without showing sensitivity to its logical structure.

For instance, Smith and DeCoster (2000) assume the existence of two qualitatively different modes of processing that tap two separate databases representing knowledge in different formats. Whereas the associative mode of processing reflected by implicit evaluations is hypothesized to draw solely on patterns of features built up over time, rule-based processing reflected by explicit evaluations is thought to be subserved by symbolically encoded propositions. Similarly, DeCoster et al. (2006) argue that "[…] the contents of [the implicit] system simply represent what elements have been paired together in the environment and may therefore fail to capture inferences and conclusions deriving from conscious processing of the events" (p. 19).

Similar arguments have also been advanced in philosophy. Perhaps most famously, Gendler (2008) makes a distinction between aliefs and beliefs, with the former being associative and arational and the latter being propositional and rational. Implicit evaluations are thought to be an expression of the former type of mental content and explicit evaluations of the latter. More recently, Madva (2016) has argued that implicit evaluations "[…] seem to be insensitive to the logical form of an agent's thoughts and perceptions" (specifically, operators such as negations and conditionals), and merely reflect "spatiotemporal relations in thought and perception" (p. 2659; including the co-occurrence structure embedded in language).

In line with the co-occurrence hypothesis advanced by these theories, a number of studies have found implicit evaluations to be impervious to the logical structure of language and to encode only co-occurrence information. For instance, DeCoster et al. (2006) administered an impression formation task to participants in which half of the trait words paired with novel targets were negated. At test, implicit evaluations selectively reflected co-occurrence information without encoding the implications of the logical operator: Pairings with positive traits resulted in positive implicit evaluations and pairings with negative traits resulted in negative implicit evaluations. In a different design, Gawronski, Deutsch, Mbirkou, Seibt, and Strack (2008) presented pairings of Black and White faces with stereotype-consistent trait words to participants. Instructing participants to negate such pairings remained ineffective in modulating implicit evaluations.

1.2. In favor of the sensitivity of implicit cognition to logical structure

More recently, an emerging set of theories in social cognition have argued that both implicit and explicit evaluations are subserved by the same basic types of computation. Specifically, over the past decade, De Houwer and colleagues (De Houwer, 2014; De Houwer, Van Dessel, & Moran, 2020; Hughes et al., 2011; Mitchell et al., 2009) have made the case that all social evaluation, including its varieties long assumed to be associative, is more adequately characterized as propositional. Specifically, under these accounts, both explicit and implicit evaluation are thought to emerge from propositional representations that are responsive to the logical form of language, including propositional inferences that one makes from it.

Similar arguments have also been made in philosophy. Most prominently, Mandelbaum (2016) has rejected Gendler's argument that implicit evaluations are associative and arational. Instead, according to Mandelbaum, implicit evaluation emerges from propositional beliefs that are responsive to logical structure. Relatedly, Levy (2014) has questioned the degree to which implicit evaluations encode logical structure; however, he seems to agree with Mandelbaum's

characterization that arational and associative structures are insufficient to account for the pattern of acquisition and updating observed on implicit measures of evaluation.

In line with the inferential hypothesis advanced by these theorists, some empirical investigations have found implicit evaluations to be sensitive to logical structure, although such sensitivity seems to be subject to certain boundary conditions. For instance, in a set of studies by Boucher and Rydell (2012), participants guessed whether positive or negative statements were characteristic of a novel target and received feedback (e.g., "You are incorrect. Bob would not do this."). Implicit evaluations were in line with the propositional implications of the negated feedback but only to the extent that it was made especially visually salient (i.e., presented in large font).

Johnson, Kopp, and Petty (2016) replicated the finding of insensitivity to negation obtained by Gawronski et al. (2008), but also found that a more meaningful negation condition (responding "That's wrong" to pairings of Black faces and negative trait adjectives) was successful in shifting implicit evaluations. Finally, Peters and Gawronski (2011) exposed participants to descriptions of novel targets that were either positive or negative and were subsequently revealed to be either true or false. In line with the inferential hypothesis, both explicit and implicit evaluations reflected the propositional implications of the information rather than merely its co-occurrence structure.

Gast and De Houwer (2012) probed the sensitivity of implicit evaluations to logical structure in a different framework, relying on the idea of second-order conditioning. Specifically, in the first part of the learning phase, participants were exposed to pairings of intrinsically positive images (US_{pos}) with gray squares labeled with the number 1 and intrinsically negative images (US_{neg}) with gray squares labeled with the number 2. Following this intervention, participants learned that during conditioning square no. 1 had covered one neutral image (CS_{pos}) and square no. 2 had covered a different neutral image (CS_{neg}). Under a co-occurrence hypothesis, this kind of learning should not be able to shift implicit evaluations given that it involves symbolic learning about equivalence relationships rather than repeated exposure to stimulus pairings. However, contrary to this hypothesis and in line with an inferential view, implicit evaluations of the neutral images reflected participants' inferences about the covert US–CS pairings, which they had never directly experienced.

1.3. Alternative interpretations of existing evidence

To summarize, empirical evidence on whether implicit measures of social evaluation are sensitive to propositional reasoning is mixed, with some studies providing evidence against such sensitivity (DeCoster et al., 2006; Gawronski et al., 2008) and others providing evidence in favor, even if subject to certain boundary conditions (Boucher & Rydell, 2012; Gast & De Houwer, 2012; Johnson et al., 2016; Peters & Gawronski, 2011). Such mixed evidence would in and of itself constitute sufficient reason to conduct further investigations of this issue.

However, even more importantly, existing studies supporting the sensitivity of implicit evaluation to logical structure are subject to alternative interpretations. Specifically, when it comes to traits with clear opposites, such as the ones used in the studies reviewed above, participants may encode an association between the target and the opposite of the presented trait (Mayo, Schul, & Burnstein, 2004). Crucially for the present purposes, such recoding may well unfold in a way that sidesteps any inferential transitions between mental representations.

For example, upon reading the statement "Laapians are not nice," participants may encode the representation Laapians–Mean in one of two ways involving purely associative structures. First, given that "not nice" and "mean" are synonyms of each other, the representations Not Nice and Mean may co-activate each other in an associative network. Indeed, there is some evidence to suggest that especially commonly used negations, such as "not nice," may be stored as discrete lexical units in long-term memory (Deutsch, Gawronski, & Strack, 2006).

Second, if one allows for inhibitory connections in a conceptual network, then exposure to "not nice" may weaken the connection between the conceptual node for Nice and the conceptual node for Laapians. This type of inhibition may, in turn, strengthen the connection between Laapians and Mean even if the word "mean" had never been used in the stimuli to which the participant was exposed. To the degree that lexical processes, processes of associative inhibition, or a combination of both can account for the effects of negation on implicit evaluation, the results of past studies may not provide clear evidence in favor of the sensitivity of implicit evaluations to the logical structure implied by language.

As recognized by the authors themselves, similar arguments about associative processes can be made about the findings by Gast and De Houwer (2012). Specifically, contrary to the authors' preferred inferential interpretation, the second half of the learning phase in which the numbered gray squares were presented together with the neutral conditioned stimuli may have resulted in one-shot associative learning. Mutatis mutandis, similar arguments apply to a host of other studies that have used relational qualifiers going beyond the negation operator to investigate the sensitivity of implicit evaluations to relational information (for reviews, see De Houwer et al., 2020; Kurdi & Dunham, 2020).

For example, Hu, Gawronski, and Balas (2017) have found standard (assimilative) evaluative conditioning effects on implicit measures when a drug was described as causing the negative symptoms with which it was paired. However, when participants believed that the drugs prevented the same negative symptoms, evaluative conditioning effects, as reflected by implicit measures of evaluation, were reversed. Although the authors interpreted these findings in terms of the effects of propositional reasoning on implicit evaluations, they could be reinterpreted in relatively simple associative terms. Specifically, it is conceivable to account for these results by appealing to the possibility of such information strengthening inhibitory connections between the conceptual nodes corresponding to the drug and to the negative symptoms paired with it in an associative network.

1.4. Overview of the present project

In the present project we investigated the sensitivity of implicit cognition to more complex propositional structures than those implemented in previous work. Across five studies, participants were exposed to conditional statements about novel social targets (e.g., "If you see a purple pentagon, you can conclude that Yimoolap is trustworthy"), each of which was followed by a disambiguating stimulus (e.g., a purple pentagon) that helped participants make propositional inferences about the target. At test, explicit (self-reported) and implicit (indirect) measures of evaluation were administered.

In Study 1, each target co-occurred with both a positive and a negative trait. However, use of *modus ponens* (inferring B from A → B and A) would reveal that one group was characterized exclusively by positive traits and the other group exclusively by negative traits. As such, under the co-occurrence hypothesis, implicit evaluations of both groups should be equal; under the inferential hypothesis, an implicit preference in line with the propositional inference should emerge. Studies 2, 3A, 3B, and 4, in turn, relied on normative errors in propositional reasoning. Specifically, in these studies, participants were exposed to statements that did not license accurate propositional inferences. However, we expected that a subset of participants would commit normative errors in propositional reasoning, specifically affirming the consequent, i.e., inferring A from A → B and B (Study 2) or denying the antecedent, i.e., inferring ¬B from A → B and ¬A (Studies 3A, 3B, and 4).

Crucially, the current studies differ from previous work in some important details, which would make an associative reinterpretation of any potential effects of propositional inference on implicit evaluation considerably more challenging to sustain. Specifically, the colored shapes used as disambiguating stimuli in the present work were completely arbitrary and derived their contextual meaning exclusively

from the conditional statements to which participants were exposed. This feature of the design makes it unlikely that that preexisting lexical representations (e.g., semantic connections between trait adjectives and their antonyms) could account for any influence of inferential reasoning on implicit evaluation.

Second, unlike the negation operator NOT used in prior experiments, it is difficult to see how the conditional operator IF could be represented in a purely associative structure. For example, under an associative account, each conditional statement to which participants in Study 1 are exposed is posited to create (or strengthen) two competing associative representations, e.g., LAAPIANS–GOOD and LAAPIANS–BAD. It is unclear how a mental symbol for the subsequently presented disambiguating stimulus could be attached to these conceptual associations, let alone how the role of the conditional operator, which confers meaning upon the otherwise arbitrary disambiguating stimulus, could be accounted for in associative terms.

If in Studies 2, 3A, 3B, and 4 implicit evaluations were found to be modulated by whether a participant makes errors in propositional reasoning, such a result may be particularly challenging to interpret in a purely associative framework. After all, in these studies, participants were, by design, exposed to identical co-occurrences of novel social stimuli and valenced traits and even to identical disambiguating stimuli. Moreover, the inferential error that some participants in Studies 3A, 3B, and 4 committed involves concluding that the opposite of the trait mentioned in a previously read statement is true (e.g., inferring that Yimoolap is untrustworthy after having read a statement in which the word "Yimoolap" co-occurred with the trait "trustworthy"), thus making an associative interpretation even more dubious.

2. Study 1

Study 1 provided an initial test of the sensitivity of implicit evaluations to logical structure going beyond the simplest possible case of negation used in relevant previous work (Boucher & Rydell, 2012; DeCoster et al., 2006; Gawronski et al., 2008; Johnson et al., 2016; Peters & Gawronski, 2011). Specifically, in a learning phase, participants were exposed to conditional statements about two novel groups in which members of both groups co-occurred equally often with both positive and negative trait adjectives. However, use of *modus ponens* should reveal that one group was characterized only by positive adjectives and the other group only by negative adjectives. At test, explicit (self-reported) and implicit (indirectly measured) evaluations of both groups were assessed. According to all relevant theoretical perspectives, explicit evaluations should reflect the results of propositional reasoning. The predictions of the co-occurrence and inferential hypotheses diverge for implicit evaluations: Under the former, implicit evaluations should reflect only the co-occurrence information embedded in the statements, whereas under the latter, implicit evaluations should be sensitive to their propositional implications.

2.1. Method

2.1.1. Open science practices

All materials, data files, and analysis scripts for this and all remaining studies are available for download from Kurdi and Dunham (2021). We report how we determined the sample size, all data exclusions, all manipulations, and all measures throughout the paper.

2.1.2. Participants and design

Participants were 643 adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu/). Participants were recruited separately for Study 1A ($N = 258$) and Study 1B ($N = 385$). However, the procedures of both studies were nearly identical (see below). Moreover, the study variable (Study 1A vs. Study 1B) did not participate in any significant main effects or interactions with explicit or implicit evaluations as the dependent variable (combined posterior probabilities of the models including the study variable $p = .017$ for explicit evaluations and $p = .010$ for implicit evaluations). As such, we report results collapsed across these two studies.

In line with standard practice, participants who did not complete the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which constituted the focal dependent measure ($n = 21$), as well as participants whose response latencies were below 300 ms on at least 10% of IAT trials ($n = 39$), were excluded from consideration (Greenwald, Nosek, & Banaji, 2003). This resulted in a final sample size of 583. Participants were randomly assigned to an experimental condition or a control condition in a between-participant design (see below).

2.1.3. Materials

Ten names from two novel groups each (the Laapians and the Niffians; Gregg, Seibt, & Banaji, 2006) served as the target stimuli in both the learning and test phases of the experiment (see below). As required for research using the IAT, these stimuli were designed to be easily categorizable. Specifically, Laapian names ended with the syllable *–lap* (e.g., "Neenolap," "Omeelap") and Niffian names ended with the syllable *–nif* (e.g., "Ibbonif," "Yossanif"). Novel social stimuli were selected as targets of learning in this and all remaining studies because novel material provides a relatively pure measure of inferential reasoning uncontaminated by relevant prior knowledge. Moreover, ten clearly positive (e.g., "dependable," "sincere") and ten clearly negative (e.g., "cruel," "malicious") trait adjectives were retrieved from Anderson (1968) for use in both the learning and test phases. Finally, as described below, five color drawings of geometric shapes (a blue square, a gray blob, a green circle, an orange triangle, and a purple pentagon) served as disambiguating stimuli in the learning phase.

2.1.4. Procedure and measures

The study consisted of a learning phase and a test phase. In the learning phase, participants were exposed to a series of two-step trials in the course of which they learned about the two novel target groups (Laapians vs. Niffians) via propositional inference. Crucially, in both the control and experimental conditions, Laapian and Niffian stimuli were paired with positive and negative trait adjectives the same number of times, thus eliminating the possibility that learning could have unfolded as a result of differences in co-occurrence information. In the test phase, participants completed an Implicit Association Test (IAT; Greenwald et al., 1998) measuring implicit evaluations of the two target groups, followed by a set of Likert items measuring explicit evaluations.

2.1.4.1. Learning phase. At the beginning of the learning phase, participants were introduced to the two target groups, read an explanation of the learning task (see below), and were instructed to form a general impression of Laapians and Niffians.

2.1.4.1.1. Experimental condition. 20 pairs of conditional statements were individually generated for each participant and presented in randomized order. Each pair of conditional statements was of the form "If you see [target shape], you can conclude that [target individual] is [target trait]; if you see [alternative shape], you can conclude that [target individual] is [alternative trait of opposing valence]." For instance, a specific trial might read, "If you see a gray blob, you can conclude that Oballnif is obnoxious; if you see a green circle, you can conclude that Oballnif is open-minded." In Study 1A, participants were exposed to all 20 statements, whereas in Study 1B, to shorten the procedure, a randomly selected subset of 12 statements were presented to participants.

Each trial started with a pair of conditional statements displayed above the midpoint of the screen. Upon reading the statements, the participant was able to reveal the shape (disambiguating stimulus) below the statement by hitting the space bar, and the conditional statements and the disambiguating stimulus remained simultaneously

on screen for 3500 ms. Afterward, the program advanced automatically to the next trial. Trials were constructed in such a way that use of *modus ponens* should reveal only positive traits to be characteristic of Niffians and only negative traits to be characteristic of Laapians, although targets from both groups co-occurred with positive and negative traits the same number of times over the course of the learning task.

For each trial, a target individual was randomly selected. Over the course of the learning task, each Laapian and each Niffian stimulus served as a target individual only once. For each trial, a target shape was randomly selected such that each shape was selected four times throughout the task. Alternative shapes were selected randomly from the remaining four shapes. For each trial, a target trait and an alternative trait were randomly selected such that target traits and alternative traits were of the opposite valence. Each positive trait and each negative trait were selected twice, once as a target trait and once as an alternative trait. Whether the target shape and target trait were mentioned in the first clause and the alternative shape and alternative trait in the second clause or vice versa was randomly selected on each trial.

2.1.4.1.2. Control condition. The learning phase in the control condition was procedurally identical to the learning phase in the experimental condition, with the sole exception that five positive and five negative traits were revealed to be characteristic of Laapians and five positive and five negative traits were revealed to be characteristic of Niffians. As such, Laapians and Niffians co-occurred with positive and negative trait adjectives the same number of times across the control and experimental conditions, with the only difference consisting in the implied truth values of Laapian–positive, Laapian–negative, Niffian–-positive, and Niffian–negative propositions revealed to participants via the disambiguating stimuli.

2.1.4.2. Test phase. In the test phase, participants completed an Implicit Association Test (IAT; Greenwald et al., 1998) measuring implicit evaluations of the two target groups (Laapians vs. Niffians). The IAT was followed by a battery of 40 Likert items designed to measure explicit evaluations of the target groups. Given the theoretical focus of the present work on implicit, rather than explicit, evaluations, the IAT was always administered first to be able to obtain a relatively pure measure of implicit evaluations, uncontaminated by having reported explicit evaluations previously.

2.1.4.2.1. Implicit evaluations. Implicit evaluations of Laapian and Niffian targets were measured using a standard five-block Implicit Association Test (IAT; Greenwald et al., 1998). The IAT is a response interference task similar in logic to the Stroop task (Stroop, 1935). Specifically, implicit evaluations are inferred by comparing the speed and accuracy of responding across two sets of combined sorting trials: a first set of sorting trials on which one target group (e.g., Niffians) shares a response key with positive items and the other target group (e.g., Laapians) shares a response key with negative items, and a second set of sorting trials on which the assignment of groups to valences is reversed (e.g., Laapian–positive vs. Niffian–negative).

In block 1 (category practice; 20 trials), participants used two response keys (E and I) to sort the Laapian and Niffian names used as target stimuli in the learning phase. The words "Laapians" and "Niffians" were used as category labels. In block 2 (attribute practice; 20 trials), participants sorted the positive and negative adjectives used as target traits in the learning phase. The words "good" and "bad" were used as attribute labels. In block 3 (congruent combined block; 40 trials), participants used one response key to sort Niffian names and positive trait adjectives and a different response key to sort Laapian names and negative trait adjectives. In block 4 (reversed category practice; 20 trials), participants sorted the same Laapian and Niffian names used in blocks 1 and 3 but with the mapping of categories to response keys reversed. Finally, in block 5 (incongruent combined block; 40 trials), participants used one response key to sort Laapian names and positive trait adjectives and a different response key to sort Niffian names and negative trait adjectives. Given our interest in comparing the control and experimental conditions to each other rather than estimating the absolute magnitude of implicit evaluations, block order was not counterbalanced; rather, the congruent block was always administered first to reduce irrelevant sources of variation (see also Kurdi & Banaji, 2017, 2019).

Performance on the IAT was assessed using the improved scoring algorithm (Greenwald et al., 2003) such that higher D scores index more positive evaluations of Niffians and more negative evaluations of Laapians, in line with the inferential implications of the learning task completed in the experimental condition. Based on 100 split-half samples, the internal consistency of the IAT was found to be satisfactory, $r = 0.73$.

2.1.4.2.2. Explicit evaluations. In line with recent recommendations by Gawronski (2019), the stimuli used to assess implicit and explicit evaluations were identical. Participants completed 40 Likert items, one for each combination of target groups (Laapians vs. Niffians) and trait adjectives previously used in the learning task and on the IAT. Specifically, participants were asked to report to what extent they thought each of the 20 target traits was characteristic of Laapians and Niffians based on what they had learned in the study. Response options ranged from 1 to 7, with 1 labeled "not at all characteristic" and 7 labeled "extremely characteristic." Participants completed all items for Laapians on the same screen and all items for Niffians on the same screen, with the order of the two screens counterbalanced. The order of trait adjectives was individually randomized for each participant and within each target group. Responses for negative items were reverse scored such that higher scores reflect more positive evaluations.

The scales for Laapians and Niffians were highly internally consistent (both Cronbach's αs = 0.98). Therefore, separate composites were created for Laapians and Niffians by calculating the mean of the relevant items. Finally, explicit evaluations of Laapians were subtracted from explicit evaluations of Niffians to create an overall explicit evaluation difference score and thus make explicit and implicit evaluation scores comparable to each other.

2.1.5. Analytic strategy

The utility of frequentist analyses is severely limited in the present setting because such analyses cannot inform about the relative plausibility of different theories in light of the data obtained. This issue is further exacerbated by the inability of traditional frequentist analyses to (*a*) provide evidence in favor of the null hypothesis and (*b*) to combine evidence derived from multiple planned comparisons in a principled manner. As such, for all the main statistical analyses reported below, we adopted a Bayesian model comparison approach.[1]

Specifically, we placed different theoretically derived ordinal constraints on condition means (see Table 1) and used Bayes Factors to compare the plausibility of the co-occurrence and inferential hypotheses relative to the null hypothesis (under which all condition means are equal), relative to the full model (under which condition means are allowed to vary in any theoretically non-specified manner), and, crucially, relative to each other in light of the data (Haaf & Rouder, 2017; Rouder, Haaf, & Aust, 2018). As such, instead of providing information about specific planned contrasts one by one, the Bayes Factors derived from the models below represent a single measure of the relative strength of the evidence for the co-occurrence vs. inferential hypotheses.

A Bayes Factor (BF) of 1 suggests that both hypotheses are equally likely given the data (and the researcher's prior beliefs) and, as such, the experiment is theoretically uninformative. In line with established

[1] Although we see traditional frequentist analyses of the data to have limited utility, we conducted such analyses and made them available in Kurdi and Dunham (2021). Moreover, interested researchers are free to conduct their own analyses using the data files, including trial-level IAT data, which we have also made openly available.

Table 1
Theoretically specified equality and ordinal constraints imposed on patterns of condition means.

Study	Co-occurrence hypothesis	Inferential hypothesis
Study 1	$M_{Experimental} = M_{Control}$	$M_{Experimental} > M_{Control}$
Study 2	$M_{Control} < M_{Experimental\ Accurate} = M_{Experimental\ Error}$	$M_{Control} = M_{Experimental\ Accurate} < M_{Experimental\ Error}$
Study 3A	$M_{Experimental\ Accurate} = M_{Experimental\ Error} < M_{Control}$	$M_{Control} = M_{Experimental\ Accurate} < M_{Experimental\ Error}$
Study 3B	$M_{Experimental\ Accurate} = M_{Experimental\ Error} < M_{Control}$	$M_{Control} = M_{Experimental\ Accurate} < M_{Experimental\ Error}$
Study 4	$M_{Affirm} = M_{Ambiguous} = M_{Deny\ Accurate} = M_{Deny\ Error} < M_{Control}$	$M_{Affirm} < M_{Control} = M_{Ambiguous} = M_{Deny\ Accurate} < M_{Deny\ Error}$

guidelines, we consider $3 > BF > 1$ to provide anecdotal evidence, $10 > BF > 3$ to provide moderate evidence, and $BF > 10$ provide strong evidence for one hypothesis over the other (Lee & Wagenmakers, 2013, p. 105). In some cases, the theoretically possible range of Bayes Factors was restricted. In such cases, the cutoffs mentioned above do not apply, and Bayes Factors should be interpreted in relation to the theoretically possible range reported. Bayes Factors were calculated using the BayesFactor package (Morey, Rouder, & Jamil, 2015). We used the default priors recommended by Rouder, Morey, Speckman, and Province (2012) for one-way ANOVA designs.

2.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 1. Descriptively, both explicit and implicit evaluations seemed to reflect propositional inferences, with a positive shift from the control to the experimental condition.

2.2.1. Explicit evaluations

In the present study, the co-occurrence and null hypotheses make the same prediction; as such, the data cannot distinguish between the two ($BF = 1$). However, as shown in Table 2, the full model was strongly favored by the data relative to the co-occurrence hypothesis ($BF > 10^9$), suggesting that the co-occurrence hypothesis fails to adequately capture the pattern of means revealed by explicit evaluations. In contrast, Bayes Factors revealed strong evidence in favor of the inferential hypothesis compared to both the full model ($BF = 2 \in [0,2]$) and the co-occurrence (null) hypothesis ($BF > 10^9$). As such, this result should increase confidence in the soundness of the experimental design and manipulation.

2.2.2. Implicit evaluations

In line with the inferential hypothesis, the pattern of Bayes Factors was highly similar for implicit evaluations. Specifically, the full model was strongly favored by the data relative to the co-occurrence hypothesis ($BF > 14$), suggesting that the co-occurrence hypothesis fails to capture the pattern of means revealed by implicit evaluations. In contrast, Bayes Factors revealed strong evidence in favor of the inferential hypothesis compared to both the full model ($BF = 1.99 \in [0, 2]$) and, crucially, the co-occurrence (null) hypothesis ($BF > 28$).

2.3. Discussion

In Study 1, members of two novel groups (Niffians and Laapians) were paired with the same number of positive and negative traits; however, only the conditional statements describing Niffians as positive and Laapians as negative were subsequently revealed to be true. As expected under all relevant theoretical accounts, explicit evaluations of the two groups reflected the propositional implications of the information presented during the learning task. Crucially, implicit evaluations were also found to reflect inferential reasoning. As such, this study provides initial evidence in favor of the inferential hypothesis, i.e., the sensitivity of implicit evaluations to logical structure. Notably, the present results may be more challenging to explain in associative terms than previous work: Unlike a single statement involving a negation, it is difficult to see how two conditional statements referring to arbitrary disambiguating stimuli and carrying conflicting evaluative implications could have been recoded into a simple association without the involvement of propositional processes.

3. Study 2

Study 2 was designed to offer a further test of the sensitivity of implicit evaluations to the logical structure of language. Specifically, participants were exposed to statements of the form A → B (e.g., "If Ibbonif is trustworthy, you will see a purple pentagon"). The statements consistently paired Niffians with positive trait adjectives and Laapians with negative trait adjectives. As such, under the co-occurrence hypothesis, implicit evaluations should reveal a preference for Niffians over Laapians.

As in Study 1, conditional statements were followed by disambiguating stimuli; however, in the present study, B was always revealed to be true (e.g., via presentation of a purple pentagon in the example above). This situation licenses no normatively accurate propositional inferences: A (i.e., Ibbonif being trustworthy) and ¬A (i.e., Ibbonif being untrustworthy) are equally compatible with the information presented.[2] Nevertheless, we expected that at least some participants would erroneously conclude that A is true (affirming the consequent). Under the inferential hypothesis, implicit evaluations should reflect the inferences that participants have made from the verbal statements: Participants displaying normatively accurate reasoning should show no change in implicit evaluations, whereas participants displaying normatively erroneous reasoning should show a shift toward implicit preference for Niffians over Laapians.

3.1. Method

3.1.1. Participants and design

Participants were 744 adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu/). In line with standard practice, participants who did not complete the IAT ($n = 21$) and participants whose response latencies were below 300 ms on at least 10% of IAT trials ($n = 8$) were excluded from consideration. Finally, participants who failed to provide a response or provided a nonsensical response on the explicit inference item, indicating inattention (see below; $n = 43$), were also excluded from further analyses. This resulted in a final sample size of 672.

Participants were randomly assigned to one of three conditions with the following probabilities: experimental condition ($p = .66$), active control ($p = .17$), or no intervention control ($p = .17$). In line with our interest in the effects of actual, rather than normatively correct, propositional inferences, twice as many participants were assigned to the experimental condition as to the control conditions to then divide participants from the experimental condition into two groups based on their response to the explicit inference item (see below; experimental

[2] To illustrate using an intuitive example, participants were exposed to premises structurally equivalent to "If this animal is a dog [A], then [→] it has four legs [B]. This animal has four legs [B]." From this it does not follow that the animal is a dog [A]. Clearly, there are many other types of animals that have four legs.

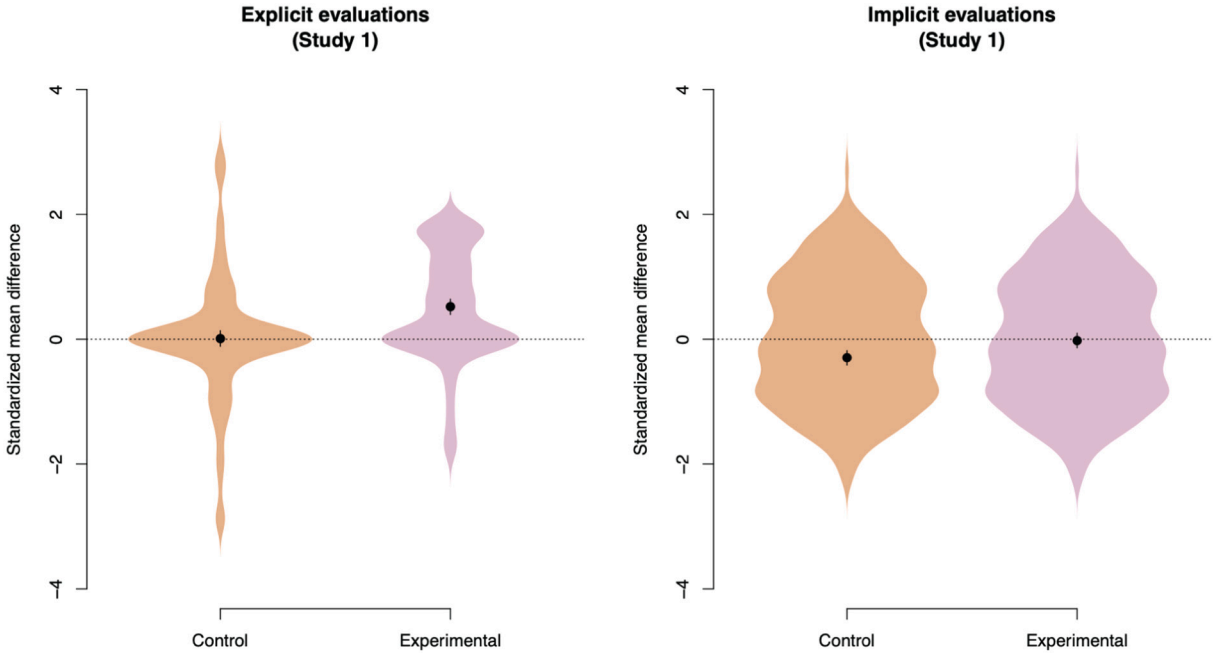

Fig. 1. Distribution of explicit and implicit evaluations by condition (Study 1). Solid dots represent condition means, error bars correspond to 95% highest density intervals (HDIs), and the dashed lines show neutrality. Scores have been standardized to ensure comparability.

Table 2
Relative evidence (Bayes Factor) in favor of the co-occurrence hypothesis vs. null hypothesis, the co-occurrence hypothesis vs. full model, the inferential hypothesis vs. null hypothesis, the inferential hypothesis vs. the full model, and the inferential vs. co-occurrence hypothesis with explicit and implicit attitudes as the dependent measure. BF > 10 is customarily thought to provide strong evidence and 10 > BF > 3 moderate evidence in favor of a hypothesis. 3 > BF > 1/3 is considered to provide anecdotal evidence (Lee & Wagenmakers, 2013, p. 105). In some of the present cases, the theoretical range of Bayes Factors is limited. In such cases, these thresholds do not apply.

Study	Co-occurrence vs. null	Co-occurrence vs. full	Inferential vs. null	Inferential vs. full	Inferential vs. co-occurrence
Explicit evaluations					
Study 1	1	1.21×10^{-9}	1.65×10^{9}	$2 \in [0, 2]$	1.65×10^{9}
Study 2	5.51×10^{4}	$6 \in [0, 6]$	8.12×10^{8}	$6 \in [0, 6]$	1.47×10^{4}
Study 3A	0.02	$2.46 \times 10^{-3} \in [0, 6]$	1.36×10^{4}	$6 \in [0, 6]$	7.98×10^{5}
Study 3B	8.23×10^{-3}	–	5.71×10^{6}	–	6.94×10^{8}
Study 4	0.06	$39.94 \in [0, 120]$	5.86×10^{20}	$119.99 \in [0, 120]$	9.20×10^{21}
Implicit evaluations					
Study 1	1	0.07	28.09	$1.99 \in [0, 2]$	28.09
Study 2	0.73	$5.73 \in [0, 6]$	0.92	$5.81 \in [0, 6]$	1.25
Study 3A	0.12	$4.22 \in [0, 6]$	3.66	$5.96 \in [0, 6]$	30.62
Study 3B	< 0.001	–	799	–	$> 7.99 \times 10^{5}$
Study 4	0.16	$90.34 \in [0, 120]$	4.40×10^{5}	$113.81 \in [0, 120]$	2.73×10^{6}

accurate vs. experimental error groups). Moreover, we expected explicit and implicit evaluations in the two control conditions not to differ from each other. In fact, responding was found to be statistically equivalent on both explicit, $t(229.01) = 1.09$, $BF_{01} = 3.99$, Cohen's $d = 0.14$, and implicit measures, $t(239.44) = 1.48$, $BF_{01} = 2.56$, Cohen's $d = 0.19$. As such, all analyses reported below collapse across the no intervention and active control conditions.

3.1.2. Materials

The materials were identical to the ones used in Study 1. In addition to the materials used in Study 1, ten German first names (e.g., Jörg, Ursula) and ten French first names (e.g., Justine, Antoine) were used as target stimuli in the active control condition.

3.1.3. Procedure and measures

The overall procedure was similar to the one implemented in Study 1, with a learning phase followed by a test phase in which implicit and explicit evaluations were measured. However, some crucial details of the learning task used in the present study differed from the learning task used in Study 1. Moreover, participants assigned to the no intervention control condition proceeded directly to the test phase without completing a learning task.

3.1.3.1. Learning phase

3.1.3.1.1. Experimental condition. Similar to Study 1, participants learned about Laapians and Niffians via propositional inferences made from combinations of conditional statements and disambiguating stimuli. Moreover, the structure of the trials presented in Study 2 was identical to the structure of the trials presented in Study 1.

However, crucially, the types of conditional statements presented to participants were different. Specifically, 20 conditional statements were individually generated for each participant. To shorten the procedure, participants were exposed only to an individually randomized subset of 16 statements, with 8 statements about Laapian targets and 8 statements about Niffian targets. Each conditional statement was of the form "If [target individual] is [target trait], you will see a [target shape]." For

instance, a specific conditional statement might read, "If Maasolap is obnoxious, you will see a blue square." Throughout the task, the conditional statements consistently paired Niffian target individuals with positive traits and Laapian target individuals with negative traits. On each trial, the target shape mentioned in the statement was revealed, thus allowing for an affirming the consequent error to emerge among some participants.

3.1.3.1.2. Active control condition. The learning phase in the active control condition was procedurally identical to the learning phase in the experimental condition, with the sole exception that instead of Niffian and Laapian names, German and French first names served as target stimuli. German names were consistently paired with positive traits and French names with negative traits. We expected this intervention not to influence responding on the implicit and explicit measures of evaluation of the Niffian and Laapian targets. As mentioned above, a comparison of the active control and no intervention control conditions confirmed this prediction.

3.1.3.1.3. No intervention control condition. In the no intervention control condition, participants proceeded directly to the test phase of the experiment.

3.1.3.2. Test phase. The test phase was identical to the test phase in Study 1, with implicit evaluations (internal consistency $r = 0.70$) measured first and explicit evaluations (both Cronbach's αs = 0.97) measured second. In addition to the implicit and explicit evaluation measures, participants also completed an explicit measure of propositional inference. Specifically, participants were asked what could be inferred about Ibbonif from seeing a blue square after being told "If Ibbonif is sincere, you will see a blue square." The three response options were "That Ibbonif is sincere" (erroneous inference), "That Ibbonif is not sincere" (nonsensical response), and "Nothing" (correct inference). Participants in the experimental condition were divided into three groups based on their responses to this measure: (*1*) those who selected the correct inference ($n = 111$; 24%) were included in the experimental accurate group; (*2*) those who selected the erroneous inference ($n = 311$; 67%) were included in the experimental error group; and (*3*) those who selected the nonsensical response or failed to respond to this item ($n = 43$; 9%) were excluded from further analyses.

3.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 2. Descriptively, explicit evaluations seemed to clearly reflect propositional inferences, with similar means in the control and experimental accurate groups and a positive shift in the experimental error group. On the implicit measure, the pattern of means was similar, although with a less pronounced difference between the control and experimental error groups.

3.2.1. Explicit evaluations

The co-occurrence hypothesis provided an adequate description of the pattern of means revealed by explicit evaluations. Specifically, the co-occurrence hypothesis was strongly preferred to both the null hypothesis ($BF > 10^4$) and the full model ($BF = 6 \in [0, 6]$). Importantly, the same was true for the inferential hypothesis, which was also strongly preferred to both the null hypothesis ($BF > 10^8$) and the full model ($BF = 6 \in [0, 6]$). Finally, the crucial comparison provided evidence that the inferential hypothesis was strongly preferred to the co-occurrence hypothesis ($BF > 10^4$), thus increasing confidence in the soundness of the experimental design and manipulation.

3.2.2. Implicit evaluations

For implicit evaluations, the data did not provide convincing evidence in favor of either hypothesis. Specifically, although the co-occurrence hypothesis was preferred to the full model ($BF = 5.73 \in [0, 6]$), the data remained equally consistent with the co-occurrence hypothesis and the null hypothesis ($BF = 1.37$). Similarly, the inferential hypothesis was preferred to the full model ($BF = 5.81 \in [0, 6]$), but the data remained equally consistent with the inferential hypothesis and the null hypothesis ($BF = 1.09$). As such, the present data cannot help conclusively arbitrate between the inferential and co-occurrence hypotheses ($BF = 1.25$).

3.3. Discussion

In Study 2, participants were exposed to conditional statements in which Niffians consistently co-occurred with positive trait adjectives and Laapians with negative trait adjectives. However, the conditional statements licensed no accurate inferences. While some participants correctly recognized this (experimental accurate group), others made inaccurate inferences in line with the co-occurrence information (experimental error group). Whereas explicit evaluations clearly reflected the (accurate and erroneous) propositional inferences made by participants, the data showed only very weak support for the inferential over the co-occurrence hypothesis with implicit evaluations as the dependent measure. The most likely reason for this ambiguity is that the co-occurrence and inferential hypotheses made the same prediction for two out of the three condition means and the one remaining comparison remained inconclusive. As such, we designed Studies 3A, 3B, and 4 in such a way that the predictions implied by the co-occurrence and inferential hypotheses diverged more clearly from each other.

4. Study 3A

Similar to Study 2, Study 3A relied on the idea that differences in implicit evaluation between participants who were exposed to the same information but made different propositional inferences from it would provide strong support for the inferential hypothesis. However, we designed the study in such a way as to make the predictions derived from the co-occurrence and inferential hypotheses more clearly different from each other. Specifically, participants read conditional statements of the form A → B (e.g., "If you see a green circle, you can conclude that Ibbonif is malicious"), with Laapians consistently paired with positive trait adjectives and Niffians with negative trait adjectives. Subsequently, ¬A (e.g., an orange triangle) was presented. This information licenses no normatively accurate propositional inferences; however, we expected that some participants would incorrectly infer ¬B (i.e., Ibbonif not being malicious; denying the antecedent).[3]

Under the co-occurrence hypothesis, all participants in the experimental condition should exhibit a change toward implicit preference of Laapians over Niffians in line with the co-occurrences to which they had been exposed. Under the inferential hypothesis, patterns of change in implicit evaluation should be modulated by participants' propositional inferences: Those making the normatively correct inference should not show updating, whereas those committing the denying the antecedent error should exhibit learning in a direction opposite from that suggested by the pairings.

4.1. Method

4.1.1. Participants and design

Participants were 1142 adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu/). As in previous studies, participants who did not

[3] To illustrate using an intuitive example, participants were exposed to premises structurally equivalent to "If this animal is a dog [A], then [→] it has four legs [B]. This animal is not a dog [¬A]." From this it does not follow that the animal does not have four legs [¬B]. Clearly, there are many other types of animals that also have four legs.

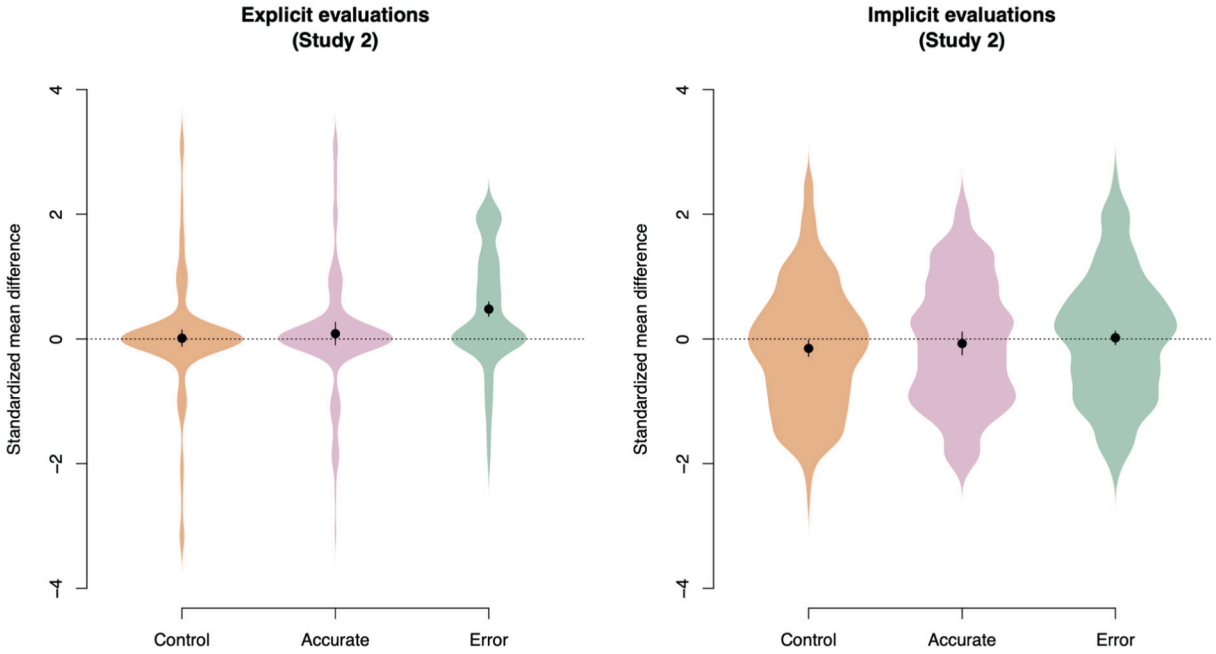

Fig. 2. Distribution of explicit and implicit evaluations by condition and group (Study 2). Solid dots represent condition means, error bars correspond to 95% highest density intervals (HDIs), and the dashed lines show neutrality. Scores have been standardized to ensure comparability.

complete the IAT (n = 22) and participants whose response latencies were below 300 ms on at least 10% of IAT trials (n = 45) were excluded from consideration. Finally, participants who failed to provide a response or provided a nonsensical response on the explicit inference item, indicating inattention (see below; n = 58), were also excluded from further analyses. This resulted in a final sample size of 1017.

Similar to Study 2, participants were randomly assigned to one of three conditions with the following probabilities: experimental condition (p = .66), active control (p = .17), or no intervention control (p = .17). Given our interest in the effects of actual, rather than normatively accurate, inferences, twice as many participants were assigned to the experimental condition as to the control conditions to then divide participants from the experimental condition into two groups based on their response to the explicit inference item (see below; experimental accurate vs. experimental error groups). Moreover, we expected explicit and implicit evaluations in the two control conditions not to differ from each other. In fact, responding was found to be statistically equivalent on both explicit, $t(347.16) = 1.41$, $BF_{01} = 3.27$, Cohen's $d = 0.15$, and implicit measures, $t(359.73) = 0.41$, $BF_{01} = 7.95$, Cohen's $d = 0.04$. As such, all analyses reported below collapse across the no intervention and active control conditions.

4.1.2. Materials and procedure

The materials and procedure were identical to those used in Study 2. Based on 100 split-half samples, the internal consistency of the IAT was found to be acceptable, although lower than in either previous study, r = 0.66. The explicit evaluation scales for Laapians and Niffians were highly internally consistent (Cronbach's αs = 0.96 and 0.97, respectively). Crucially, the conditional statements presented to participants in the experimental condition of the present study had a different structure.

Specifically, in Study 3A, each conditional statement was of the form "If you see [target shape], you can conclude that [target individual] is [target trait]." For instance, a specific conditional statement might read, "If you see a green circle, you can conclude that Ibbonif is malicious." Throughout the task, Niffian target individuals were paired only with negative traits and Laapian target individuals were paired only with positive traits. On each trial, a shape other than the target shape mentioned in the statement was revealed. The revealed shape was selected randomly from the set of four remaining shapes.

In the test phase, participants completed an explicit measure of propositional inference that was slightly different from the one used in Study 2. Specifically, participants were asked what could be inferred about Ibbonif from seeing a gray blob after being told "If you see a blue square, you can conclude that Ibbonif is sincere." The three response options were "That Ibbonif is sincere" (nonsensical response), "That Ibbonif is not sincere" (erroneous inference), and "Nothing" (correct inference). Participants in the experimental condition were divided into three groups based on their responses to this measure: (*1*) those who selected the correct inference (n = 363; 51%) were included in the experimental accurate group; (*2*) those who selected the erroneous inference (n = 290; 41%) were included in the experimental error group; and (*3*) those who selected the nonsensical response or failed to respond to this item (n = 58; 8%) were excluded from further analyses.

4.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 3. Descriptively, explicit evaluations seemed to reflect propositional inferences, with similar means in the control and experimental accurate groups and a positive shift in the experimental error group. On the implicit measure, a similar difference emerged between the control and experimental error groups. At the same time, there appeared to be some indication that the experimental error group may have shifted toward the valence implied by the pairings (as suggested by the co-occurrence hypothesis) rather than remaining constant (as suggested by the inferential hypothesis).

4.2.1. Explicit evaluations

The co-occurrence hypothesis did not provide an adequate description of the pattern of means revealed by explicit evaluations. Specifically, both the null hypothesis (BF > 50) and the full model (BF < 10^{-3}

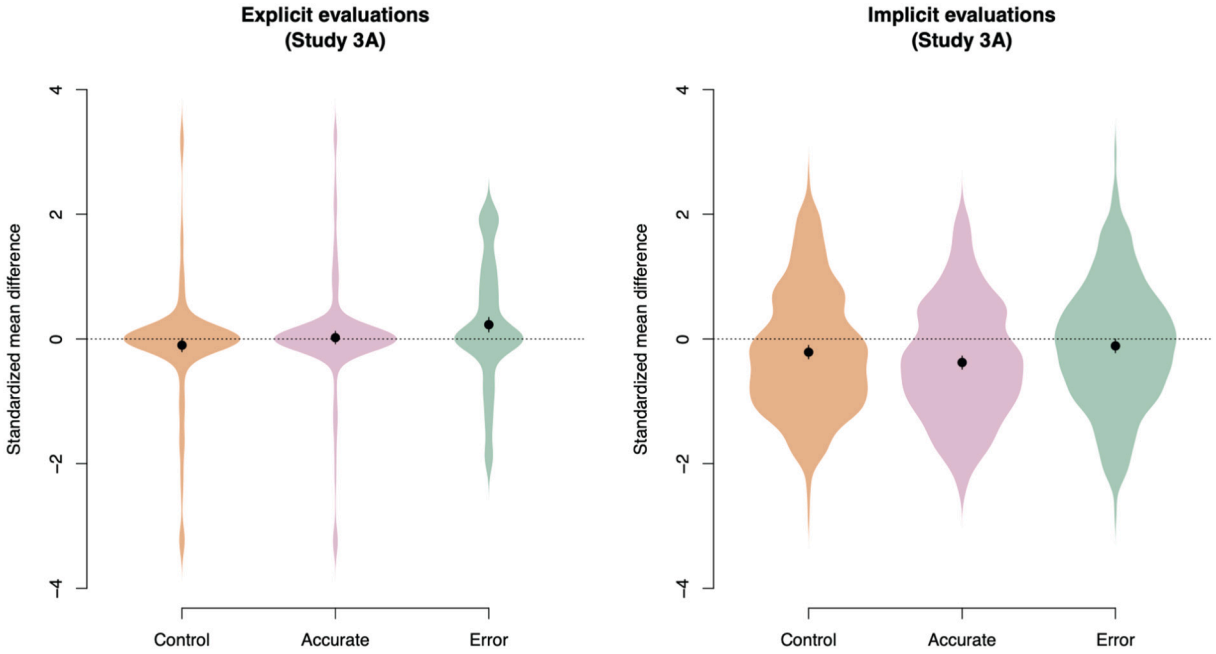

Fig. 3. Distribution of explicit and implicit evaluations by condition and group (Study 3A). Solid dots represent condition means, error bars correspond to 95% highest density intervals (HDIs), and the dashed lines show neutrality. Scores have been standardized to ensure comparability.

$\in$ [0, 6])[4] were strongly preferred to the co-occurrence hypothesis. Importantly, the same was not true for the inferential hypothesis, which was strongly preferred to both the null hypothesis (BF $> 10^4$) and the full model (BF $= 6 \in [0, 6]$). Accordingly, the crucial comparison provided evidence that the inferential hypothesis was strongly preferred to the co-occurrence hypothesis (BF $> 10^5$), thus increasing confidence in the soundness of the experimental design and manipulation.

4.2.2. *Implicit evaluations*

Similarly, the co-occurrence hypothesis did not provide an adequate description of the pattern of means revealed by implicit evaluations. Specifically, the null hypothesis was moderately preferred to the co-occurrence hypothesis (BF > 8), whereas the co-occurrence hypothesis was preferred to the full model (BF $= 4.22 \in [0, 6]$). Importantly, the same ambiguity did not characterize the inferential hypothesis, which was moderately preferred to the null hypothesis (BF > 3) and strongly preferred to the full model (BF $= 5.96 \in [0, 6]$). Accordingly, the crucial comparison provided evidence that the inferential hypothesis was strongly preferred to the co-occurrence hypothesis (BF > 30).

4.3. *Discussion*

In Study 3A, participants were exposed to conditional statements in which Niffians consistently co-occurred with positive trait adjectives and Laapians with negative trait adjectives. However, the conditional statements licensed no accurate inferences. Some participants correctly recognized this (experimental accurate group), while others made inaccurate inferences opposite in direction to the one suggested by the co-occurrence information (experimental error group). In this study, both explicit and implicit evaluations reflected the (accurate and erroneous) propositional inferences made by participants, thus lending credence to the inferential hypothesis. In fact, given the data, the inferential hypothesis was found to be over 30 times more likely to be true than the co-occurrence hypothesis. Remarkably, this pattern of results emerged although all participants in the experimental condition were exposed to the same stimuli, thus making an account of the present data in terms of purely associative processes difficult to defend.

5. Study 3B

The results of Studies 1 and 3A have provided support for the inferential hypothesis, suggesting that implicit evaluations can reflect the logical structure of linguistic input above and beyond mere co-occurrence information. However, both of these studies relied on the IAT as their sole measure of implicit evaluation. Therefore, the results may be specific to this measure (or perhaps a larger set of measures operating on the basis of response competition mechanisms), and as such might not generalize to implicit evaluations more broadly. To address this possible limitation of Studies 1 and 3A and to probe the generalizability of the findings obtained above, in Study 3B we tested whether implicit evaluations measured by the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) are also sensitive to inferential reasoning.

The choice of the AMP as the dependent measure in the present study was guided by both practical and theoretical considerations. At present, the AMP is the second most widely used implicit measure behind the IAT; as such, generalizing the findings obtained in Studies 1 and 3A to this measure may be of inherent practical interest. Moreover, beyond a host of relatively more superficial differences, the AMP and the IAT are thought to be characterized by fundamentally different mechanisms of operation, with the former relying on response competition and the latter on misattribution of affect (e.g., De Houwer & Moors, 2010). Accordingly, the AMP and the IAT have been shown to correlate only modestly with each other (e.g., Bar-Anan & Vianello, 2018), and the two measures do not always respond identically to experimental manipulations (e.g., Van Dessel, Ye, & De Houwer, 2019). As such, if the pattern of results obtained in Study 3A were to replicate using an AMP, we would consider this finding particularly strong evidence for the generalizability of those results across different types of implicit measures and, ultimately, to the theoretical construct of implicit evaluation.

[4] In this case, given that the co-occurrence hypothesis is more specific than the full model, values of the Bayes Factor closer to zero represent stronger evidence in favor of the full model over the co-occurrence hypothesis.

5.1. Method

5.1.1. Open science practices

The hypotheses, design, sample size, and participant exclusions were formally preregistered (https://aspredicted.org/u4ue5.pdf). Any deviations from the preregistration document are explicitly noted below.

5.1.2. Participants and design

Participants were 1003 adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu/). We preregistered a target sample size of 800; however, due to a technical error on the website, study completions were not registered for a few hours, thus resulting in a larger sample. Participants who did not complete the AMP ($n = 47$) and participants who pressed the same key on all AMP trials ($n = 165$), suggesting noncompliance with instructions, were excluded from consideration. Finally, participants who failed to provide a response or provided a nonsensical response on the explicit inference item, indicating inattention (see below; $n = 84$), were also excluded from further analyses. This resulted in a final sample size of 707.

Similar to relevant past research (e.g., Cone & Ferguson, 2015; Mann & Ferguson, 2015, 2017), instead of relying on a separate control condition, along with the experimental prime, we included control primes on the AMP as a within-subjects measure of participants' baseline tendency to evaluate the targets positively. In addition, similar to Study 3A, participants were divided into three groups on the basis of whether their response to the explicit inference item (see below) was correct or erroneous: (*1*) those who selected the correct inference ($n = 397$; 50%) were included in the accurate group; (*2*) those who selected the erroneous inference ($n = 310$; 39%) were included in the error group; and, as mentioned above, (*3*) those who selected the nonsensical response or failed to respond to this item ($n = 84$; 11%) were excluded from further analyses. As such, the design of the study was a mixed 2 × 2 factorial, with accuracy of propositional inference (accurate vs. erroneous) measured between participants and type of AMP prime (control vs. experimental) manipulated within participants.

5.1.3. Materials

We used the same materials as in all previous studies, with three exceptions. First, instead of novel groups, the learning phase of the experiment featured one of six facial images of young White men drawn from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). The AMP featured all six images as primes. Second, we used one of ten male names (including Anthony, Christopher, David, Ethan, Henry, James, Lucas, Michael, Oliver, and Ryan) to refer to the experimental target during the learning phase and on the explicit items. Third, 80 abstract images were borrowed from Katz, Mann, Ferguson, Shen, and Goncalo (2020) for use as targets on the AMP.

5.1.4. Procedure and measures

The procedure was similar to Study 3A: Participants learned about a target via two-step trials consisting of propositional statements and a disambiguating stimulus. Unlike in Study 3A and similar to relevant past research (e.g., Cone & Ferguson, 2015; Mann & Ferguson, 2015, 2017), a single novel individual served as the target of learning. Moreover, crucially, in the test phase we used an AMP instead of an IAT to measure implicit evaluations.

5.1.4.1. Learning phase. For the purposes of the learning phase, participants were not assigned to different conditions; instead, they all underwent a similar learning experience. This learning experience was modeled after the learning phase of Study 3A, with minor changes implemented to make the paradigm more suitable for measuring implicit evaluations via the AMP.

Specifically, instead of names drawn from two novel groups, participants were introduced to a single novel individual and were asked to form an impression of him. Participants were then exposed to 10 learning trials whose structure was the same as that of the learning trials in Study 3A, with the exception that before the participant pressed the space bar to reveal the disambiguating stimulus, the facial image representing the target was displayed below the conditional statement. For each participant, this facial image was randomly drawn from the set of six images described above.

To simplify the procedure, only conditional statements including negative adjectives were used, e.g., "If you see a green circle, you can conclude that Oliver is malicious." Similar to Study 3A, when the participant pressed the space bar, a disambiguating stimulus other than the disambiguating stimulus mentioned in the statement was revealed. As such, based on the rules of propositional logic, participants should have inferred that the conditional statements are uninformative with regard to the target's character. However, participants who committed the denying the antecedent reasoning error would erroneously conclude that the target is characterized by the opposite of the negative adjectives included in the conditional statements, thus resulting in more positive evaluations.

5.1.4.2. Test phase. In the test phase, participants completed an Affect Misattribution Procedure (AMP; Payne et al., 2005) measuring implicit evaluations of the novel target. The AMP was followed by a battery of 20 Likert items designed to measure explicit evaluations of the target as well as by an explicit measure of propositional inference.

5.1.4.2.1. Implicit evaluations. Implicit evaluations were measured using a standard Affect Misattribution Procedure (AMP; Payne et al., 2005). On each trial of the AMP, a prime stimulus was displayed for 75 ms, followed by a blank screen for 125 ms, a target stimulus for 100 ms, and a noise image until the participant entered a response. At the outset of the task, participants were instructed to judge the pleasantness of the target stimulus while resisting any biasing influence of the primes. Participants were asked to press the I key if they believed that the target was more pleasant than average and the E key if they believed that the target was less pleasant than average.

The AMP consisted of 50 trials. On 25 of these trials, the facial image that had served as the target of learning during the learning phase was used as the prime (experimental primes); on the remaining 25 trials, one of the remaining five facial images was used as a prime such that each image appeared five times over the course of the task (control primes). A randomly selected subset of the 80 abstract images described above was used as target stimuli on the AMP such that each image appeared on a single trial. The two types of trial appeared intermixed in an individually randomized order.

5.1.4.2.2. Explicit evaluations and propositional inference. The explicit measures were similar to the explicit measures administered in Study 3A, with the exception that explicit evaluations were collected for only one target given that only one target had been introduced during the learning phase. Internal consistency of the explicit evaluation scale was excellent (Cronbach's $\alpha = 0.96$). The propositional inference item was identical to the one used in Study 3A, with the exception that it used the name "Eric" rather than "Ibbonif."

5.1.5. Analytic strategy

Implicit evaluations were investigated using Bayesian generalized linear mixed-effects models implemented in the rstanarm package (Goodrich, Gabry, Ali, & Brilleman, 2020) in the R statistical computing environment. The binary response variable from the AMP (pleasant vs. unpleasant) served as the dependent variable and random effects included random intercepts for participants, prime images, and target images. In the main effect model (model 1), prime type (control vs. experimental) was the only fixed effect. In the interaction model (model 2), additional fixed effects included accuracy of propositional reasoning (accurate vs. inaccurate) and the Prime Type × Accuracy interaction.

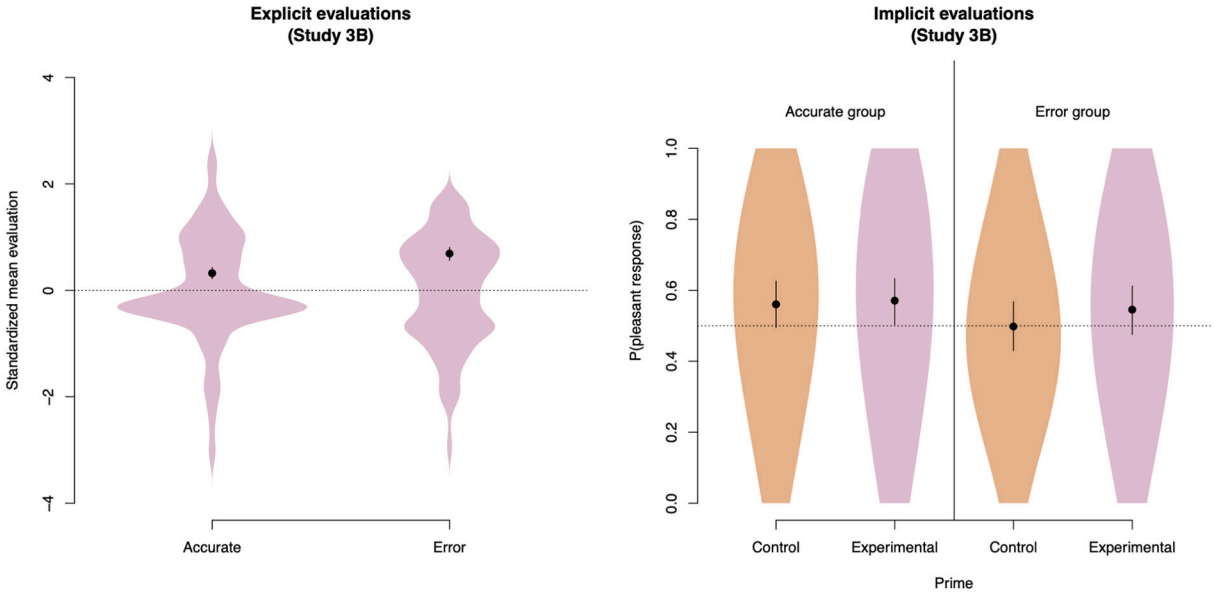

Fig. 4. Distribution of explicit and implicit evaluations by target and group (Study 3B). On the explicit measure, participants evaluated only one target and on the implicit measure, the target along with control faces. Solid dots represent condition means, error bars correspond to 95% highest density intervals (HDIs), and the dashed lines show neutrality. Scores on the two measures are not strictly comparable given that the implicit evaluation measure is bounded at [0; 1].

The co-occurrence hypothesis predicted only a negative main effect of prime type in model 1, corresponding to more negative implicit evaluations of the experimental than of the control primes irrespective of the accuracy of a participant's propositional reasoning. By contrast, the inferential hypothesis predicted a positive interaction effect in model 2, corresponding to more positive implicit evaluations of the experimental than of the control primes in the erroneous reasoning group and no difference between the two types of primes in the accurate reasoning group. Relative support for these two hypotheses was calculated via Bayes Factors obtained using the brms package (Bürkner, 2017).

5.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 4. Descriptively, explicit evaluations seemed to reflect propositional inferences, with the error group expressing more positivity toward the target than the accurate group. On the implicit measure, we obtained the pattern suggested by the inferential hypothesis: Participants in the accurate group evaluated control and experimental primes equally positively, whereas participants in the error group exhibited an implicit preference for experimental over control primes.

5.2.1. Explicit evaluations

Explicit evaluations were considerably more positive in the error than in the accurate group. As such, the inferential hypothesis was preferred both to the null hypothesis (BF $> 10^6$) and to the co-occurrence hypothesis (BF $> 10^8$), thus increasing confidence in the soundness of the experimental design and manipulation.

5.2.2. Implicit evaluations

The Bayesian mixed-effects model yielded a significant and positive interaction effect (BF = 799) such that participants in the accurate group evaluated control and experimental primes equivalently, $b = 0.04$, 95% HDI: [−0.02, 0.11], whereas participants in the error group exhibited an implicit preference for experimental over control primes, $b = 0.19$, 95% HDI: [0.12, 0.26]. Accordingly, the data were found to be incompatible with a negative main effect of prime type irrespective of a participant's propositional reasoning accuracy, BF < 0.001. As such, the present results provided decisive evidence in favor of the inferential over the co-occurrence hypothesis, BF $> 10^5$.

5.3. Discussion

In Study 3B, we replicated the same pattern of results obtained in Study 3A using a different implicit measure of evaluation. Specifically, we found that participants who correctly recognized that a set of conditional statements were uninformative with regard to the character of a novel target showed the same amount of positivity toward this target and control individuals on the AMP. By contrast, participants who erroneously concluded that the novel target was characterized by the opposite of the negative adjectives included in the conditional statements exhibited an implicit preference for the novel target over control individuals. Given fundamental differences in operating conditions between the IAT used as the key dependent measure in Study 3A and the AMP used as the key dependent in the present study, we are confident in concluding that the pattern of learning observed in the present set of experiments characterizes implicit evaluation more broadly and cannot be explained by some incidental feature of the IAT alone.

6. Study 4

Studies 1, 3A, and 3B provided strong evidence in favor of the inferential over the co-occurrence hypothesis, whereas in Study 2 the degree of support remained anecdotal. Notably, in Studies 3A and 3B, implicit evaluations were found to be modulated by individual differences in participants' inferences made from the very same information. Given the novelty of this finding, in Study 4 we replicated the same design and included two additional conditions. In the affirm condition, the disambiguating shapes consistently indicated that the propositional statements were true. As such, theoretical predictions for this condition did not differ. Accordingly, this condition served as an additional check on the robustness of the design and manipulation. More importantly, in a newly added ambiguous condition, participants were merely exposed to conditional statements without any disambiguating stimuli. In this condition, the co-occurrence hypothesis predicted that implicit evaluations should track the co-occurrence information, whereas the inferential hypothesis predicted no updating given that the information did not license any propositional inferences.

6.1. Method

6.1.1. Participants and design

Participants were 1051 adult volunteers from the United States recruited via the Project Implicit educational website (http://implicit.harvard.edu/). As in previous studies, participants who did not complete the IAT ($n = 19$) and participants whose response latencies were below 300 ms on at least 10% of IAT trials ($n = 38$) were excluded from consideration. Finally, participants who failed to provide a response or provided a nonsensical response on the explicit inference item, indicating inattention (see below; $n = 45$), were also excluded from further analyses. This resulted in a final sample size of 949.

Participants were randomly assigned to a no intervention control condition ($p = .2$) or one of three experimental conditions: the affirm condition ($p = .2$), the ambiguous condition ($p = .2$), or the deny condition ($p = .4$). Because the no intervention and active control conditions were found not to differ from each other in Studies 2 and 3A, the active control condition was dropped from the present study. The deny condition was identical to the experimental condition of Study 3A. In the two newly added conditions, the same conditional statements were presented to participants as in the deny condition but were followed by a different set of disambiguating stimuli, thus providing us with further opportunities to probe the effects of propositional inferences beyond mere co-occurrence of targets with positive and negative trait adjectives. Similar to Study 3A, twice as many participants were assigned to the deny condition as to the other conditions to then divide participants from the deny condition into two groups based on their response to the explicit inference item (see below; deny accurate vs. deny error groups).

6.1.2. Materials and procedure

The materials and procedure were similar to those used in Study 3A. Based on 100 split-half samples, the internal consistency of the IAT was found to be satisfactory, $r = 0.72$. The explicit evaluation scales for Laapians and Niffians were highly internally consistent (Cronbach's αs = 0.96 and 0.97, respectively).

The learning phase was similar to the learning phase of Study 3A, with two exceptions. First, at the beginning of each trial, the image of a curtain obscuring the to-be-revealed shape was displayed on the screen. Upon the participant hitting the space bar, the curtain parted in the middle, with the left part moving leftward and the right part moving rightward for 1500 ms. Once the curtains had parted, the revealed shape remained on screen for an additional 2500 ms. We implemented this change to make sure that participants in the ambiguous condition (see below) were able to correctly detect the absence of a disambiguating stimulus.

Similar to Study 3A, Niffian targets were consistently paired with negative traits and Laapian targets were consistently paired with positive traits throughout the learning task. However, depending on assignment to condition, different disambiguating stimuli were revealed. In the affirm condition, the target shape mentioned in the conditional statement was revealed on all trials. In the ambiguous condition, no shape was revealed, i.e., the screen remained blank after the parting of the curtains. In the deny condition, similar to Study 3A, a shape other than the target shape mentioned in the statement was revealed. The revealed shape was selected randomly from the set of four remaining shapes.

In the test phase, unlike in previous studies, the order of Niffian/good–Laapian/bad and Laapian/good–Niffian/bad critical blocks on the IAT was counterbalanced given that different directions of change were theoretically expected in different conditions. Moreover, as in Studies 3A and 3B, a propositional inference measure was administered and used to divide participants in the deny condition into three groups based on their responses: those who selected the correct inference ($n = 220$; 58%) were included in the deny accurate group; (*2*) those who selected the erroneous inference ($n = 116$; 30%) were included in the deny error group; and (*3*) those who selected the nonsensical response or failed to respond to this item ($n = 45$; 12%) were excluded from further analyses.

6.2. Results

The distribution of explicit and implicit evaluations is shown in Fig. 5. Descriptively, both explicit and implicit evaluations seemed to reflect propositional inferences.

6.2.1. Explicit evaluations

The co-occurrence hypothesis did not provide an adequate characterization of the pattern of means revealed by explicit evaluations. Specifically, the co-occurrence hypothesis was outperformed by the null hypothesis (BF > 16) but was preferred to the full model (BF = 39.94 ∈ [0,120]) in accounting for the data. Importantly, the same ambiguity did not characterize the inferential hypothesis, which was strongly preferred to both the null hypothesis (BF > 10^{20}) and the full model (BF = 119.99 ∈ [0, 120]). Accordingly, the crucial comparison provided evidence that the inferential hypothesis was strongly preferred to the co-occurrence

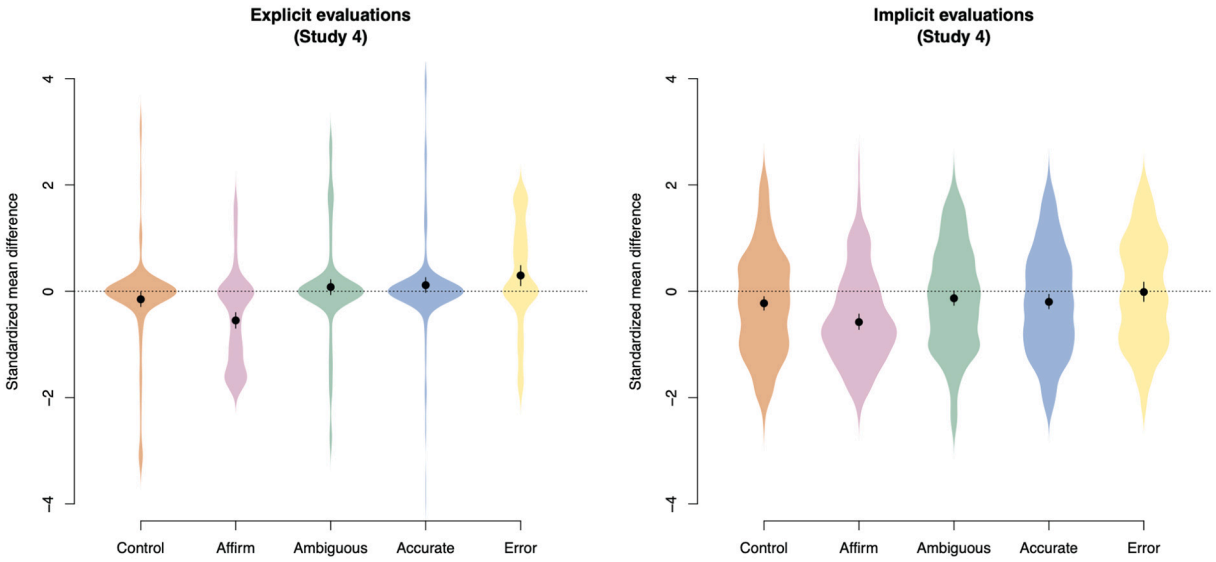

Fig. 5. Distribution of explicit and implicit evaluations by condition and group (Study 4). Solid dots represent condition means, error bars correspond to 95% highest density intervals (HDIs), and the dashed lines show neutrality. Scores have been standardized to ensure comparability.

hypothesis (BF $> 10^{21}$), thus increasing confidence in the soundness of the experimental design and manipulation.

6.2.2. Implicit evaluations

Similarly, the co-occurrence hypothesis did not provide an adequate characterization of the pattern of means revealed by implicit evaluations. Specifically, although the null hypothesis was moderately preferred to the co-occurrence hypothesis (BF > 6), the co-occurrence hypothesis outperformed the full model (BF $= 90.34 \in [0, 120]$). Importantly, the same ambiguity did not characterize the inferential hypothesis, which was strongly preferred both to the null hypothesis (BF $> 10^5$) and the full model (BF $= 113.81 \in [0, 120]$). Accordingly, the crucial comparison provided evidence that the inferential hypothesis was strongly preferred to the co-occurrence hypothesis (BF $> 10^6$).

6.3. Discussion

In Study 4, we replicated and extended the results obtained in Studies 3A and 3B showing that implicit evaluations can differ based on the propositional inferences made by participants even if co-occurrence information or, for some comparisons, both co-occurrence information and disambiguating information, are held constant. In this study, the data produced overwhelming evidence in favor of the inferential hypothesis, which is all the more remarkable given the highly specific pattern of means predicted by this hypothesis. As such, this study provides especially strong evidence that implicit evaluations can be sensitive to the logical structure implied by language as opposed to merely the co-occurrence structure embedded in it.

7. General discussion

In the present project, we conducted five studies to probe whether implicit (indirectly measured) evaluations are sensitive to the propositional inferences made on the basis of verbal statements, or if they merely reflect the co-occurrence structure of such statements. The results emerging from four of the five studies provided strong evidence in favor of the inferential hypothesis: In Study 1, implicit evaluations differed across two conditions that contained identical co-occurrence information but licensed different propositional inferences. Remarkably, in Studies 3A, 3B, and 4, implicit evaluations exhibited differences in response to the very same material depending on whether participants made normatively correct or erroneous propositional inferences from this material. The latter pattern of results, a novelty of the present work, is especially challenging to reconcile with a view under which implicit evaluations merely track co-occurrence information and are immune to high-level reasoning. Finally, Study 2 remained inconclusive but provided some (extremely limited) evidence for the inferential hypothesis. In no study did we obtain any evidence in favor of the co-occurrence hypothesis.

7.1. Theoretical implications

Taken together, the results emerging from the present studies are difficult to reconcile with several dual-process accounts that have dominated theorizing about implicit evaluation both in social cognition research (Evans, 2003; Gawronski & Strack, 2004; Lieberman et al., 2002; Sloman, 1996; Smith & DeCoster, 2000; Strack & Deutsch, 2004) and in philosophy (Gendler, 2008; Madva, 2016). Specifically, these theories posit that implicit evaluations should uniquely reflect the effects of co-occurrence information, without any modulation by high-level inferential reasoning.[5] At the same time, the present findings are clearly in line with the idea that logical structure and inferential reasoning can influence implicit cognition, thus providing evidence in favor of propositional theories (De Houwer, 2014; De Houwer et al., 2020; Hughes et al., 2011; Mandelbaum, 2016; Mitchell et al., 2009).

As noted above, some previous work did not find effects of the negation operator on implicit evaluations (DeCoster et al., 2006; Gawronski et al., 2008), whereas the effects obtained by others (Boucher & Rydell, 2012; Johnson et al., 2016; Peters & Gawronski, 2011) are open to reinterpretation in associative terms. The same limitation does not appear to apply to the present project: In Study 1, participants could not have arrived at the normatively correct inference without (at least temporarily) representing two propositions, which had competing evaluative implications. In Studies 3A, 3B, and 4, participants exhibited different patterns of implicit evaluation depending on what inferences they had made from the very same statements. And, finally, it is difficult to see how the conditional operator If, used in all statements across the present studies, could be used by a purely associative process.

In addition, we see the present findings as complementary to previous work that has established the sensitivity of implicit evaluations to relational information other than negation (for reviews, see De Houwer et al., 2020; Kurdi & Dunham, 2020). Overall, these studies have generally suggested that specifying the relationship between two co-occurring stimuli via relational qualifiers, such as "cause" vs. "prevent" or "like" vs. "dislike," modulates the effects of those stimulus pairings on implicit evaluation. For example, contrary to the evaluative implications of mere co-occurrence information, describing a stimulus as preventing a negative event can result in positive, rather than negative, implicit evaluations of that stimulus (e.g., Hu et al., 2017).

At the same time, we note that the present results also seem to go beyond this set of studies given that, similar to the work relying on negation, most of these studies could be reinterpreted in associative terms. For instance, it may be argued that a statement such as "melatonin prevents sleep issues," or at least its evaluative implications, may be represented by strengthening an inhibitory connection between the conceptual nodes Melatonin and Sleep Issues (and, indirectly, negative valence) in an associative network of concepts. However, the same limitation appears to be less directly applicable to some recent work that has investigated the modulation of the updating of implicit evaluations via gradations of some relationship (e.g., a person being unrelated to, predictive of, or causally responsible for the appearance of the same valenced stimulus; Hughes, Ye, Van Dessel, & De Houwer, 2019) rather than relational qualifiers that are direct opposites of each other in meaning.

Despite the apparent dominance of relational information, some may wish to argue that associative processes could have contributed to the findings obtained in at least some conditions of the present studies. For instance, in Study 1, upon establishing which of the two propositions was valid using the disambiguating stimulus, participants may have mentally rehearsed the accurate statement (e.g., "This means that Yimoolap is open-minded") or a conceptual association derived from it (e.g., Yimoolap–Open-Minded). This way, they may have self-generated additional stimulus pairings to which they were not exposed during the task itself.

[5] To the extent that these theories allow for interactions between explicit and implicit evaluation, they tend to do so under a limited set of conditions involving protracted practice leading to habit formation. Although none of these theories provide a formal definition of protracted practice, it seems safe to assume that the present studies, whose learning phase took no more than five minutes to complete, cannot be adequately characterized as giving rise to habitual responding.

Although the present data are not inconsistent with this possibility, this account begs the question of how participants selected which statement or mental association to rehearse in the first place. Moreover, given that, in the presence of two statements with competing evaluative implications, this selection process must have involved use of the arbitrary disambiguating stimulus in combination with the conditional operator, it is difficult to see how it could be described in purely associative terms (Study 1). Moreover, mental rehearsal accounts would also be hard-pressed to explain other findings from the present project: In conditions where participants were exposed to co-occurrence information embedded in conditional statements but concluded that no valid propositional inference was possible (Studies 3A, 3B, and 4), it seems unlikely that additional rehearsal could have been responsible for the resistance of implicit evaluations to updating.

Finally, some readers may interpret the present results as providing evidence against the discriminant validity of implicit measures of evaluation relative to their explicit counterparts. In response to such potential interpretations we note that claims of construct validity become meaningful only in the context of some substantive theory of the phenomenon under investigation (e.g., Campbell & Fiske, 1959). As such, to the extent that a theorist views relative sensitivity to co-occurrence information vs. relational information as a fundamental difference between implicit and explicit measures of evaluation, then she will no doubt construe the current studies as providing evidence against discriminant validity. However, theories not relying on a dual-process framework tend to characterize the distinction between explicit and implicit evaluation differently, such as in terms of the automaticity of the retrieval of propositional representations (e.g., De Houwer, 2014), the degree to which each has the ability to dynamically and iteratively integrate different sources of evaluative information with each other (e.g., Cunningham, Zelazo, Packer, & Bavel, 2007), or the extent to which each compresses the same underlying evaluative information (e.g., Kurdi & Dunham, 2020). Crucially, the present findings are not in any way incompatible with any of these presumed differences between explicit and implicit evaluation.

7.2. Open questions and future directions

Although we believe that the present studies are theoretically informative, several open questions remain, which we hope will be addressed in future work.

First, the argument presented here should be understood as an existence proof. That is, we think of the present data as convincingly demonstrating the possibility that implicit evaluations can reflect the effects of logical structure. However, the current results do not provide any indication as to how ubiquitous this effect is. Specifically, in designing the present studies, we created conditions that may have made it more likely for the effects of propositional reasoning to emerge: Participants were given the instruction to focus on forming an impression of the groups rather than to track co-occurrences (Moran, Bar-Anan, & Nosek, 2015); learning was intentional rather than incidental; and participants were presumably motivated to encode the information presented during the learning phase.

In fact, existing evidence using word embeddings suggests that co-occurrence information encoded in vast repositories of natural language shows striking similarities with the representations revealed by individual participants on implicit measures of social cognition (Caliskan, Bryson, & Narayanan, 2017; Kurdi, Mann, Charlesworth, & Banaji, 2019). For instance, Caliskan et al. (2017) found that co-occurrence-based measures of semantic distance between categories such as Male and Female and attributes such as Science and Arts calculated from the Common Crawl database of online text were consistent both in direction and size with measures of semantic distance calculated from Implicit Association Tests administered to millions of individual participants. Such convergence raises the possibility that, under conditions different from the ones created in the present studies, implicit evaluations may reflect low-level associative processes without sensitivity to propositional reasoning.

Second, beyond introducing a paradigm in which learning effects are more challenging to attribute to associative processes than in prior work, the present studies are relatively silent regarding the specific nature of the inferential processes unfolding during the learning phase. Notably, classic theories of deductive reasoning (e.g., Evans, 2003; Stanovich & West, 2001) tend to assume that propositional reasoning is a conscious and effortful process. This view is echoed by most theories of implicit evaluation (e.g., Gendler, 2008; Smith & DeCoster, 2000; Strack & Deutsch, 2004).

However, contrary to this view, recent work has produced convincing evidence that propositional inferences can also be produced automatically (for reviews, see De Neys & Pennycook, 2019; Quilty-Dunn & Mandelbaum, 2018). These findings raise the question of whether implicit evaluations are equally capable of reflecting the outputs of relatively effortful and relatively automatic propositional processes, or whether they are primarily sensitive to the latter and not to the former. These competing possibilities should be investigated in future empirical work. The results emerging from such investigations would either be able to considerably constrain propositional accounts of implicit evaluation, or they would provide further evidence for their robustness and scope.

We also see fruitful potential areas of overlap between Bayesian accounts of inferential reasoning (e.g., Oaksford & Chater, 2020) and propositional approaches to implicit evaluation, for multiple reasons. Notably, procedures used in the latter literature often rely on behavioral statements about novel targets to create positive or negative implicit evaluations. For example, Cone and Ferguson (2015) informed participants that an individual called Bob “recently mutilated a small, defenseless animal” and found that implicit evaluations of Bob became markedly negative.

It stands to reason that Bayesian approaches, which have the ability to accommodate uncertainty over the resulting inferences, provide a better normative account of propositional processes giving rise to implicit evaluations in such designs than relatively rigid binary propositional logic. Specifically, given that any directly observable behavior is compatible with an infinite number of latent causes, it is not difficult to see that having mutilated a small, defenseless animal need not formally imply that Bob is a bad person. To name just one example, Bob may have been a veterinarian who saved a puppy's life by amputating an infected limb. The background assumptions and lay theories of person perception entertained by participants, as a result of which most of them are considerably more likely to conclude that Bob is a sadist rather than a vet in this particular case, are easily accommodated by a Bayesian framework.

In addition, Bayesian approaches also have the ability to incorporate the communicative context in which propositional reasoning occurs into formal probabilistic models of the inference process. This feature seems eminently advantageous when it comes to experimental situations in which participants are well aware of the fact that the information provided to them has been selected by the experimenter with the goal of achieving a particular pedagogical goal. For example, in the present paper, we refer to instances of affirming the consequent and denying the antecedent as errors in inferential reasoning and they certainly qualify as such under classic definitions of logical consistency. However, whether inferences of this kind are also irrational or reflect a pragmatically rational use of conversational maxims, such as the maxim of relevance (Grice, 1975), is an open question that could be fruitfully addressed in a Bayesian framework.

Third, in addition to positing that implicit evaluations are sensitive to logical structure, most propositional theories of implicit evaluation make a stronger claim. Specifically, they postulate that implicit evaluation emerges from automatically activated propositional representations (De Houwer, 2014; De Houwer et al., 2020; Hughes et al., 2011; Mandelbaum, 2016; Mitchell et al., 2009). Although the present results

are by no means inconsistent with this possibility, they could also be accounted for by different theoretical frameworks.

For example, the present findings may also be explained under the "common currency" hypothesis recently proposed by Kurdi and Dunham (2020). According to this hypothesis, explicit and implicit evaluations do not differ from each other in terms of their relative sensitivity to inferential reasoning but rather in terms of the degree to which they compress the outputs of such reasoning. For instance, the statement "If you see a green circle, Oballnif is open-minded" followed by a green circle implies the relatively simple conclusion that Oballnif is open-minded (or Oballnif is good). This conclusion, in turn, could ultimately be encoded by updating a highly compressed representation (such as [Oballnif]–[Open-Minded] or even [Oballnif]–[+5]), which does not contain any information about the premises from which the conclusion was originally derived. Thus, future work will be necessary to more directly probe the format of the representations resulting from the propositional inferences that were found to affect responding on implicit measures of evaluation in the present studies.

The present data also appear broadly consistent with hybrid theories of implicit evaluation, including the associative–propositional evaluation (APE) model, which suggests that propositional reasoning may sometimes exert an indirect influence on implicit evaluations (Gawronski & Bodenhausen, 2006). However, it should be pointed out that the APE model accords a central role to associative processes in implicit social cognition and does not anticipate a broad array of cases in which inferential reasoning should dominate over conflicting co-occurrence information in the updating of implicit evaluations. Not precluding the possibility that the increased flexibility of the APE model may be required to account for patterns of data more complex than the ones obtained here, we believe that based on considerations of parsimony and falsifiability, single-process theories should be preferred in explaining the current results. Similar arguments apply to the broader class of "second-generation Klingon" models (De Houwer, 2018) assuming that low-level association formation mechanisms are embedded in and modulated by high-level cognitive processes.

Fourth, it has been suggested that the relative importance of co-occurrence information versus propositional information in the updating of implicit evaluations may be moderated by the temporal order in which the two types of information are provided to participants. Specifically, a recent review by Kurdi and Dunham (2020) has found that implicit evaluations are likely to reflect the effects of relational information when relational information is provided simultaneously with, or at least in close temporal proximity to, co-occurrence information (e.g., Kurdi & Banaji, 2019; Peters & Gawronski, 2011). By contrast, when participants are required to productively recombine multiple pieces of information learned on separate occasions, such recombination tends to affect explicit, but not implicit, evaluations (e.g., Gregg et al., 2006; Mann, Kurdi, & Banaji, 2020).

This distinction seems germane to the present work in at least two ways. On the one hand, it is conceivable that implicit evaluations would have been found less responsive to relational information in the present studies if relational information had not appeared in close temporal proximity to co-occurrence information. On the other hand, this specific type of inflexibility in the updating of implicit (but not explicit) evaluations may also be productively used for theory building. Specifically, this pattern of results is consistent with the possibility that, in line with the "common currency" hypothesis explained above, implicit evaluations may be subserved by compressed summary representations of past experience that are impervious to updating if such updating requires access to high-dimensional details of the original experience. At the same time, without making auxiliary assumptions, such findings seem relatively less consistent with propositional theories and particularly difficult to reconcile with associative accounts of implicit evaluation. As such, we believe that it would be especially important to investigate the robustness of such findings in future work, including in studies using variations of the current paradigm.

Fifth, performance on the Implicit Association Test (IAT) and the Affect Misattribution Procedure (AMP) used in the present studies has been demonstrated to emerge from a combination of relatively more automatic and relatively more controlled processes (Calanchini, Sherman, Klauer, & Lai, 2014; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Payne, Hall, Cameron, & Bishara, 2010). As such, future investigations using a process dissociation framework may provide further clarity on the degree of automaticity with which the outputs of propositional reasoning processes can be activated (for initial evidence on automatic effects of mere instructions in a similar framework, see Hütter & De Houwer, 2017). In addition, although both the IAT and the AMP have been designed to reflect responding under relatively automatic conditions, automaticity is not a unitary construct (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009). As such, we encourage extensions and replications of the present work using further implicit measures of cognition characterized by structural features different from those of the IAT and the AMP.

Sixth, the present studies involve assigning participants to groups on the basis of an individual difference measure of acuity in propositional reasoning. Similar to any correlational design, this approach raises the specter of a third-variable problem, i.e., the possibility that a variable correlated with the propensity for making incorrect inferences may be causally responsible for the present findings. Although we believe that this possibility is unlikely, future studies may use fully randomized assignment to experimental conditions to conclusively eliminate it. Specifically, such work may aim to demonstrate that training participants not to commit inferential errors (e.g., denying the antecedent) can be causally responsible for reducing the effects of these inferential errors on subsequently expressed implicit evaluations.

8. Conclusion

Across five studies, we found evidence for the sensitivity of implicit (indirectly measured) evaluations to the logical structure implied by language above and beyond the co-occurrence information embedded in it. Notably, in three studies, patterns of updating clearly differed as a function of whether participants made normatively accurate or normatively erroneous inferences from the very same information. Such results are difficult to reconcile with any theoretical account assuming that implicit evaluations merely track co-occurrence information experienced in the environment and are impervious to inferential reasoning. By contrast, the present results are compatible with theories that allow for inferential reasoning to influence not only explicit but also implicit evaluations. However, whether similar results would emerge under different learning conditions and, crucially, whether the effects observed here are subserved by genuinely propositional representations remains to be explored in future work.

Declarations of interest

Benedek Kurdi is a member of the Scientific Advisory Committee of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition.

References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*(3), 272–279. https://doi.org/10.1037/h0025907.

Bar-Anan, Y., & Vianello, M. (2018). A multi-method multi-trait test of the dual-attitude perspective. *Journal of Experimental Psychology: General, 147*(8), 1264–1272. https://doi.org/10.1037/xge0000383.

Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during attitude formation. *Personality and Social Psychology Bulletin, 38*(10), 1329–1342. https://doi.org/10.1177/0146167212450464.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. doi:10.18637/jss.v080.i01.
Calanchini, J., Sherman, J. W., Klauer, K. C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin, 40*(10), 1285–1296. https://doi.org/10.1177/0146167214540723.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230.
Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016.
Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37–57. https://doi.org/10.1037/pspa0000014.
Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology, 89*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469.
Cunningham, W. A., Zelazo, P. D., Packer, D. J., & Bavel, J. J. V. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition, 25*(5), 736–760. https://doi.org/10.1521/soco.2007.25.5.736.
De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass, 8*(7), 342–353. https://doi.org/10.1111/spc3.12111.
De Houwer, J. (2018). A functional–cognitive perspective on the relation between conditioning and placebo research. *Neurobiology of the Placebo Effect Part I, 138*, 95–111. Elsevier Inc https://doi.org/10.1016/bs.irn.2018.01.007.
De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski, & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–196). New York, NY: Guilford Press.
De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*(3), 347–368. https://doi.org/10.1037/a0014211.
De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology, 61*, 127–183. Elsevier Inc https://doi.org/10.1016/bs.aesp.2019.09.004.
De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science, 28*(5), 503–509. https://doi.org/10.1177/0963721419855658.
DeCoster, J., Banner, M. J., Smith, E. R., & Semin, G. R. (2006). On the inexplicability of the implicit: Differences in the information provided by implicit and explicit tests. *Social Cognition, 24*(1), 5–21. https://doi.org/10.1521/soco.2006.24.1.5.
Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as reflective operation. *Journal of Personality and Social Psychology, 91*(3), 385–405. https://doi.org/10.1037/0022-3514.91.3.385.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037//0022-3514.56.1.5.
Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt, Brace, & Janovich.
Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012.
Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229.
Gast, A., & De Houwer, J. (2012). Evaluative conditioning without directly experienced pairings of the conditioned and the unconditioned stimuli. *Quarterly Journal of Experimental Psychology, 65*(9), 1657–1674. https://doi.org/10.1080/17470218.2012.665061.
Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science, 14*(4), 574–595. https://doi.org/10.1177/1745691619826015.
Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692.
Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44*(2), 370–377. https://doi.org/10.1016/j.jesp.2006.12.004.
Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40*(4), 535–542. https://doi.org/10.1016/j.jesp.2003.10.005.
Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy, 105*(10), 634–663. https://doi.org/10.5840/jphil20081051025.
Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1. https://mc-stan.org/rstanarm.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037//0033-295X.102.1.4.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464.
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.
Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*(1), 1–20. https://doi.org/10.1037/0022-3514.90.1.1.
Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Vol. 3. Syntax and semantics* (pp. 41–58). New York, NY: Academic Press.
Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods, 22*(4), 779–798. https://doi.org/10.1037/met0000156.
Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. the effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43*(1), 17–32. https://doi.org/10.1177/0146167216673351.
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.
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.
Hütter, M., & De Houwer, J. (2017). Examining the contributions of memory-dependent and memory-independent components to evaluative conditioning via instructions. *Journal of Experimental Social Psychology, 71*(C), 49–58. https://doi.org/10.1016/j.jesp.2017.02.007.
Johnson, I. R., Kopp, B. M., & Petty, R. E. (2016). Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes. *Group Processes & Intergroup Relations, 21*(1), 88–110. https://doi.org/10.1177/1368430216647189.
Katz, J., Mann, T. C., Ferguson, M. J., Shen, X., & Goncalo, J. A. (2020). *Implicit impressions of creative people*. Manuscript submitted for publication.
Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109. https://doi.org/10.1037/a0020762.
Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? *Journal of Experimental Psychology: General, 146*(2), 194–213. https://doi.org/10.1037/xge0000239.
Kurdi, B., & Banaji, M. R. (2019). Attitude change via repeated evaluative pairings versus evaluative statements: Shared and unique features. *Journal of Personality and Social Psychology, 116*(5), 681–703. https://doi.org/10.1037/pspa0000151.
Kurdi, B., & Dunham, Y. (2020). Propositional accounts of implicit evaluation: Taking stock and looking ahead. *Social Cognition, 38*(supplement), s42–s67. https://doi.org/10.1521/soco.2020.38.supp.s42.
Kurdi, B., & Dunham, Y.. *Sensitivity of implicit evaluations to accurate and erroneous propositional inferences*. (2021, May 13) [Dataset] Retrieved from osf.io/cv59s.
Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences, 21*. https://doi.org/10.1073/pnas.1820240116, 201820240–10.
Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.
Levy, N. (2014). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs, 49*(4), 800–823. https://doi.org/10.1111/nous.12074.
Lieberman, M. D., Gaunt, R., Gilbert, D. T., & Trope, Y. (2002). Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. In *, Vol. 34. Advances in Experimental Social Psychology* (pp. 199–249). Elsevier Inc. https://doi.org/10.1016/S0065-2601(02)80006-5.
Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5.
Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese, 193*(8), 2659–2684. https://doi.org/10.1007/s11229-015-0874-2.
Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs, 50*(3), 629–658. https://doi.org/10.1111/nous.12089.
Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823–849. https://doi.org/10.1037/pspa0000021.
Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology, 68*(C), 122–127. https://doi.org/10.1016/j.jesp.2016.06.004.
Mann, T. C., Kurdi, B., & Banaji, M. R. (2020). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General, 149*(6), 1169–1192. https://doi.org/10.1037/xge0000701.
Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs. "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology, 40*(4), 433–449. https://doi.org/10.1016/j.jesp.2003.07.008.
McGuire, W. J. (1985). Attitudes and attitude change. In G. Lindzey, & E. Aronson (Eds.) (3rd ed.,, *Vol. 2. The handbook or social psychology* (pp. 233–346). New York, NY: Random House.
Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*(02), 183–198. https://doi.org/10.1017/S0140525X09000855.
Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co-occurrence on automatic evaluation. *Journal of Experimental Social Psychology, 60*(C), 157–162. https://doi.org/10.1016/j.jesp.2015.05.009.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015, September 19). Package "BayesFactor.". Retrieved from http://bayesfactorpcl.r-forge.r-project.org/.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology, 71*(1), 305–330. https://doi.org/10.1146/annurev-psych-010419-051132.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*(3), 277–293. https://doi.org/10.1037/0022-3514.89.3.277.

Payne, B. K., Hall, D. L., Cameron, C. D., & Bishara, A. J. (2010). A process model of affect misattribution. *Personality and Social Psychology Bulletin, 36*(10), 1397–1408. https://doi.org/10.1177/0146167210383440.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 37*(4), 557–569. https://doi.org/10.1177/0146167211400423.

Quilty-Dunn, J., & Mandelbaum, E. (2018). Inferential transitions. *Australasian Journal of Philosophy, 96*(3), 532–547. https://doi.org/10.1080/00048402.2017.1358754.

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs, 85*(1), 41–56. https://doi.org/10.1080/03637751.2017.1394581.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2006). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37*(5), 867–878. https://doi.org/10.1002/ejsp.393.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01.

Stanovich, K. E., & West, R. F. (2001). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651.

Van Dessel, P., Ye, Y., & De Houwer, J. (2019). Changing deep-rooted implicit evaluation in the blink of an eye: Negative verbal information shifts automatic liking of Gandhi. *Social Psychological and Personality Science, 10*(2), 266–273. https://doi.org/10.1177/1948550617752064.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual Review of Psychology, 51*(1), 539–570. https://doi.org/10.1146/annurev.psych.51.1.539.