Social Cognition, Vol. 38, Supplement, 2020, pp. S42–S67

PROPOSITIONAL ACCOUNTS OF IMPLICIT EVALUATION: TAKING STOCK AND LOOKING AHEAD

Benedek Kurdi
Cornell University

Yarrow Dunham
Yale University

Associative accounts suggest that implicit (indirectly measured) evaluations are sensitive primarily to co-occurrence information (e.g., pairings of gorges with positive experiences) and are represented associatively (e.g., GORGE–NICE). By contrast, recent propositional accounts have argued that implicit evaluations are also responsive to relational information (e.g., gorges causing vs. preventing ennui) and are represented propositionally (e.g., "I find gorges fascinating"). In a review of 30 empirical papers involving exposure to contradictory co-occurrence information and relational information, we found overwhelming evidence for the latter dominating the updating of implicit evaluations, supporting the propositional perspective. However, unlike explicit evaluations, implicit evaluations seem recalcitrant in the face of relational information that requires retrospective revaluation of already encoded co-occurrence information. These findings may be jointly explained by a "common currency" hypothesis under which implicit evaluations emerge from compressed summary representations, which are sensitive to relational information but are not fully propositional.

We thank Bertram Gawronski, Yoav Bar-Anan, and an anonymous reviewer for their insightful comments on previous drafts of this article. The seeds of the ideas explored in this article were first presented at the annual meeting of the Person Memory Interest Group in Severn Bridge, Ontario, Canada, in October 2019.

In addition to the 25th anniversary of the publication of his review article on implicit social cognition with Mahzarin Banaji in *Psychological Review*, this year also marks the occasion of Anthony Greenwald retiring from the University of Washington after a 34-year career there and a total of 57 years spent in psychological research. Given Tony's well-known distaste for any kind of adulation, we refrain from mentioning any of his accomplishments here. But we do hope that he will read this article and tell us why he thoroughly disagrees with our arguments.

Correspondence concerning this article should be addressed to Benedek Kurdi, Department of Psychology, Cornell University, 243 Uris Hall, Ithaca, NY 14850. E-mail: bk493@cornell.edu

Keywords: attitudes, associative theories, dual-process theories, implicit evaluations, propositional theories

In his classic *Handbook* chapter, Allport (1935) described attitudes as "the most distinctive and indispensable construct in contemporary American social psychology" (p. 798). At least to attitude researchers, Allport's words ring just as true today as they did three generations ago. Nonetheless, the landscape of contemporary attitude research would presumably be all but unrecognizable to early attitude theorists, not the least because of fundamental changes in the methods via which attitudes are being measured. Even though reliance on self-report was not traditionally a part of the definition of the attitude construct (Greenwald & Banaji, 1995), for the better part of the 20th century, direct (i.e., self-report) measures remained the most important instrument used to index attitudes (Thurstone, 1928).

Inferring the structure of nonsocial (Meyer & Schvaneveldt, 1971; Neely, 1976) and social (Devine, 1989; Dovidio, Evans, & Tyler, 1986; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Gaertner & McLaughlin, 1983) category knowledge without relying on self-report became possible in the 1970s thanks to the increasingly widespread use of personal computers and the precise recording of response latencies they enabled. In social cognition work, these newly introduced indirect measures revealed that social group stimuli, for instance stimuli related to the category *Black*, facilitated responding to evaluatively or semantically congruent stimuli, such as *death* or *welfare*, and did so even in participants who did not express any negative attitudes or stereotypes toward the category on direct measures.

The year 1995 saw the publication of two highly influential papers on indirectly measured social category knowledge. In their review, Greenwald and Banaji (1995) coined the term *implicit social cognition* and laid the conceptual groundwork for the introduction of the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) a few years later. In their own article, Fazio, Jackson, Dunton, and Williams (1995) provided evidence for the validity of evaluative priming as a measure of racial attitudes.

Since 1995, attitude research using indirect measures has advanced at breakneck speed and has permeated areas of psychology far removed from the relatively narrow original applications, including the study of development (Dunham, Chen, & Banaji, 2013), regional-level phenomena (Hehman, Calanchini, Flake, & Leitner, 2019), and psychopathology (Teachman, Clerkin, Cunningham, Dreyer-Oren, & Werntz, 2019). Remarkably, despite these advances, it remains a matter of vigorous theoretical debate how evaluative knowledge reflected by indirect measures is acquired and updated, how it is represented in the mind, and what processes are involved in its retrieval. These contentious issues are our main focus in this brief review.

PRELIMINARY REMARKS

For the purposes of this review, we define attitudes as "evaluative knowledge" (Fazio, 2007, p. 603). However, unlike Fazio (1995; 2007), we believe that the format

in which evaluative knowledge is represented (e.g., associatively vs. propositionally) should be a matter of empirical research rather than a matter definition. To distinguish latent knowledge structures from their behavioral expressions, we refer to the former using the label "attitudes" and to the latter using the label "evaluations" (e.g., Cunningham & Zelazo, 2007). Moreover, to accommodate the diversity of theoretical perspectives, we define implicit evaluations as the behavioral expression of evaluative knowledge on indirect measures and explicit evaluations as the behavioral expressions of evaluative knowledge on direct measures.[1] As such, we do not presuppose the existence of separate explicit and implicit attitude representations in memory. Moreover, to avoid confusion, we refer to expressions of evaluative knowledge using the distinction of explicit vs. implicit and to the measures used to index such expressions as direct vs. indirect.

In the present article, we review two broad classes of accounts that have been put forth to explain how the knowledge structures from which implicit evaluations emerge are initially acquired and subsequently updated, represented in the mind, and retrieved: associative accounts (e.g., McConnell & Rydell, 2014; Rydell & McConnell, 2006; Smith & DeCoster, 2000; Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000) and propositional accounts (e.g., De Houwer, 2014; 2018; De Houwer & Hughes, 2016; Mandelbaum, 2016; Van Dessel, Hughes, & De Houwer, 2018). We recognize that specific associative theories differ in myriad details from each other, as do their more recent propositional counterparts. However, for our current purposes, we see considerable heuristic value in the associative–propositional distinction given that the two broad classes of accounts make widely divergent claims about a number of central issues. Specifically, they differ in their stance about whether implicit evaluations are sensitive mostly to co-occurrence information or also relational information (especially when the two are in conflict), whether the evaluative knowledge from which implicit evaluations emerge is represented associatively or propositionally, and the types of processes that are involved in the retrieval of such knowledge.

ASSOCIATIVE ACCOUNTS OF IMPLICIT EVALUATION

From the very inception of the field, the construct of implicit evaluation was linked to the idea of associative learning and representation. Initial methodological development in this area relied on existing empirical and theoretical work from experimental cognitive psychology (Collins & Loftus, 1975; Meyer & Schvaneveldt, 1971; Neely, 1976). As the idea of implicit evaluations took hold and indirect measures became more widespread in social cognition research, the field started developing

1. Although we believe that direct and indirect measures differ from each other in terms of automaticity features, we do not make automaticity part of the definition of the implicit vs. explicit construct, for two reasons. First, no measure is process pure (Jacoby, 1991), and, as such, equating *implicit* with *automatic* and *explicit* with *controlled* seems problematic. Second, we acknowledge that different indirect measures create different automaticity conditions (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009), and we believe that such different automaticity conditions should be a matter of empirical research rather than a matter of definition.

its own more specific theories, especially with regard to implicit evaluations and, to a lesser degree, implicit stereotypes.

Associative accounts of implicit evaluation (McConnell & Rydell, 2014; Rydell & McConnell, 2006; Strack & Deutsch, 2004; Wilson et al., 2000) share the core tenet that implicit evaluations are fundamentally different from their explicit counterparts, not merely in terms of the measures used to capture them, but also in how the knowledge structures underlying them are acquired and updated, represented, and retrieved from memory. Specifically, inherited from the historical precursors of implicit social cognition research, including spreading activation models of memory (e.g., Collins & Loftus, 1975) and the priming paradigms used to test them (Meyer & Schvaneveldt, 1971; Neely, 1976), these accounts posit that the knowledge from which implicit evaluations emerge is associative in at least three senses of the word, namely in its acquisition and updating, its representation, and its retrieval.

First, associative accounts postulate that implicit evaluations are associative in the sense that they are primarily sensitive to co-occurrence information. That is, implicit evaluations are thought to track what usually goes together with what in the environment. Such statistical relationships[2] include both innocuous ones, such as bread co-occurring with butter or thunder with lightning, and ones with far-reaching societal implications, such as co-occurrences of gender and science (Nosek et al., 2009) or race and weapons (Correll et al., 2007).

Because co-occurrence information is best computed over longer periods of time, associative accounts predict that implicit evaluations will be especially responsive to learning that is based on repeated experience. For instance, decongestants, coughing, and sneezing often co-occur in everyday life. As such, to the degree that coughing and sneezing are unpleasant, associative accounts posit that implicit evaluations of decongestants should also become negative over time. Crucially, implicit evaluations are not thought to reflect the way in which decongestants are related to coughing, specifically the fact that taking decongestants can alleviate coughing (which is a pleasant outcome) or that coughing does not alleviate decongestants (that is, that the relationship is directional).

To the extent that language is involved in the acquisition and updating of the knowledge underlying implicit evaluations, under associative theories, learning is hypothesized to respond exclusively to the co-occurrence structure of language (e.g., group labels such as *women* often appearing close to attribute labels such as *wealthy*) but not to the relational information embedded in it. For example, implicit evaluations are posited to be insensitive to differences between statements describing actual versus hypothetical states of the world as long as those statements contain the same co-occurrences (such as "*women* are *wealthy*" vs. "I wish *women* were *wealthy*").

2. The external world as such does not contain co-occurrence information; rather, it is more appropriate to say that, under associative accounts, the cognitive processes underlying implicit evaluation chunk incoming information into classes or kinds, and then analyze the relationships between those classes or kinds in terms of co-occurrence structure.

Second, relatedly, associative accounts suggest that implicit evaluations emerge from associative mental representations that reflect only the fact that two concepts are related to each other and the degree of their relatedness (e.g., Bread–Butter, Women–Wealth, Robitussin–Sneezing). Such representations are relatively impoverished because they do not reveal *how* the two concepts are related, including any propositional attitudes[3] that one might hold toward this relationship (such as whether one accepts or denies, believes or contests it; McGrath & Devin, 2018). For instance, complexities of the world and language, such as whether bread is usually on top of butter or vice versa, whether women should be or already are wealthy, and whether Robitussin alleviates coughing or coughing alleviates Robitussin, are thought to be reflected only by explicit evaluations but not by their implicit counterparts.

Third, under associative accounts, the retrieval processes characterizing implicit evaluation are also associative. Specifically, borrowing from spreading activation models of memory (e.g., Collins & Loftus, 1975), associative accounts posit that encountering a stimulus automatically activates the conceptual node corresponding to that stimulus in long-term memory (e.g., Bread). Once the node for the stimulus comes online, activation spreads to the conceptual nodes representing related constructs (e.g., Butter), with the degree of relatedness determined by the co-occurrence statistics of the environment. As such, the retrieval of the knowledge underlying implicit evaluation is posited to involve very little online computation.

PROPOSITIONAL ACCOUNTS OF IMPLICIT EVALUATION

During the first few decades of research on implicit evaluation, associative accounts dominated the field. However, around 15 years ago, De Houwer (2006) published a curious finding: Contrary to the expectation that implicit evaluations should not be responsive to one-shot language-based learning, merely instructing participants about upcoming stimulus pairings ("Laapians will be paired with pleasant images") influenced responding on a subsequent IAT (see also Kurdi & Banaji, 2017). The first building block for propositional accounts (De Houwer, 2014; 2018; De Houwer & Hughes, 2016; Mandelbaum, 2016; Van Dessel et al., 2018) was laid. These accounts depart from associative accounts along all three dimensions discussed above, including the types of information to which implicit evaluations are posited to be sensitive, the mental representations from which they are thought to emerge, and the computations involved in their retrieval from memory.

First, propositional accounts suggest that implicit evaluations can be acquired and updated not only as a result of protracted exposure to co-occurrence information but also as a result of single-shot language-based learning. Such learning is hypothesized to be sensitive to the structure of the input beyond mere co-occurrence, including bread being above versus below butter, women who are wealthy versus women who wish to be wealthy, and Robitussin preventing coughing versus coughing preventing Robitussin. Crucially, under propositional

3. Propositional attitudes are not to be confused with *attitudes* as used in social psychology, that is, evaluative knowledge (Fazio, 2007).

accounts, implicit evaluations are hypothesized to show sensitivity to relational information even when contradictory co-occurrence information about the same attitude object is also available.

Second, the representations from which implicit evaluations emerge are posited to be propositional rather than associative. That is, they are thought to contain symbols reflecting the way in which two concepts are related to each other, rather than merely the degree of their relatedness, such as Below (Bread, Butter), Wishes (Woman, Be Wealthy), or Causes (Cold, Sneezing) (De Houwer, Van Dessel, & Moran, 2020; Gawronski & Strack, 2004). Moreover, unlike associations, which can be strong or weak, propositions can be true or false. As such, they have the ability to encode differences between two propositions that are characterized by the same semantic content but different propositional attitudes (McGrath & Devin, 2018), such as "I believe that propositional theories are largely accurate" vs. "I doubt that propositional theories are largely accurate." The adaptive nature of such representations is clearly demonstrated by these two examples: Different propositional attitudes attached to the same semantic content can have diametrically opposed evaluative implications.

Third, propositional accounts suggest that implicit evaluation emerges from the automatic (and sometimes incomplete) retrieval of propositions (Van Dessel, Gawronski, & De Houwer, 2019) rather than the spreading of activation through an associative network. As such, incomplete retrieval of propositions is the primary explanatory tool that propositional accounts have at their disposal to explain dissociations between explicit and implicit evaluations, including their differential sensitivity to certain experimental manipulations (e.g., Lai et al., 2014). Indeed, if explicit and implicit evaluations are sensitive to the same types of information and both emerge from propositional representations, then any differences between them must be explained in terms of retrieval processes.

THE ROLE OF RELATIONAL INFORMATION IN IMPLICIT EVALUATION

Propositional accounts of implicit evaluation have been highly impactful in a way that transcends whether all empirical predictions derived from them turn out to be accurate: By introducing the field to the idea that implicit evaluations may be more sensitive to inferential reasoning and relational information than even the most flexible dual-process theories available at the time had recognized (Gawronski & Bodenhausen, 2006), these accounts have reinvigorated research on evaluative learning to a degree that seemed unlikely if not inconceivable 15 years ago. After all if, as posited by associative accounts, implicit evaluations simply track the long-term co-occurrence statistics of the environment, then there is not much to be investigated about mechanisms of acquisition and change.

Thanks in large part to propositional accounts, empirical research has started probing connections between implicit evaluation and high-level cognition going beyond the passive recording of co-occurrence information, including reasoning about diagnosticity (e.g., Cone & Ferguson, 2015; Cone, Flaharty, & Ferguson,

2019), the reinterpretation of previous evidence (e.g., Mann & Ferguson, 2015), and, crucially for the present purposes, reasoning about relational information (e.g., Zanon, De Houwer, Gast, & Smith, 2014). A comprehensive summary of this literature is beyond the scope of the present article (for recent reviews, see Cone, Mann, & Ferguson, 2017; De Houwer et al., 2020). Instead, we retain a relatively narrow focus on empirical evidence that has investigated the sensitivity of implicit evaluations to relational information in the presence of contradictory co-occurrence information.

One central point of contention between associative and propositional accounts concerns the types of information that should be capable of shifting implicit evaluations. Specifically, despite many differences between them, associative accounts converge on the idea that implicit evaluations reflect primarily the co-occurrence statistics of the environment; when co-occurrence information and relational information conflict, the former should determine implicit evaluations. Propositional accounts agree that implicit evaluations may be sensitive to co-occurrence information; however, under these accounts, implicit evaluations should not be impervious to relational information when co-occurrence information and relational information conflict. Given the substantial degree of divergence between associative and propositional accounts on this issue, below we focus on studies involving exposure to conflicting co-occurrence information and relational information with subsequent measurement of implicit and explicit evaluations (see Table 1).[4]

Some typical studies featured in this review include 1) Rydell, McConnell, Mackie, and Strain (2006), in which a person named Bob was paired with subliminally presented trait adjectives (co-occurrence information) of one valence and then described via behavioral statements of the opposite valence (relational information); 2) Moran and Bar-Anan (2013), in which families of alien creatures co-occurred with pleasant and unpleasant sounds, with some creatures starting and other creatures stopping these sounds (relational information); 3) Mann, Kurdi, and Banaji (2020), in which a novel social target was first paired with aversive screams (co-occurrence information) and then described positively using diagnostic behavioral information; and 4) Kurdi, Morris, and Cushman (2020), in which participants observed the operation of a causal system in which two stimuli were equally statistically associated with reward (co-occurrence information) but only one of those stimuli was causally responsible for it (relational information).

As should be clear based on these examples, the studies in this literature differ from each other considerably in terms of the attitude objects investigated (individuals, social groups, and nonsocial targets), the type of co-occurrence information (valenced

4. We focus on the effects of relational information on implicit evaluation for two reasons. First, associative and propositional accounts agree that explicit evaluations should be sensitive to relational information. Based on our review, this prediction is overwhelmingly supported by empirical evidence. Interestingly, explicit evaluations sometimes seem to additionally reflect the effects of co-occurrence information (e.g., Moran & Bar-Anan, 2013; Peters & Gawronski, 2011), a finding that should be addressed in future work. Second, current propositional accounts do not make precise claims about the relative importance of co-occurrence vs. relational information in shifting implicit evaluations; therefore, sensitivity of implicit evaluations to relational information in the presence of conflicting co-occurrence information favors propositional over associative accounts.

TABLE 1. Summary of published studies in which co-occurrence information and relational information have contradictory evaluative implications. The first section contains studies that provide evidence for the dominance of co-occurrence information over relational information in the updating of implicit evaluations, the second section contains studies that provide evidence for the dominance of relational information over co-occurrence information, and the third section contains studies in which the pattern of evidence is mixed. Studies sharing the same superscript are closely related to each other. For instance, Heycke and colleagues (2018) is a (failed) direct replication of Rydell and colleagues (2006).

Citation	Co-occurrence information	Relational information
Dominance of co-occurrence information over relational information		
DeCoster, Banner, Smith, & Semin (2006)	Novel targets paired with valenced traits (e.g., "smart" and "lazy")	Descriptions contain either no negation (e.g. "Phil is smart") or negation (e.g., "John is not lazy")
Deutsch, Gawronski, & Strack (2006), Exp. 4–6[a]	Positive primes (e.g., "sunshine") vs. negative primes (e.g., "garbage")	Affirmative (e.g., "a" or "an") or negating (e.g., "no") qualifier
Gawronski, Deutsch, Mbirkou, Seibt, & Strack (2008)[b]	Exposure to stereotype-consistent (men/strong–women/weak, White/good–Black/bad) and stereotype-inconsistent (women/strong–men/weak, Black/good–White/bad) pairings	Affirming stereotype-inconsistent pairings or negating stereotype-consistent pairings
Gregg, Seibt, & Banaji (2006), Exp. 3–4[c]	Narrative about novel groups (Luupites vs. Niffites) in which one group is described as positive and the other group as negative	Instruction to suppose that roles in the story had been reversed (Exp. 3–4) or narrative about the groups having switched character
Hu, Gawronski, & Balas (2017b)	Evaluative conditioning of shapes with different color patterns (CSs) with valenced images (USs)	Counterconditioning instructions
Moran & Bar-Anan (2013)[d]	Evaluative conditioning of two families of alien creatures (CSs) with positive sounds (USs) and two families of alien creatures (CSs) with negative sounds (USs)	Two families of CSs starting USs and two families of CSs ending USs
Moran & Bar-Anan (2019)	Evaluative conditioning of two families of alien creatures (CSs) with positive sounds (USs) and two families of alien creatures (CSs) with negative sounds (USs)	Two families of CSs starting USs and two families of CSs ending USs
Rydell, McConnell, Mackie, & Strain (2006)[e]	Novel social target (Bob) paired with subliminally presented positive (negative) words	The same target described as negative (positive) using supraliminally presented behavioral statements
Dominance of relational information over co-occurrence information		
De Houwer & Vandorpe (2010)	Compound cues (EF and GH) paired with outcomes (O1 and O2)	Context pairs establishing that compound cues are associated with same (e.g., A–O1, B–O1, AB–O1) or different (e.g., A–O1, B–O1, AB–O2) outcomes as elemental cues
Gast & De Houwer (2013), Exp. 2B	Evaluative conditioning of nonwords ("UDIBNON" and "ENANWAL"; CSs) with valenced images (USs)	Counterconditioning instructions

(*continued*)

TABLE 1. (*continued*)

Heycke, Gehrmann, Haaf, & Stahl (2018)[e]	Novel social target (Bob) paired with subliminally presented positive (negative) words	The same target described as negative (positive) using supraliminally presented behavioral statements
Hughes, Ye, Van Dessel, & De Houwer (2019)	Novel social target co-occurring with positive or negative adjectives	Target described as causing, predicting, or being unrelated to the adjectives
Kurdi & Dunham (2019)	Novel groups (Laapians and Niffians) embedded in conditional statements containing positive and negative trait adjectives	Truth value associated with conditional statements revealed via disambiguating stimuli presented following the statement
Kurdi, Morris, & Cushman (2020)	Positive (e.g., diamond) and negative (e.g., toxic sludge) outcomes associated with neutral colored shapes	Colored shapes either causally responsible or merely spatiotemporally associated with outcomes
Siegel, Sigall, & Huber (2011)	Evaluative conditioning of novel groups (Luupites vs. Niffites; CSs) with positive and negative adjectives (USs)	Stimulus pairings described as accurate or randomly generated (Exp. 1–2) or presented with or without negation (Exp. 3)
Whitfield & Jordan (2009), Exp. 2–4	Evaluative conditioning of novel (Exp. 2–3) and known (Exp. 4) social targets (CSs) with valenced words and images (USs)	Same targets presented with positive (e.g., "does his grandparents' yard work") and negative (e.g., "is rude to his mother") behavioral statements
Zanon, De Houwer, & Gast (2012)	Evaluative conditioning of compound nonwords (CSs) with positive and negative outcomes (wins vs. losses, USs)	Context pairings indicating whether compound cues have the same (e.g., A+, B+, AB+) or the opposite (e.g., C+, D+, CD-) valence as elemental cues
Mixed patterns of evidence		
Boucher & Rydell (2012)	Valenced behavioral statements about (a) novel social target(s)	Behavioral statements described as true (e.g., "Bob would do this") or false (e.g., "Bob would not do this")
Deutsch, Kordts-Freudinger, Gawronski, & Strack (2009)[a]	Positive primes (e.g., "sunshine") vs. negative primes (e.g., "garbage")	Affirmative (e.g., "a" or "an") or negating (e.g., "no") qualifier
Gawronski, Walther, & Blank (2005)	Positive (e.g., "likes to help new colleagues") and negative (e.g., "often insults the secretary") behavioral statements about source individuals	Target individuals described as liking or disliking source individuals paired with behavioral statements
Hu, Gawronski, & Balas (2017a)	Evaluative conditioning of pharmaceutical products (CSs) with positive and negative health-related conditions (USs)	CS described as causing or preventing US
Hughes, Ye, & De Houwer (2018)	Evaluative conditioning of nonwords ("ambik" vs. "safrom"; CSs) with positive and negative words (USs)	Context pairings identical (e.g., "up–up") or opposite (e.g., "up–down") in meaning
Johnson, Kopp, & Petty (2016)[b]	Exposure to stereotype-consistent (White/good–Black/bad) and stereotype-inconsistent (Black/good–White/bad) pairings	Affirming stereotype-inconsistent pairings or negating stereotype-consistent pairings
Kurdi & Banaji (2019), Exp. 2	Evaluative conditioning of novel groups (Laapians vs. Niffians; CSs) with positive and negative images (USs)	Stimulus pairings described as diagnostic ("revealing the nature of the groups") or nondiagnostic ("randomly generated by the computer")

Mann, Kurdi, & Banaji (2019)[c]	Evaluative conditioning of novel social target (CS) with unpleasant screams (US); control target presented without screams	Target described positively via diagnostic behaviors (Exp. 1–4) or instruction to suppose that the roles of target and control target were reversed (Exp. 4)
Moran, Bar-Anan, & Nosek (2015)[d]	Evaluative conditioning of two families of alien creatures (CSs) with positive sounds (USs) and two families of alien creatures (CSs) with negative sounds (USs)	Two families of CSs starting USs and two families of CSs ending USs
Moran, Bar-Anan, & Nosek (2016)	Four novel social targets described via positive and negative behavioral statements	Instruction to suppose that behavioral statements about one target should be switched with behavioral statements about a different target
Peters & Gawronski (2011)	Four novel social targets described via positive and negative behavioral statements	Feedback about accuracy of the statement (true vs. false)
Wyer (2016), Exp. 2A and 2B[c]	Narrative about novel groups (Luupites vs. Niffites) in which one group is described as positive and the other group as negative	Instruction to suppose that roles in the story had been reversed, either with or without re-exposure to narrative with reversed roles
Zanon, De Houwer, Gast, & Smith (2014)	Evaluative conditioning of nonwords (CSs) with positive and negative adjectives (USs)	Verbal descriptions indicating whether pairings represent equivalence or opposition relationship

words, melodies and screams, pleasant and unpleasant images), the type of relational information (the temporal or causal structure of events, short verbal instructions, and narratives), the order of co-occurrence and relational information, and several other features. Given the number and diversity of operationalizations of central constructs, we believe that the inferences that can be drawn from this set of studies should be quite informative with regard to the underlying theoretical questions.

As shown in Table 1, the relevant studies can be assigned to three different categories on the basis of the main finding emerging from them: one group of studies suggests that co-occurrence information dominates the updating of implicit evaluations, a second group of studies suggest that relational information dominates the updating of implicit evaluations, and a third group of studies finds a mixed pattern of dominance. We now turn to discussing the overall picture emerging from these three sets of studies, which seems to favor the propositional over the associative perspective.

EVIDENCE FOR THE DOMINANCE OF CO-OCCURRENCE INFORMATION IN UPDATING IMPLICIT EVALUATIONS

A first, relatively small, set of studies (see the first section of Table 1) provides evidence for the dominance of co-occurrence information in the presence of conflicting relational information. Specifically, Deutsch, Gawronski, and Strack (2006) and Gawronski, Deutsch, Mbirkou, Seibt, and Strack (2008) investigated the effects of negation; Gregg, Seibt, and Banaji (2006) the effects of supposing that the roles of the two protagonists in a previously read narrative had been reversed; Moran and Bar-Anan (2013) the effects of starting vs. stopping a valenced stimulus; and Rydell and colleagues (2006) the effects of behavioral descriptions contradicting subliminally presented valenced words. These studies found that the effects of co-occurrence information on implicit evaluation were stronger than those of relational information, thus supporting associative accounts.[5]

However, closely related follow-up work has put considerable constraints on the generalizability of each of these findings. Specifically, Deutsch, Kordts-Freudinger, Gawronski, and Strack (2009) found sensitivity of implicit evaluations to negation on a different indirect measure and Johnson, Kopp, and Petty (2016) observed effects of a stronger manipulation of negations on implicit evaluation. With regard to the study by Gregg and colleagues (2006), it has been demonstrated that abstract supposition can influence implicit evaluations if participants are re-exposed to the relevant prior information (Wyer, 2016). Moreover, recent findings by Mann and colleagues (2020) suggest that although the supposition instruction seems to be ineffective in shifting implicit evaluations, other types of relational information, such as diagnostic behavioral information, can have impact. Moran and Bar-Anan

5. Notably, these findings do not necessarily contradict propositional accounts given that such accounts are not inconsistent with the possibility of co-occurrence information dominating over relational information to the extent that co-occurrence information gives rise to propositional inferences. As such, cases in which relational information dominates over co-occurrence information are more diagnostic in terms of their ability to arbitrate between the two types of account.

have conducted follow-up work themselves to demonstrate that implicit evaluations can respond to the distinction between starting and stopping a stimulus to the degree that initial instructions do not direct attention to co-occurrence information (Moran, Bar-Anan, & Nosek, 2015). Finally, the dissociation observed by Rydell and colleagues (2006) according to which implicit evaluations are uniquely responsive to subliminally presented co-occurrence information could not be confirmed in an independent replication attempt (Heycke, Gehrmann, Haaf, & Stahl, 2018). Rather, in the new set of studies, both explicit and implicit evaluations were in line with supraliminally presented behavioral information.

To summarize, to date, studies on the updating of implicit evaluations offer little convincing evidence for the dominance of co-occurrence information over conflicting relational information. To clarify, based on a large body of work, it is evident that implicit evaluations can be sensitive to co-occurrence information (e.g., Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). However, when co-occurrence information and relational information are in conflict, the former does not necessarily exert a larger influence on implicit evaluation than the latter.

EVIDENCE FOR THE DOMINANCE OF RELATIONAL INFORMATION IN UPDATING IMPLICIT EVALUATIONS

The overwhelming majority of the studies included in this review demonstrate the possibility of relational information dominating over co-occurrence information in the updating of implicit evaluations. A complete list of these studies is provided in the second and third sections of Table 1; here we briefly review some evidence from our own labs to highlight recent advances in this area. It should be noted that the mere fact that relational information can drown out the effects of co-occurrence information can be seen as supporting propositional over associative accounts given that associative accounts categorically reject this possibility, whereas propositional theories make an existence proof type of argument in favor of it. The wide range of conditions under which such a pattern of results can be obtained is, as it were, only the icing on the propositional cake.

For instance, in some recent studies by Mann and colleagues (2019), a novel social target was first paired with highly aversive scream stimuli (co-occurrence information), with a subsequent brief verbal narrative revealing positive diagnostic information about the target (relational information). Initially formed negative implicit evaluations were substantially updated under a wide range of conditions, including when diagnostic behavioral information provided an explanation for the earlier scream pairings (e.g., the screams were described as the screams of abused women that the target supported as a social worker) and when it did not (e.g., the target was described as an animal welfare worker). Moreover, this study included a first direct comparison between two different types of relational information in shifting implicit evaluations in the presence of conflicting co-occurrence information: Whereas the diagnostic behavioral information mentioned above produced considerable shifts in implicit evaluation, the mere instruction to suppose that the target person was not paired with screams did not have impact.

In another recent set of studies, Kurdi and Dunham (2020) probed the sensitivity of implicit evaluations to inferential reasoning above and beyond the co-occurrence structure of language. Specifically, participants were exposed to conditional statements in which names drawn from two novel social groups co-occurred equally frequently with positive and negative traits; however, one group was subsequently revealed to be characterized only by positive traits and the other group only by negative traits. Implicit evaluations reflected the propositional implications of the verbal statements rather than merely the co-occurrence information embedded in them.

In subsequent studies, the evidence emerging in favor of the propositional perspective was even stronger: In these studies, changes in implicit evaluation were modulated by participants' propensity to commit normative errors in inferential reasoning. For instance, participants who committed denying the antecedent error (i.e., inferred ¬B from A → B and ¬A) exhibited updating in a direction opposite of that suggested by the pairings embedded in the verbal material. By contrast, participants who reasoned accurately did not show updating. As such, in line with propositional accounts, these studies provide clear evidence for the sensitivity of implicit evaluations to complex operations of (correct and erroneous) inferential reasoning, including the use of conditionals, above and beyond co-occurrence information.

Finally, Kurdi and colleagues (2020) tested the sensitivity of implicit evaluations to causal relationships. Across five experiments, participants observed the operation of physical systems in which two sets of objects were equally predictive of the appearance of valenced objects (co-occurrence information) but differed in their causal status (i.e., causal vs. merely associated; relational information). Implicit evaluations were found to be consistently sensitive to causal status: For instance, when outcomes were positive, causal stimuli were implicitly preferred to merely associated stimuli. Crucially, in these studies, either no verbal instructions about the operation of the machine were provided or the verbal instructions were redundant with the events that participants could directly observe. As such, this work provides evidence that relational information can dominate over co-occurrence information in the updating of implicit evaluations even when the former is inferred purely from observation.

MIXED PATTERNS OF EVIDENCE

Although the preponderance of the evidence suggests that relational information can have a stronger effect on implicit evaluations than conflicting co-occurrence information under a wide range of conditions, a third set of studies demonstrates that such a pattern of results is not universal. Specifically, we identified five groups of moderator variables that seem to predict whether relational information dominates the updating of implicit evaluations. These moderator variables include the strength of the manipulation, the indirect measure used, high-level processing instructions, the timing of co-occurrence vs. relational information, and the nature of the relational information itself (see Table 2).

To summarize, it seems that sometimes more obvious manipulations of relational information are needed to influence implicit evaluations than their explicit

TABLE 2. Summary of studies in which implicit evaluations show a mixed pattern of sensitivity to contradictory co-occurrence information vs. relational information. The first of each pair of conditions is associated with the dominance of relational information over co-occurrence information and the second one is associated with the dominance of co-occurrence information over relational information. For instance, in Boucher and Rydell (2012), when negations were visually salient during encoding, implicit evaluations reflected the propositional implications of those negations, whereas when they were not visually salient, implicit evaluations reflected the pairings of targets and trait adjectives.

Strength of manipulation	Indirect measure used	Timing of co-occurrence vs. relational information	Nature of relational information	High-level processing instructions
Boucher & Rydell (2012): Salience of negation (salient vs. non-salient) during encoding Hughes, Ye, & De Houwer (2018): Nature of context pairs (valenced vs. non-valenced) Johnson, Kopp, & Petty (2016): Nature of negation (meaningful vs. simple)	Deutsch, Kordts-Freudinger, Gawronski, & Strack (2009): AMP vs. EPT Moran, Bar-Anan, & Nosek (2016): IAT vs. EPT or AMP	Gawronski, Walther, & Blank (2005): Relational information provided simultaneously vs. after co-occurrence information Hu, Gawronski, & Balas (2017a): Relational information provided simultaneously with vs. before co-occurrence information Kurdi & Banaji (2019), Exp. 2: Relational information provided before vs. after co-occurrence information Peters & Gawronski (2011): Relational information provided before vs. after co-occurrence information Zanon, De Houwer, Gast, & Smith (2014): Relational information provided before vs. after co-occurrence information	Mann, Kurdi, & Banaji (2020): Unrelated relational information vs. suppose opposite pairings Wyer (2016), Exp. 2A and 2B: Re-exposure vs. no re-exposure to original information	Moran, Bar-Anan, & Nosek (2015): Impression formation vs. contingency learning

counterparts. For instance, Boucher and Rydell (2012) found implicit evaluations to be sensitive to negation only when negation was made visually salient, whereas Johnson and colleagues (2016) found a moderating effect for the meaningfulness of the negation manipulation. Similarly, in Hughes, Ye, and De Houwer (2018), evaluative conditioning effects on implicit evaluation were modulated by context pairings only if the context pairings were valenced and, as such, more clearly applicable to the interpretation of the CS–US pairings. Another set of studies (Deutsch et al., 2009; Moran, Bar-Anan, & Nosek, 2016) found implicit evaluations to be more sensitive to relational information when measured with some indirect measures rather than others. Moreover, as mentioned above, the effects of relational information on implicit evaluations can depend on high-level processing instructions (Moran et al., 2015).

Crucially for our argument below, the sensitivity of implicit evaluations to relational information also seems to be modulated by the temporal order of and dependency between co-occurrence information and relational information. Specifically, implicit evaluations appear to reflect relational information when such information is provided simultaneously with or directly before or after co-occurrence information but not when such information would have to be used to retroactively revaluate co-occurrence information. For instance, in Peters and Gawronski (2011), validity information modulated the effects of stimulus pairings only when such validity information was provided immediately following the pairings but not after a delay. Similarly, Kurdi and Banaji (2019) found that evaluative conditioning effects were moderated by information about the diagnosticity of CS–US pairings when such information was provided before but not after exposure to the pairings. Other findings in a similar vein have been obtained in a relatively large set of studies (including Gawronski, Walther, & Blank, 2005; Hu, Gawronski, & Balas, 2017a; Zanon et al., 2014).

Moreover, as mentioned above, Gregg and colleagues (2006) found that implicit evaluations were impervious to the abstract supposition that the roles of the two protagonists in a previously read narrative were reversed. In follow-up work, Mann and colleagues (2020) confirmed that the abstract supposition instruction was also ineffective following aversive evaluative conditioning, however, not because implicit evaluations were generally resistant to change. Specifically, a manipulation involving diagnostic behavioral information had impact. In other follow-up work, Wyer (2016) concluded that re-exposure to the Gregg and colleagues' narrative with reversed contingencies did modulate implicit evaluations, whereas abstract supposition without re-exposure produced no effect.

A CHALLENGING PATTERN OF RESULTS FOR PROPOSITIONAL ACCOUNTS

Taken together, these studies provide convincing evidence for a central tenet of propositional accounts, namely that relational information can have a stronger influence on implicit evaluations than conflicting co-occurrence information.

Moreover, it seems that such a pattern of updating, although subject to certain boundary conditions, is the rule rather than the exception. By contrast, this body of work is extremely difficult, if not impossible, to reconcile with associative accounts under which implicit evaluations should be impervious to relational information, at least in the presence of conflicting co-occurrence information.

However, without auxiliary assumptions, current propositional accounts have difficulty explaining the common pattern of findings described above in which implicit evaluations seem to be selectively impervious to manipulations that involve the retrospective revaluation of (co-occurrence) information already encoded in long-term memory. For instance, under propositional accounts, the initial narrative read by participants in Gregg and colleagues (2006) should lead to encoding of the propositional representation, "I believe that Niffians are peaceful." The subsequent abstract supposition manipulation, in turn, should change either the content of these representations ("I believe that Niffians are *not* peaceful") or the propositional attitudes attached to them ("I *do not* believe that Niffians are peaceful"). In either case, following the abstract supposition manipulation, implicit evaluations should reflect the newly updated proposition. Instead, implicit evaluations seem to exhibit selective recalcitrance to manipulations involving retroactive revaluation of prior co-occurrence information without any re-exposure to that co-occurrence information.

Notably, in the same set of studies, explicit evaluations have been shown to shift in response to such manipulations. Although this pattern is not universal (Gawronski et al., 2005; Kurdi & Banaji, 2019; Zanon et al., 2014), it is clear that explicit evaluations have the ability to shift in response to a type of manipulation that leaves already acquired implicit evaluations intact (Gregg et al., 2006; Hu et al., 2017a; Mann et al., 2020; Peters & Gawronski, 2011; Wyer, 2016). This pattern of dissociation is all the more curious given that, under propositional accounts, explicit and implicit evaluations emerge from the same propositional representations. However, if direct and indirect measures of attitude access the same propositional representations, then why do implicit evaluations show a selective lack of sensitivity to retrospective revaluation of already encoded information?

As mentioned above, propositional accounts explain patterns of explicit–implicit dissociation at the level of retrieval processes (Van Dessel et al., 2019). According to this argument, both explicit and implicit evaluations are fully propositional; however, some propositions may not be (fully) retrieved under the automaticity conditions created by indirect measures (De Houwer et al., 2009). Although this possibility is consistent with the empirical evidence, the challenge that any such proposal faces is addressing the issue of which propositions can and cannot be retrieved on indirect measures. Absent such an account, such proposals do not seem to be falsifiable. Indeed, within this general framework, any convergence between explicit and implicit evaluations can be explained by propositions that were activated on both direct and indirect measures, whereas any divergence can be attributed to a failure to retrieve some propositions on the indirect, but not on the direct, measure. To circumvent this problem, future iterations of the

propositional account should specify the types of propositions or the conditions under which propositions are and are not retrievable when attitudes are measured indirectly.[6]

A PROPOSAL FOR INTEGRATION: THE COMMON CURRENCY HYPOTHESIS

The sensitivity of implicit evaluations to relational information at acquisition, combined with their recalcitrance in the face of certain types of updating that require access to details of the original proposition, gives rise to an intriguing hypothesis, which we believe deserves empirical testing: Namely, implicit evaluations may serve as a "common currency" to make otherwise incommensurable propositions characterizing the same attitude object commensurable with each other and thereby enable rapid and smooth automatic responding to attitude objects in social behavior.

This idea has its roots in economic accounts of human decision-making as well as empirical research on its neural substrates. If one likes both apples and oranges, how should one choose between them? Under the common currency proposal (Levy & Glimcher, 2012; Rustichini, 2009; Sugrue, Corrado, & Newsome, 2005), decisions of this kind are supported by assigning reward values to different options on a common scale. For instance, if an apple is worth a reward of +5 and an orange a reward of +8, then one should prefer two apples to one orange but one orange to one apple.

Despite their many virtues, propositions exacerbate the problem of comparing apples and oranges: Complex propositional representations frequently carry contradictory evaluative implications not only as they apply to different attitude objects but even to the same one. For instance, if an attitude object is known to cause euphoria, to interfere with sleeping, and to go together with headaches, should it be approached or avoided? The need to integrate incommensurable pieces of propositional information online seems to be fundamentally incompatible with the ability of evaluative knowledge to swiftly guide behavior.

In line with propositional accounts, the common currency hypothesis (for a schematic overview, see Table 3) proposes that, at initial acquisition, the knowledge structures underlying explicit and implicit evaluations are sensitive to both co-occurrence information and relational information. As such, unlike associative accounts, the common currency hypothesis does not accord any special importance to co-occurrence information in influencing implicit evaluations. However, in line with associative accounts, the common currency hypothesis posits that explicit and implicit evaluations emerge from different mental representations, specifically with regard to their level of compression: Whereas explicit evaluations are

6. Similar arguments apply to other types of dissociation, including different patterns of long-term change in explicit and implicit evaluations (Charlesworth & Banaji, 2019), the presence of implicit ingroup preference in the absence of any explicit preference (Ratliff et al., 2020), and dissociations in responsiveness to experimental manipulations not specifically addressed here.

TABLE 3. Schematic summary of the common currency hypothesis. Both implicit and explicit evaluations are shown to be sensitive to the same basic types of information, including co-occurrence information and relational information. Whereas implicit evaluations emerge from representations of evaluative knowledge as a compressed scalar value, explicit evaluations emerge from representations of the same knowledge via high-dimensional sentence embeddings. Correspondingly, the retrieval of the knowledge underlying explicit evaluations requires considerably more online computation than does the retrieval of the knowledge underlying implicit evaluations. While explicit evaluations are potentially amenable to change in response to all types of co-occurrence and relational information, implicit evaluations are not predicted to change as a result of relational information that refers to the high-dimensional details of past experience given that such details are not encoded in the compressed summary representation.

	Acquisition		Representation	Retrieval	Updating
Implicit evaluations	Co-occurrence information: (a) Pairings with images (b) Pairings with sounds (c) Pairings with words (d) Pairings with wins/losses …	Relational information: (a) Abstract supposition (b) Behavioral statements (c) Context pairings (d) Diagnosticity information (e) Instructed pairings (f) Type of relationship (g) Validity information …	Running tally of the difference in cosine similarity of sentence embeddings corresponding to relevant propositions (A_1, A_2, . . ., A_n) to "A is good" vs. "A is bad" (scalar, or 1-dimensional quantity): $\sum_{i=1}^{n} \cos(A_i, A\,is\,good) - \cos(A_i, A\,is\,bad)$	Lookup of compressed value associated with attitude object; involves little or no online computation	Possible based on co-occurrence information or relational information, as long as relational information does not refer to high-dimensional details of initial proposition
Explicit evaluations			Matrix of relevant sentence embeddings, with each relevant proposition corresponding to a row of a matrix with k columns, depending on the dimensionality of the embeddings ($n \times k$ dimensional): $\begin{bmatrix} -0.02 & \dots & \dots & 0.10 & 0.18 \\ 0.22 & \dots & \dots & -0.05 & -0.08 \\ -0.12 & \dots & \dots & -0.33 & 0.06 \end{bmatrix}$	Strategic integration of relevant propositions; involves large amount of online computation	Possible based on any relevant co-occurrence information or relational information

subserved by high-dimensional propositional representations of relevant experience, implicit evaluations are subserved by compressed summary representations of the same experience.

Although this proposal is preliminary at this point, we illustrate the idea with reference to an existing computational[7] framework that captures the notion of different levels of compressing information and so points the way towards future formalization: the method of sentence embeddings (e.g., Cer et al., 2018). Sentence embeddings use distributional statistics of text data to algorithmically represent the meaning of sentences in relatively high-dimensional semantic space. For instance, the universal sentence encoder (Cer et al., 2018), a specific sentence embeddings model, has the ability to transform any English sentence into a 512-dimensional vector of real numbers. One of several benefits of this method is that it allows users to perform any operation that can generally be performed using vectors: Notably, semantic similarity between two propositions can be computed by calculating the cosine of the angle (correlation) between them in 512-dimensional space.

Applying a technique like this to propositional representations, one might imagine explicit evaluations emerging from knowledge represented via large matrices consisting of 512-dimensional vectors, with each row corresponding to a proposition characterizing the attitude object A. For instance, row 1 of the matrix might encode A Causes Euphoria, row 2 might encode A Interferes With Sleep, and row 3 might encode A Goes Together With Headaches. This idea illustrates the extent of the problem facing propositional accounts: Well-known attitude objects are associated with large amounts of relevant experience and, as such, high numbers of corresponding propositions characterizing them. Specifying how quick retrieval of implicit attitudes unfolds given such high-dimensional representations is a nontrivial task. And, as mentioned above, if retrieval is thought to be incomplete (Van Dessel et al., 2019), it is unclear what (elements of) propositions are thought to be amenable to automatic retrieval.

To address this issue, we propose that implicit evaluations emerge from scalar value representations that provide a running tally of the subjective value of the totality of one's experience with an attitude object, including previously encountered co-occurrence information and relational information. In the context of word embeddings, one might conceive of this running tally as the relative cosine distance (semantic similarity) of each relevant proposition characterizing an attitude object from the propositions A Is Good and A Is Bad. In other words, propositions that express positive content overall (e.g., A Causes Euphoria) are represented by adding a positive value to the running tally associated with the attitude object, and propositions that express negative content overall (e.g., A Goes Together With Headaches) are represented by adding a negative value to the running tally associated with the attitude object.

This type of representation retains many benefits of propositional representations (including, critically, their sensitivity to relational information) without creating

7. We use the term *computational* in line with Marr's (1982) definition to denote the high-level function of a cognitive system. We do not believe that the human mind, let alone the brain, computes cosine similarities between 512-dimensional vectors to calculate semantic relatedness.

an insurmountable challenge of performing complex online computations at retrieval, which seems to be incommensurable with the very idea of implicit evaluation. Moreover, unlike with propositional representations whose dimensionality increases with each additional proposition, a compressed running tally of this kind can continuously update its value as a function of increasing experience without any increase in complexity. As such, the automatic retrieval of the knowledge structures underlying implicit evaluations of highly familiar attitude objects becomes computationally tractable.

In addition to addressing the issue of how relational information can influence responding on indirect measures of cognition without requiring a large amount of online computation, this proposal readily explains the patterns of explicit–implicit dissociation described above (Gregg et al., 2006; Hu et al., 2017a; Kurdi & Banaji, 2019; Mann et al., 2020; Peters & Gawronski, 2011; Wyer, 2016; Zanon et al., 2014). Specifically, under the present proposal, interventions aimed at retrospective revaluation of past co-occurrence information, such as the instruction to suppose that already experienced pairings had been reversed or already encoded pairings with trait adjectives should be seen as invalid, cannot influence implicit evaluation because implicit evaluations are unable to access high-dimensional details of the relevant earlier experience.[8] By contrast, new information that does not require access to a high-dimensional representation of earlier experience (e.g., diagnostic information unrelated to the previous conditioning experience or re-exposure to a modified narrative with reversed contingencies) should simply lead to a modification of the running tally in the corresponding direction.

This idea can be subjected to further empirical testing in myriad different paradigms (for initial evidence, see Kurdi, Gershman, & Banaji, 2019). For instance, the common currency hypothesis predicts that in the context of propositional inference, exposing participants to new propositions with countervalent implications should succeed in shifting implicit evaluations, whereas the mere instruction to reverse the truth values associated with previously encoded propositions should fail (see also the distinction between cases 3 and 4 in the APE model; Gawronski & Bodenhausen, 2006). In the context of causal learning, direct experience with reversed causal roles should have impact; however, implicit evaluations should not be modulated by an intervention that requires retroactive revaluation of the stimuli, for example, as a result of information that the stimuli had been erroneously switched by the experimenter. The reason for this proposed asymmetry is that, under the current account, completely new relational information can update the compressed representations from which implicit evaluations emerge in a relatively straightforward manner. However, once updating has occurred, the specific content of the (relational) information that led to that updating is not recoverable implicitly. Although it might be possible to track the magnitude of change associated with each new experience, any given magnitude of change would still be compatible with an essentially infinite range of experiences.

8. As such, unlike most current theories of implicit evaluation, the common currency proposal makes a distinction between initial acquisition and subsequent updating in terms of the sensitivity of the knowledge structures underlying implicit evaluations to different types of information.

Notably, the common currency proposal both builds on and differs from the two most prominent dual-process accounts of implicit evaluation in important ways. Specifically, similar to the MODE model (Fazio, 1995; 2007), the common currency hypothesis does not make distinctions between explicit and implicit evaluations in terms of their sensitivity to different types of information at initial acquisition. However, unlike the MODE model, which defines attitudes as associations between an object and an evaluation, the current hypothesis proposes that implicit and explicit evaluations emerge from different knowledge structures, specifically with regard to their levels of compression.

The current proposal also shows some similarity to the APE model (Gawronski & Bodenhausen, 2006) in that it posits that implicit evaluations emerge from compressed (associative) representations and explicit evaluations from less compressed (propositional) representations of evaluative knowledge. However, importantly, the current proposal does not accord any special role to co-occurrence information in shifting implicit evaluations. Further, in deviation from the APE model, this account predicts that relational information can shift implicit evaluations directly, that is, without such shifts being mediated by concomitant shifts in explicit evaluations.

CONCLUSION

In conclusion, we see the propositional approach to implicit evaluation as a remarkable success story in social cognition research over the past decades. Propositional accounts have been instrumental in emancipating the field from its intellectual predecessors positing that implicit evaluations are uniquely sensitive to co-occurrence information and emerge from simple associative representations linking attitude objects to positive or negative valence. The proposal that implicit evaluations may show widespread sensitivity to relational information and emerge from propositional representations reinvigorated experimental research on how evaluative knowledge is acquired, represented, and retrieved. Although propositional accounts can explain an impressive array of empirical findings, we propose that a minor modification to these accounts may be in order: Specifically, implicit evaluations may be subserved by compressed summary representations that are sensitive to relational information but are themselves not fully propositional.

REFERENCES

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

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

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., St John, R., et al. (2018). Universal sentence encoder for English. In E. Blanco & W. Lu (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing System Demonstrations* (pp. 169–174). Brussels: Association for Computational Linguistics.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes:

I. Long-term change and stability from 2007 to 2016. *Psychological Science, 30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

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

Cone, J., Flaharty, K., & Ferguson, M. J. (2019). Believability of evidence matters for correcting social impressions. *Proceedings of the National Academy of Sciences*, *116*(20), 9802–9807. https://doi.org/10.1073/pnas.1903222116

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in Experimental Social Psychology* (pp. 131–199). Amsterdam: Elsevier. https://doi.org/10.1016/bs.aesp.2017.03.001

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, *11*(3), 97–104. https://doi.org/10.1016/j.tics.2006.12.005

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*(2), 176–187. https://doi.org/10.1016/j.lmot.2005.12.002

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). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3), 49–21. https://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and persuasion. *Social Cognition*, *34*(5), 480–494. https://doi.org/10.1521/soco.2016.34.5.480

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* (pp. 127–183). Amsterdam: Elsevier. https://doi.org/10.1016/bs.aesp.2019.09.004

De Houwer, J., & Vandorpe, S. (2010). Using the Implicit Association Test as a measure of causal learning does not eliminate effects of rule learning. *Experimental Psychology*, *57*(1), 61–67. https://doi.org/10.1027/1618–3169/a000008

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

Deutsch, R., Kordts-Freudinger, R., Gawronski, B., & Strack, F. (2009). Fast and fragile: A new look at the automaticity of negation processing. *Experimental Psychology*, *56*(6), 434–446. https://doi.org/10.1027/1618-3169.56.6.434

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

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, *22*(1), 22–37. https://doi.org/10.1016/0022-1031(86)90039-9

Dunham, Y., Chen, E. E., & Banaji, M. R. (2013). Two signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. *Psychological Science*, *24*(6), 1–27. https://doi.org/10.1177/0956797612463081

Fazio, R. H. (1995). Attitudes as object–evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 247–282). Mahwah, NJ: Erlbaum.

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social Cognition*, *25*(5), 603–637. https://doi.org/10.1521/soco.2007.25.5.603

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037//0022–3514.69.6.1013

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

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. *Social Psychology Quarterly*, *46*(1), 23–30. https://doi.org/10.2307/3033657

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312–325. https://doi.org/10.1016/j.lmot.2013.03.003

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

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. *Journal of Experimental Social Psychology*, *41*(6), 618–626. https://doi.org/10.1016/j.jesp.2004.10.005

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

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

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology: General*, *148*(6), 1022–1040. https://doi.org/10.1037/xge0000623

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? A registered replication of Rydell et al. (2006). *Cognition & Emotion*, *32*(8), 1708–1727. https://doi.org/10.1080/02699931.2018.1429389

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. http://doi.org/10.1037/a0018916

Hu, X., Gawronski, B., & Balas, R. (2017a). 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

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science*, *8*(8), 858–866. https://doi.org/10.1177/1948550617691094

Hughes, S., Ye, Y., & De Houwer, J. (2018). Evaluative conditioning effects are modulated by the nature of contextual pairings. *Cognition & Emotion*, *33*(5), 871–884. https://doi.org/10.1080/02699931.2018.1500882

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

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

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

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). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. Manuscript submitted for publication.

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, *116*(13), 6035–6044. https://doi.org/10.1073/pnas.1820238116

Kurdi, B., Morris, A., & Cushman, F. A. (2020, February 1). *The role of causal structure in implicit cognition*. https://doi.org/10.31234/osf.io/r7cfa

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

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., 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

Marr, D. (1982), *Vision: A computational approach*, San Francisco : Freeman.

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 204–217). New York: Guilford.

McGrath, M., & Devin, F. (2018). Propositions. In *The Stanford Encyclopedia of Philosophy.* Retrieved April 7, 2020, from https://plato.stanford.edu/archives/spr2018/entries/propositions/

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743–752. https://doi.org/10.1080/02699931.2012.732040

Moran, T., & Bar-Anan, Y. (2019). The effect of co-occurrence and relational information on speeded evaluation. *Cognition & Emotion*, *34*(1), 144–155. https://doi.org/10.1080/02699931.2019.1604321

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

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition*, *34*(5), 435–461. https://doi.org/10.1521/soco.2016.34.5.435

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648–654. https://doi.org/10.3758/BF03213230

Nosek, B. A., Smyth, F. L., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., et al. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

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

Ratliff, K. A., Lofaro, N., Howell, J. L., Conway, M. A., Lai, C. K., O'Shea, B., et al. (2020, February 28). *Documenting bias from 2007–2015: Pervasiveness and correlates of implicit attitudes and stereotypes II*. https://osf.io/rfzhu/

Rustichini, A. (2009). Neuroeconomics: What have we found, and what should we search for. *Current Opinion in Neurobiology*, *19*(6), 672–677. https://doi.org/10.1016/j.conb.2009.09.012

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954–958. https://doi.org/10.1111/j.1467-9280.2006.01811.x

Siegel, E., Sigall, H., & Huber, D. E. (2011). The IAT is sensitive to the perceived accuracy of newly learned associations. *European Journal of Social Psychology*, *42*(2), 189–199. https://doi.org/10.1002/ejsp.859

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

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

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*(5), 363–375. https://doi.org/10.1038/nrn1666

Teachman, B. A., Clerkin, E. M., Cunningham, W. A., Dreyer-Oren, S., & Werntz, A. (2019). Implicit cognition and psychopathology: Looking back and looking forward. *Annual Review of Clinical Psychology*, *15*(1), 123–148. https://doi.org/10.1146/annurev-clinpsy-050718-095718

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–27. https://doi.org/10.1086/214483

Van Dessel, P., Gawronski, B., & De Houwer, J. (2019). Does explaining social behavior require multiple memory systems? *Trends in Cognitive Sciences*, *23*(5), 368–369. https://doi.org/10.1016/j.tics.2019.02.001

Van Dessel, P., Hughes, S., & De Houwer, J. (2018). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review*, *23*(3), 267–284. https://doi.org/10.1177/1088868318795730

Whitfield, M., & Jordan, C. H. (2009). Mutual influence of implicit and explicit attitudes. *Journal of Experimental Social Psychology*, *45*(4), 748–759. https://doi.org/10.1016/j.jesp.2009.04.006

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*(1), 101–126. https://doi.org/10.1037//0033-295X.107.1.101

Wyer, N. A. (2016). Easier done than undone . . . by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. *Journal of Experimental Social Psychology*, *63*(C), 77–85. https://doi.org/10.1016/j.jesp.2015.12.006

Zanon, R., De Houwer, J., & Gast, A. (2012). Context effects in evaluative conditioning of implicit evaluations. *Learning and Motivation*, *43*(3), 155–165. https://doi.org/10.1016/j.lmot.2012.02.003

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology*, *67*(11), 2105–2122. https://doi.org/10.1080/17470218.2014.907324