Toward a Learning Account of Implicit Attitude Acquisition and Change

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Kurdi, Benedek. 2019. Toward a Learning Account of Implicit Attitude Acquisition and Change. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029775">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029775</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Toward a learning account of implicit attitude acquisition and change

A DISSERTATION PRESENTED
BY
BENEDEK KURDI
TO
THE DEPARTMENT OF PSYCHOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE SUBJECT OF PSYCHOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MARCH, 2019
Toward a learning account of implicit attitude acquisition and change

The present dissertation seeks to characterize the learning processes by which implicit (indirectly revealed) evaluations of social and nonsocial stimuli are updated along three dimensions: (a) informational inputs to learning, (b) the nature of the learning process, and (c) the content and format of the resulting mental representations.

Paper 1 provides evidence that implicit evaluations can be updated both as a result of direct experience with the environment (evaluative conditioning) and purely verbal information about upcoming stimulus pairings. Moreover, it shows that the two interventions are redundant in creating implicit attitude change, suggesting that they give rise to similar evaluative representations. Paper 2 offers further insight into the process by which evaluative conditioning shifts implicit evaluations, revealing that (a) learning can asymptote quickly, after as few as four stimulus pairings, and (b) information from stimulus pairings is productively combined with purely verbal information on the diagnosticity of those stimulus pairings. Furthermore, it demonstrates that direct experience with stimulus pairings creates more durable change in implicit evaluations than verbal descriptions of stimulus pairings. Finally, Paper 3 takes a reinforcement learning approach to show that whereas implicit evaluations are responsive to model-free learning, they are impervious to model-based learning.

Overall, the current studies suggest that implicit attitudes can shift (a) as a result of certain (but not other) kinds of direct experience and certain (but not other) kinds of verbal interventions, (b) slowly or quickly, depending on the parameters of the task, but generally in a context-dependent manner, and (c) via creating highly compressed representations of value, be they associative (group A–good) or propositional (“group A is good”) in nature. As such, these results
are partly compatible with propositional theories under which implicit attitudes can shift (a) in response to direct experience or language, (b) slowly or quickly, and (c) via creating propositional representations, but largely incompatible with associative theories under which implicit attitudes should shift exclusively (a) as a result of direct experience, (b) slowly, and (c) via creating associative representations. We discuss the theoretical and practical implications of these findings, along with avenues for future exploration that they open up.
Table of Contents

Title page ........................................................................................................................................... i
Copyright ............................................................................................................................................... ii
Abstract................................................................................................................................................ iii
Acknowledgements ............................................................................................................................. ix

General Introduction .......................................................................................................................... 1

Associative and propositional theories of implicit evaluation........................................................... 1
The design and main findings of the present project ........................................................................... 7
Theoretical implications of the present project .................................................................................. 13


Abstract ............................................................................................................................................... 18
Introduction .......................................................................................................................................... 19
Shifting implicit attitudes via repeated evaluative pairings and evaluative statements ...................... 22
Comparing REP and ES ..................................................................................................................... 24
Overview of the present studies .......................................................................................................... 28
Study 1 ............................................................................................................................................... 29
Method ............................................................................................................................................... 30
Results ............................................................................................................................................... 36
Discussion .......................................................................................................................................... 38
Study 2 ............................................................................................................................................... 39
Method ............................................................................................................................................... 40
Results ............................................................................................................................................... 42
Discussion .......................................................................................................................................... 42
Study 3 ............................................................................................................................................... 43
Method ............................................................................................................................................... 44
Results ............................................................................................................................................... 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discussion</td>
<td>46</td>
</tr>
<tr>
<td>Study 4</td>
<td>47</td>
</tr>
<tr>
<td>Method</td>
<td>48</td>
</tr>
<tr>
<td>Results</td>
<td>50</td>
</tr>
<tr>
<td>Discussion</td>
<td>51</td>
</tr>
<tr>
<td>Study 5</td>
<td>51</td>
</tr>
<tr>
<td>Method</td>
<td>52</td>
</tr>
<tr>
<td>Results</td>
<td>53</td>
</tr>
<tr>
<td>Discussion</td>
<td>54</td>
</tr>
<tr>
<td>Study 6A</td>
<td>54</td>
</tr>
<tr>
<td>Method</td>
<td>54</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>55</td>
</tr>
<tr>
<td>Study 6B</td>
<td>55</td>
</tr>
<tr>
<td>Method</td>
<td>56</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>56</td>
</tr>
<tr>
<td>Study 6C</td>
<td>57</td>
</tr>
<tr>
<td>Method</td>
<td>57</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>57</td>
</tr>
<tr>
<td>Combined Analyses Across Experiments</td>
<td>58</td>
</tr>
<tr>
<td>General Discussion</td>
<td>60</td>
</tr>
</tbody>
</table>

**Paper 2: Attitude Change via Repeated Evaluative Pairings versus Evaluative Statements: Shared and Unique Features** ............................................................. 69

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>70</td>
</tr>
<tr>
<td>Introduction</td>
<td>71</td>
</tr>
<tr>
<td>What kind of information is most effective in shifting implicit evaluations?</td>
<td>76</td>
</tr>
<tr>
<td>By what kind(s) of process is evaluative learning produced?</td>
<td>77</td>
</tr>
<tr>
<td>Number of stimulus pairings presented during learning</td>
<td>79</td>
</tr>
<tr>
<td>Diagnosticity</td>
<td>81</td>
</tr>
<tr>
<td>What kind of information is most effective in durably shifting implicit evaluations?</td>
<td>82</td>
</tr>
</tbody>
</table>
Overview of the present project ........................................................................................................ 85
Study 1 ............................................................................................................................................ 86
Method ........................................................................................................................................... 87
Results .......................................................................................................................................... 91
Discussion .................................................................................................................................... 93
Study 2 ............................................................................................................................................ 94
Method ........................................................................................................................................... 95
Results .......................................................................................................................................... 98
Discussion .................................................................................................................................... 100
Study 3 .......................................................................................................................................... 102
Method ........................................................................................................................................... 104
Results .......................................................................................................................................... 107
Discussion .................................................................................................................................... 111
Study 4 .......................................................................................................................................... 112
Method ........................................................................................................................................... 112
Results .......................................................................................................................................... 114
Discussion .................................................................................................................................... 117
General Discussion .......................................................................................................................... 118
Theoretical implications .................................................................................................................. 120
Methodological implications ........................................................................................................ 124
A call for computational modeling .................................................................................................. 125
Implications for long-term change .................................................................................................. 128

Paper 3: Model-Free and Model-Based Learning Processes in the Updating of Explicit and
Implicit Evaluations .......................................................................................................................... 131

Abstract ...................................................................................................................................... 132
Introduction ..................................................................................................................................... 133
The present project .......................................................................................................................... 141
Study 1 .......................................................................................................................................... 144
Design .......................................................................................................................................... 144
Results .......................................................................................................................................... 148
Acknowledgements

For their love from afar these past five years, I am extremely grateful to my family: my mom and dad, my sister and brother, and my grandparents. I know that life would often be easier without these 4,162 miles separating us. I hope they will find retrospective solace in the fact that Ithaca is even further away (updated distance: 4,371 miles).

Since the beginning of my twenties, I’ve been moving from city to city (Budapest to Berlin to Budapest to Annandale to Budapest to Cambridge), inadvertently creating a situation where no matter where I happen to be, I am too far away from many of my friends. Whether from the same office or from across some ocean, Ádám Németh, Barbara Császár, Barbara Lozar, Brandon Woo, Cili Pető, Daniel Ramírez, Elyse Neubauer, Franckie Ramírez, Gabriel Braz, István Szokonya, Kati Koncz, Kiera Hudson, Mark Thornton, Patrick Chen, Réka Eszter Bodó, Tessa Charlesworth, and Tom Mann have supported me in too many ways to list.

From my early years as a precocious and slightly annoying child through my later years as a precocious and slightly annoying (young) adult, I’ve always been enormously lucky to have mentors who took me seriously, pushed me, gave me books and papers to read, commented on drafts, wrote letters for me, and always had an open ear for my ramblings about some topic that I typically knew very little about: At my high school László Arató and Mariann Weltner, at Eötvös Lorand University Anna Bácskai and Edit Király, at Bard College Barbara Luka, Cecile Kuznitz, and Joel Perlmann, and at Central European University, whose recent struggles should be a warning to everyone who cares about academic freedom, Darren Schreiber and Levente Littvay.

On February 27, 2014, I got arguably the most important email of my life from Mahzarin Banaji telling me that I was admitted to the PhD program at Harvard. I had a sense that I was going to join a great department, but just how great this department is still amazes me every day.
As a member of this department, one has the great luxury of being surrounded by others who
genuinely care about how the mind works and who are genuinely invested not only in their own
research but also in others’.

For the countless hours spent listening to me giving practice talks and reading drafts, I
would like to thank past and present members of and visitors at the Banaji lab, including Allison
Seitchik, Calvin Lai, Caroline Wilmuth, Christopher Dial, Emily Cogsdill, Jack Cao, Joe Vitriol,
Kiera Hudson, Kirsten Morehouse, Linda Chang, Maddalena Marini, Michael Brownstein, Niv
Reggev, Olivia Kang, Sean Hughes, Steve Lehr, Tessa Charlesworth, and Tom Mann, as well
faculty in the department, including Jason Mitchell, Jim Sidanius, Josh Greene, Mina Cikara, and
Susan Carey. Over the past five years, I have also had the great fortune of being able to mentor
extraordinary research assistants who have influenced my thinking and contributed to my work
in innumerable ways. I would like to thank all of them, and especially those who took the deep-
est dives into the science of social cognition, including Catherine Kim, Ece Hakim, Harrison
Satcher, Ruolin Lu, Sarah Ryan Shayn Lozano, and Victor Yang.

For their invaluable feedback on previous drafts of my dissertation, I would like to thank
Fiery Cushman, Sam Gershman, and Liz Phelps. Fiery and Sam have been fantastic collabora-
tors, and without their teaching and mentorship over the past five years, I would never have been
able to articulate the ideas that form the basis of this dissertation. I am also deeply indebted to
Patrick Mair who not only taught me almost everything I know about statistics but also most of
what I know about teaching.

Last but not least, although I have exchanged far too many paper drafts with her to count,
it is still beyond my writing skills to do justice to the mentorship that I have received from my
advisor Mahzarin Banaji. At our first meeting in July 2014, Mahzarin told me that success in

x
graduate school was the function of a single variable: time on task. However, she forgot to mention four things: 

(a) that time on task is measured combined across student and advisor, 

(b) that I would spend almost every day of the next five years pursuing the goal of doing half as much work as she does, 

(c) that this goal was obviously unattainable, and 

(d) that it would still be worth every minute of trying.
General Introduction

Attitudes, or evaluations of entities along a positive–negative continuum (e.g., “I like others” or “I prefer Federer to Djokovic”), are central to structuring affect, behavior, and cognition. As such, research on attitudes has been a crucial endeavor within social psychology since the very inception of the field (Allport, 1935; Allport & Schanck, 1936; LaPiere, 1934; Thurstone, 1928; 1931). After decades of work relying on self-report measures to index evaluative mental content, since the 1980s much of attitude research has been guided by the recognition that evaluative representations can be activated automatically upon encountering a stimulus (Bargh, 1989; Devine, 1989; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Greenwald & Banaji, 1995). These automatically activated evaluative representations, termed implicit attitudes, are usually measured using response interference tasks (Greenwald, McGhee, & Schwartz, 1998), as opposed to explicit attitudes, which are measured using self-report.

Associative and propositional theories of implicit evaluation

Beyond these obvious and relatively superficial differences in measurement, it has been one of the central theoretical questions guiding implicit social cognition research over the past three decades whether explicit and implicit attitudes also differ from each other in more profound ways (Cone, Mann, & Ferguson, 2017; De Houwer, 2014; 2018; De Houwer & Hughes, 2016; Gawronski & Bodenhausen, 2006; McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004). This debate has been dominated by two, diametrically opposed positions: associative theories (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) versus the more recent propositional account (De Houwer, 2014; 2018; De Houwer & Hughes, 2016). The first set of these theories, usually labeled associative, posits profound differences between explicit and implicit attitudes in terms of
(a) the informational inputs to which they respond, (b) the unfolding of the learning processes that give rise to them, and (c) the evaluative representations that are created as a result. By contrast, propositional theories claim that explicit and implicit attitudes (a) respond to the same informational inputs, (b) arise as a result of similar learning processes, and (c) are subserved by the same evaluative representations.

More specifically, when it comes to inputs to evaluative learning, dominant associative theories of implicit evaluation (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) hypothesize that implicit attitudes should preferentially respond to experiential learning, specifically to stimulus pairings experienced in the environment. As such, evaluative conditioning (Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010; Levey & Martin, 1975; Martin & Levey, 1978), i.e., repeated pairings of an initially neutral conditioned stimulus (CS) with an intrinsically positive or negative unconditioned stimulus (US), is posited to be the primary modality via which change in implicit evaluations can be achieved. In line with this idea, numerous investigations have provided robust evidence that implicit evaluations of both novel and preexisting stimuli can change as a result of evaluative conditioning (Gibson, 2008; Grumm, Nestler, & Collani, 2009; C. J. Mitchell, Anderson, & Lovibond, 2003; Olson & Fazio, 2001; 2002; 2006; Prestwich, Perugini, Hurling, & Richetin, 2009). By contrast, according to the same theories, explicit evaluations should shift preferentially, or perhaps exclusively, in response to language-based information, such as verbal messages differing in levels of persuasive content (Petty & Briñol, 2010; Wood, 2000).

By contrast, propositional theories (De Houwer, 2014; 2018; De Houwer & Hughes, 2016) do not assign a special role to stimulus pairings experienced in the environment in implicit attitude change. Rather, they posit that evaluative conditioning is one of many learning modali-
ties in the face of which implicit evaluations can shift and that other kinds of information should also be able to serve as inputs to evaluative learning. Notably, according to propositional theories, purely verbal manipulations, such as informing participants of upcoming stimulus pairings without actual exposure should also be effective (De Houwer, 2006). In line with this idea, it has been demonstrated numerous times that purely language-based interventions that do not involve any personal experience with valenced stimuli are able to modulate implicit attitudes (De Houwer, 2006; Gast & De Houwer, 2013; Van Dessel, De Houwer, Gast, & Smith, 2015b; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a; Van Dessel, Gawronski, Smith, & De Houwer, 2017a; Van Dessel, Mertens, Smith, & De Houwer, 2018). Results of this kind are difficult to reconcile with associative theories under which such kinds of learning should remain completely ineffective when it comes to changing implicit evaluations. By contrast, the propositional perspective is able to account for both sets of results demonstrating the effectiveness of (a) evaluative conditioning and (b) purely verbal interventions in shifting implicit attitudes by arguing that exposure to either kind of information gives rise to propositional inferences about the valence of attitude objects (see below).

Beyond the kind of information that can serve as input to evaluative learning, associative and propositional theories also differ in their assumptions about the manner in which evaluative learning processes giving rise to implicit attitudes should unfold. Specifically, associative theories posit that implicit attitudes should change in a slow, incremental, and stimulus-driven manner (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004). By contrast, under the propositional account (De Houwer, 2014; 2018; De Houwer & Hughes, 2016), implicit attitudes are subject to updating in a quick, sudden, and inferential manner. Again, the bulk of the evidence on this issue seems to favor propositional theories: It has been
demonstrated numerous times that implicit attitudes can shift quickly, without a need for a protracted learning experience (Blair, 2002; Lai et al., 2014). Moreover, in line with the idea that implicit attitude change requires inferences about the attitude object and cannot unfold in a purely bottom-up and stimulus-driven manner, it is well-established that participants’ ability to report the stimulus contingencies to which they have been exposed is a major moderator of implicit evaluative learning effects (Van Dessel, De Houwer, & Gast, 2015a; Van Dessel, De Houwer, Roets, & Gast, 2016b; Van Dessel, Mertens, Smith, & De Houwer, 2017b). On the other hand, whether implicit attitude change in the absence of conscious awareness is at all possible is highly contested (Greenwald & De Houwer, 2017; Hödgen, Hütter, & Unkelbach, 2018; Hütter, Sweldens, Stahl, Unkelbach, & Klauer, 2012).

Finally, associative (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) and propositional (De Houwer, 2014; 2018; De Houwer & Hughes, 2016) theories also differ from each other in the mental representations posited to underlie implicit evaluations. Specifically, according to associative theories, implicit evaluations are subserved by unqualified links between conceptual nodes in long-term memory (e.g., OTTERS–GOOD, DJOKOVIC–BAD). By contrast, explicit evaluations are assumed to emerge from information that is represented in propositional form (e.g., “I like otters” or “I’m not a fan of Djokovic”). On the other hand, propositional theories hypothesize that the same propositional representations (e.g., “I like otters” or “I’m not a fan of Djokovic”) underlie both explicit and implicit evaluations. As such, explicit and implicit evaluations are hypothesized to differ from each other not in underlying mental representations but rather in the automaticity with which the same kinds of mental representation are being activated on explicit versus implicit measures.
Surprisingly little research has directly investigated whether implicit attitudes are represented in associative (e.g., Otters–Good) or propositional (e.g., “I like otters”) format, presumably in part because it is not entirely clear what patterns of empirical data might be able to conclusively arbitrate between these two accounts (but for some initial ideas see Kurdi & Dunham, in prep, described in the General Conclusion chapter). However, it seems clear that if the representations underlying explicit versus implicit evaluations are fundamentally different, then there is little reason to expect congruence between implicit and explicit evaluations to emerge. On the other hand, if explicit and implicit evaluations are subserved by the same underlying representations, then they should generally be consistent with each other. Of course, associative theorists may argue that if associative (e.g., Otters–Good) and propositional (“I like otters”) representations imply the same valenced judgments, explicit and implicit evaluations should converge. Conversely, propositional theorists may posit that explicit and implicit evaluations can diverge from each other when different sets of evaluative propositions are being activated effortfully on explicit measures versus automatically on implicit measures. However, associative theories are conspicuously silent on why fundamentally different learning processes should give rise to associative and propositional representations with the same evaluative content. Conversely, propositional theories are conspicuously silent on why certain propositions can be activated automatically, whereas others cannot. And, in fact, explicit and implicit attitudes have been repeatedly found to be correlated, but usually not fully redundant, with each other (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005; Nosek, 2005), thus contradicting both theoretical positions.

To summarize, existing associative theories of implicit evaluation (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) are extremely difficult to reconcile with available evidence on the acquisition and change of implicit attitudes. First, in terms of
the information that can serve as input to evaluative learning, associative theories posit the pri-
macy of direct experience with stimulus pairings in the environment; and yet, in empirical inves-
tigations, other kinds of information, including purely language-based interventions, have been
repeatedly demonstrated to have impact. Second, when it comes to the manner in which evalua-
tive learning is assumed to unfold, associative theories argue that implicit evaluations should be
updated in a slow, incremental, and stimulus-driven manner; and yet, empirical investigations
have provided ample evidence for quick learning based on propositional inference. Third, under
associative theories, implicit evaluations are hypothesized to be subserved by associative repre-
sentations that are fundamentally different from the propositional representations underlying ex-

dplicit evaluations; and yet, more often than not, the two are correlated with each other and can
respond similarly to the same manipulations.

On the other hand, propositional theories of implicit evaluation (De Houwer, 2014; 2018;
De Houwer & Hughes, 2016) are able to account for several aspects of the data on evaluative
learning that associative theories cannot explain. First, propositional theories allow for implicit
attitudes to shift in response to purely verbal information, and the data are in line with this posi-
tion. Second, propositional theories allow for implicit evaluations to shift quickly and via propo-
sitional inference, and, again, the data are in line with this position. Third, propositional theories
allow for overlap between explicit and implicit evaluations given that both are posited to be sub-
served by the same (propositional) representations. As such, it seems that propositional theories
provide a considerably better fit to the data than do associative theories. However, as mentioned
above, propositional theories appear not particularly well-equipped to explain dissociations be-
tween explicit and implicit attitudes; moreover, as discussed below and in more detail in the
General Conclusion chapter, they are often not sufficiently specific to falsify.
The design and main findings of the present project

Given their ability to inform about the very nature of the construct, the learning processes that give rise to implicit evaluations are of crucial theoretical interest to the basic science of social cognition. Moreover, once the basic mechanisms via which implicit evaluations are acquired and subsequently updated are better understood, such deeper understanding may form the basis of theoretically guided interventions for achieving long-lasting change in implicit evaluations—an endeavor that, so far, has remained remarkably unsuccessful (Forscher, Mitamura, Dix, Cox, & Devine, 2017; Lai et al., 2016). As discussed above, existing associative approaches to implicit evaluation fail to account for most aspects of empirical data on acquisition and change. By contrast, existing propositional approaches appear to have considerably better explanatory power; however, as explained below, they are (a) silent on a whole host of issues central to implicit attitude acquisition and change and (b) difficult to reconcile with certain empirical data. As such, the present dissertation relies on a wider set of empirical findings (Olsson & Phelps, 2004; 2007) and theoretical approaches (Sutton & Barto, 1998), drawn from the literature on human learning and memory in general and thus going beyond both existing associative and existing propositional theories, in attempting to understand when, how, and why implicit evaluations change.

Paper 1 (Kurdi & Banaji, 2017, *J Exp Psychol Gen*) investigated the separable and joint effects of evaluative conditioning, referred to as repeated evaluative pairings (REP), and verbal information about upcoming stimulus pairings, referred to as evaluative statements (ES), in shifting implicit attitudes. Prior to publication of this paper, it had already been well-established that both REP (Gibson, 2008; Grumm et al., 2009; C. J. Mitchell et al., 2003; Olson & Fazio, 2001; 2002; 2006; Prestwich et al., 2009) and ES (De Houwer, 2006; Gast & De Houwer, 2013; Gregg, Seibt, & Banaji, 2006; Ranganath & Nosek, 2008; Van Dessel et al., 2018; Van Dessel, De
Houwer, Gast, & Smith, 2015b; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a; Van Dessel, Gawronski, Smith, & De Houwer, 2017a) have the ability to shift implicit evaluations; however, the relative effectiveness of REP versus ES as well as their joint influence had never been investigated. As such, Paper 1 drew upon a similar design already implemented in the context of Pavlovian conditioning (Olsson & Phelps, 2004; 2007) to probe both (a) the kinds of information that (preferentially) serve as inputs to the acquisition of implicit evaluations and (b) the nature of the learning process via which implicit evaluations are acquired.

Paper 1 yielded a perhaps surprising but robust set of results: (a) evaluative statements were always at least as effective and in several studies significantly more effective than repeated evaluative pairings in shifting implicit attitudes, and (b) the joint effects of the two learning modalities never significantly outperformed evaluative statements in isolation. As is obvious, these results are difficult if not impossible to reconcile with existing associative theories of implicit evaluation given that they demonstrate (a) the superiority of purely verbal information over direct experience with valenced stimuli in shifting implicit attitudes and (b) redundancy between the two kinds of input in creating change, thus providing evidence for the idea of shared representations. By contrast, propositional theories can easily explain the latter finding: If both stimulus pairings and instructions about stimulus pairings give rise to propositional inferences about the targets (e.g., “Laapians are good and Niffies are bad”), experience with stimulus pairings should not add any value over and above mere verbal instructions. However, propositional theories do not offer an immediate explanation for why verbal instructions should be more effective than exposure to stimulus pairings in shifting implicit evaluations.

In view of the findings reported in Paper 1, Paper 2 (Kurdi & Banaji, 2019, *J Pers Soc Psychol*) addressed three issues: (a) the process by which repeated evaluative pairings give rise
to changes in implicit evaluation (Study 1), (b) the process by which verbal information and repeated evaluative pairings are combined to create evaluative representations (Study 2), and (c) the durability of evaluative representations created by repeated evaluative pairings, evaluative statements, and their combination (Studies 3–4). We found that REP were able to influence implicit evaluations after as few as four stimulus pairings, with additional stimulus pairings not providing any further benefit. Moreover, verbal information about the diagnosticity of stimulus pairings (i.e., stimulus pairings described as randomly generated versus reflective of the underlying character of the targets) has been found to modulate implicit learning effects. Finally, attitude acquisition from repeated evaluative pairings was found to be more durable than attitude acquisition from evaluative statements, but only when the former were presented in isolation and were not preceded by any verbal information.

The results of Studies 1 and 2 clearly contradict the predictions of existing associative theories of implicit evaluation (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) according to which learning should have unfolded slowly and without influence from verbal information. The finding that the effects of direct experience with stimulus pairings were more durable than the effects of a purely verbal intervention (Studies 3–4) seems in line with associative theories; however, these theories appear unable to explain why verbal information presented before exposure to stimulus pairings should eliminate the advantage of experience-based learning. By contrast, propositional theories (De Houwer, 2014; 2018; De Houwer & Hughes, 2016) are not irreconcilable with any of these findings given that they allow for (a) implicit evaluations to be updated via quick inferential learning, (b) experience-based and verbal information to be combined in giving rise to implicit evaluations, and (c) general theories of episodic memory to be incorporated into theories of implicit evaluation.
Specifically, under propositional theories, participants have the ability to make quick propositional inferences from data presented to them in the form of stimulus pairings and, as such, if they are able to make the intended inference about the valence of targets after as few as four stimulus pairings, then this inference should be reflected by implicit measures of evaluation. Second, propositional theories do not privilege experience with stimulus pairings over verbal instructions in shifting implicit evaluations and, therefore, the fact that both kinds of information are combined in implicit evaluative learning should not surprise from their perspective. Finally, because propositional theories do not posit separate memory representations to be underlying explicit versus implicit evaluation, they are able to draw on findings regarding episodic memory for experienced versus merely described events (Baggett, 1979; Baggett & Ehrenfeucht, 1983; Beentjes & van der Voort, 1991; Larsen & Plunkett, 1987; Toglia, Shlechter, & Chevalier, 1992) in explaining learning effects indexed by implicit measures of evaluation. Specifically, although immediate retention has been found to be superior for verbally described compared to directly experienced events, the opposite effect has been established when a delay was implemented between the learning and test phases of the experiment. This episodic memory perspective is generally in line with the present findings, suggesting that personally experienced events (e.g., exposure to stimulus pairings) may be more memorable in the long term because they require participants to make inferences about the experienced events themselves, whereas verbal descriptions of those events (e.g., evaluative statements) already provide those inferences to participants in propositional form.

However, notably, Paper 2 also provides some indication that current associative and propositional theories may both be missing important elements of implicit attitude acquisition altogether. For instance, as explained in more detail below, neither of them address the crucial
distinction between temporary shifts in implicit evaluation versus genuine long-term change. Moreover, importantly, along several important dimensions, currently available associative and propositional theories both seem to lack the level of computational specificity that would make them easily falsifiable.

Specifically, with regard to Study 1, even though current associative theories of implicit evaluation (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) posit that evaluative learning should unfold slowly, this need not be the case under associative approaches to human learning in general (Rescorla & Wagner, 1972; Rumelhart, Hinton, & Williams, 1986; Sutton & Barto, 1998). Specifically, computationally well-specified instantiations of associative learning tend to contain a learning rate parameter. If this learning parameter is high and evaluative information is consistent, learning can unfold quickly, and no existing associative theory of implicit evaluation accounts for this possibility. Conversely, under a propositional account it is unclear why a larger number of stimulus pairings (e.g., 24 as opposed to four) should not lead to less uncertainty about the inferences that one can make from the data and, correspondingly, to stronger implicit evaluations.

In the same vein, although the propositional perspective has the ability account for interactions between experience-based and verbal information in giving rise to implicit evaluations (Study 2), it remains silent on the computations by which such integration should occur. Similarly, current propositional theories of implicit evaluation are unable to explain why information on diagnosticity should modulate implicit evaluations if presented before, but not if presented after, exposure to stimulus pairings. Finally, even though propositional theories allow for evidence on episodic memory to be incorporated into the explanation of implicit evaluative learning effects, they did not a priori predict the pattern of results that emerged in Studies 3–4. Importantly, nei-
ther current associative nor current propositional theories address the issue of the temporal specificity, and more generally the context-specificity, of implicit evaluations, although the distinction between temporary shifts in accessibility versus genuine conceptual change seems central to this domain: Implicit evaluations have repeatedly been found resistant to long-term change (Forscher et al., 2017; Lai et al., 2016) even in the face of robust evidence on temporary malleability (Blair, 2002; Lai et al., 2014).

To a large degree, Paper 3 (Kurdi, Gershman, & Banaji, 2019, *Proc Natl Acad Sci USA*) has been designed to remedy the lack of computational specificity inherent to current associative (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) and propositional (De Houwer, 2014; 2018; De Houwer & Hughes, 2016) approaches to implicit evaluative learning. Specifically, after a large number of investigations probing the effects of evaluative conditioning on implicit attitudes (Gibson, 2008; Grumm et al., 2009; C. J. Mitchell et al., 2003; Olson & Fazio, 2001; 2002; 2006; Prestwich et al., 2009), in this paper we provide initial evidence for the idea that implicit evaluations of stimuli are also responsive to reinforcement learning, i.e., experience with motivationally relevant consequences of interacting with those stimuli. Moreover, importantly, we show that implicit evaluations are updated via model-free, but not via model-based, reinforcement learning. By contrast, in line with previous work (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, Niv, & Dayan, 2005), we found explicit evaluations to reflect a mixture of model-free and model-based processes. Importantly, along with providing a computationally tractable framework for investigating implicit attitude acquisition and generating novel empirically testable predictions, this finding offers an immediate explanation for why explicit and implicit attitudes are usually found to be correlated but not redundant: Whereas the latter arise exclusively from model-free processes, the former reflect both model-free and model-
based learning. Moreover, among numerous other advantages, a reinforcement learning framework can account for (a) slow or fast learning and (b) the context-specificity of evaluative learning effects in a computationally tractable way.

**Theoretical implications of the present project**

The three papers included in this dissertation have been designed to inform and constrain theories of implicit attitude acquisition and change in an attempt to advance basic understanding of the nature of implicit evaluations. The results that have emerged from these papers have implications for how (a) informational inputs to implicit evaluative learning, (b) the nature of the learning process giving rise to implicit evaluations, and (c) the resulting mental representations should be appropriately characterized.

Current associative and propositional theories of implicit evaluation differ in terms of their claims about the kinds of information that should serve as inputs to the updating of implicit evaluations. Specifically, under the associative approach (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004), implicit evaluations should be primarily, or even exclusively, responsive to direct experience with valenced stimuli. By contrast, the propositional approach (De Houwer, 2014; 2018; De Houwer & Hughes, 2016) posits that both direct experience and purely symbolic learning should shift implicit evaluations. Papers 1–2 are clearly at odds with the associative approach given that they found significant changes in implicit attitudes as a result of both repeated evaluative pairings of targets with valenced images and purely verbal descriptions of such stimulus pairings. However, prior work (Lai et al., 2014) has shown that implicit evaluations do not respond to all kinds of verbal information. Moreover, Paper 3 demonstrates that implicit evaluations shift in the face of certain kinds of direct experience (allowing for model-free learning) but not in the face of others (requiring a model of the environ-
ment to achieve change). Overall, the current work suggests that, given their undue emphasis on direct experience, associative theories are inadequate in their characterization of informational inputs to implicit evaluative learning. At the same time, propositional theories do not seem sufficiently specific. Rather than dwelling on the distinction of experience-based versus language-based learning, future research should place more emphasis on the kinds of associative and propositional information that are capable of shifting implicit evaluations. Some initial ideas on how to do so are presented in the General Conclusion chapter.

When it comes to the learning processes posited to give rise to change in implicit evaluations, the results of Papers 1–3 are exceedingly difficult to reconcile with associative theories (Rydell & McConnell, 2006; E. R. Smith & DeCoste, 2000; Strack & Deutsch, 2004) given that these theories posit that such change should happen slowly and incrementally. In fact, across all three papers, we have observed quick shifts in implicit evaluation, either as a result of language-based learning (Paper 1) or exposure to as few as four valenced stimuli (Papers 2–3). At the same time, the present results point to two important, as yet not sufficiently explored, aspects of implicit attitude change. First, learning effects measured immediately after exposure to evaluative information cannot be expected to persist into the future and, importantly, learning effects emerging from different interventions can show different rates of decay over time (Paper 2). Second, under a reinforcement learning perspective (Paper 3), changes in evaluative representations should be expected to be tied to the particular context (i.e., state) in which learning took place. The implications of this aspect of the reinforcement learning approach are discussed in more detail in the General Conclusion.

Finally, with regard to the mental representations underlying responding on implicit measures of evaluation, the present project suggests that the distinction between model-free and
model-based representations may be more instructive than the distinction between associative and propositional representations: Central to model-free reinforcement learning approaches is the idea that such learning creates highly compressed evaluative representations that assign a single scalar value to each relevant stimulus or action. Under a classic model-free reinforcement learning approach, such compressed value representations should emerge only from direct experience with the environment. However, Papers 1 and 2 (Lai et al., 2014) demonstrate that implicit evaluations are also responsive to certain, but not all, kinds of verbal information. Specifically, reexamining the results of Lai et al. (2016) from the perspective of model-free versus model-based learning leads to the conclusion that only interventions (a) allowing for model-free learning or (b) involving the simplest possible kind of model of the form $P(\text{good} \mid A) = P(\text{bad} \mid Y) = 1$ had impact (see also Paper 1). By contrast, interventions involving complex mental models remained ineffective. Taken together, these different sources of evidence suggest that only highly compressed value representations, be they action–value pairs or simple propositions, may be able to influence responding on implicit measures of attitude. This idea, along with additional implications of the findings of Paper 3, is further developed in the General Conclusion.

Taken together, the present papers suggest that an approach steeped in prior work on Pavlovian learning (Olsson & Phelps, 2004; 2007), episodic memory for verbally described versus personally experienced events (Baggett, 1979; Baggett & Ehrenfeucht, 1983; Beentjes & van der Voort, 1991; Larsen & Plunkett, 1987; Toglia et al., 1992), and human reinforcement learning (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman, Markman, & Otto, 2014; Otto, Gershman, Markman, & Daw, 2013) can be highly informative in characterizing the learning processes giving rise to implicit evaluations. Specifically, going beyond existing associative and propositional perspectives, we found that implicit evaluations reflect highly compressed value
representations that can be created \((a)\) either via personal experience with motivationally relevant stimuli, purely verbal interventions, or a combination of both, and \((b)\) often rather quickly, depending on the parameters of the task. By contrast, less compressed value representations requiring complex online computations of value may not be able to be activated automatically on implicit measures. Finally, the initial source of compressed value representations may modulate their durability, with personal experience having more lasting impact than purely verbal information.
Paper 1 • Repeated Evaluative Pairings and Evaluative Statements: How Effectively Do They Shift Implicit Attitudes?
Abstract

6 experiments, involving a total of 6,492 participants, were conducted to investigate the relative effectiveness of *repeated evaluative pairings* (REP; exposure to category members paired with pleasant or unpleasant images), *evaluative statements* (ES; verbally signaling upcoming pairings without actual exposure), and their combination (ES + REP) in shifting implicit social and non-social attitudes. Learning modality (REP, ES, and ES + REP) was varied between participants and implicit attitudes were assessed using an Implicit Association Test (IAT). Study 1 (*N* = 675) used fictitious social groups (NIFFs and LAAPs), Study 2 (*N* = 1,034) used novel social groups (humans with long vs. square faces), Study 3 (*N* = 1,072) used nonsocial stimuli (squares vs. rectangles), and Study 4 (*N* = 848) and Study 5 (*N* = 958) used known social groups (young vs. elderly; American vs. foreign). ES were more effective than REP and no less superior than ES + REP in producing implicit attitude change. Results were robust across social and non-social domains and for known and novel groups. Study 6 (*N* = 1,905) eliminated time on intervention, levels of construal, and expectancy effects as possible explanations for these findings. Associative theories of implicit evaluation posit that implicit attitudes should shift piecemeal over time; yet, in these experiments, one-shot language-based learning led to larger shifts in implicit attitude than exposure to stimulus pairings. Moreover, the redundancy observed in REP + ES suggests that attitude acquisition from repeated pairings and evaluative instructions may rely on shared mental representations.
Introduction

Learning about conspecifics is among the most fundamental and momentous tasks performed by humans. In order to decide whether a person is friend or foe, and whether to approach or avoid them, access to the products of such learning in memory is essential. Such decisions, moreover, often depend upon immediate responses in complex and rapidly shifting social environments. Accordingly, the process of social learning requires maintenance of a delicate balance between holding on to an initial or even well-learned evaluation and adaptively changing that evaluation when novel, and potentially contradictory, information becomes available. Learning and updating are therefore critical to achieving social accuracy and even to survival.

Learning in the social world, at its core, involves the formation of attitudes, defined as “a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor” (Eagly & Chaiken, 1993, p. 1). Fazio and colleagues (Fazio et al., 1986; Fazio, Jackson, Dunton, & Williams, 1995), Bargh and colleagues (Bargh, 1989; Bargh, Chaiken, Govender, & Pratto, 1992), and Greenwald and Banaji (1995) focused attention on the implicit nature of social evaluation. In this framework, implicit attitudes are relatively less consciously accessible, less controllable, and more automatic than their explicit counterparts (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Greenwald & Banaji, 1995; Kim, 2003). Moreover, they uniquely guide behavior in a wide range of situations (Devine, 1989; Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997; Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

But how are implicit attitudes acquired and how do they shift in the face of new information? Social attitudes and modalities of acquisition and change can be categorized in myriad ways. A recent distinction that has emerged is that between associative (Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Rydell & McConnell, 2006) and propositional (De
Houwer, 2009; 2014; C. J. Mitchell, De Houwer, & Lovibond, 2009) processes of learning and change. The present research relies on the distinction between associative and propositional learning to ask a basic question that is yet to be investigated directly: Which one of these two learning modalities is more effective in shifting implicit attitudes? Answering this question can teach us about the nature of attitudes in their most fundamental form and has the potential to lead to method development for changing attitudes via theoretically guided interventions.

The nature of the mental representations underlying evaluative learning has been contested (De Houwer, 2009; 2014; Gawronski & Bodenhausen, 2006; 2011), and many proposals have emerged from this debate. We are in agreement with De Houwer (2009) that conceptual confusion may be avoided by characterizing methods of learning in terms of actual operations rather than using more abstract labels. Far too often, abstract psychological constructs, including the most studied and central ones (such as perception or memory), are used to refer to a host of diverse phenomena and can map onto a range of different operationalizations, thus obscuring meaningful, and possibly unintended, differences across possible instantiations. To avoid fuzzy thinking about the learning mechanisms introduced here, we created novel labels that hew closely to the actual learning manipulations that were used in the experiments.

The present research compares the separate and combined effects of two different forms of learning. The first learning procedure will be referred to as repeated evaluative pairings (REP), a learning modality commonly characterized as associative in nature. We define repeated evaluative pairings as “pairing [a] stimulus [the conditioned stimulus; CS] with other positive or negative stimuli [unconditioned stimuli; US]” (De Houwer, Thomas, & Baeyens, 2001). This learning modality, also known as evaluative conditioning (Baeyens & De Houwer, 1995; De Houwer et al., 2001; Hofmann et al., 2010), is rooted in well-known associative learning proce-
dures such as classical conditioning (Pavlov, 1927). Using this method, learning has been observed in numerous species, including goldfish, quail, hamsters, rats, dogs, cats, pigeons (Rescorla & Holland, 1982) and humans (Delgado, Olsson, & Phelps, 2006; C. J. Mitchell et al., 2009). In such paradigms, learning is assumed to occur via the transfer of valence between an intrinsically positive or negative stimulus and an initially neutral stimulus or action. Specifically, in evaluative conditioning, changes in the liking of a stimulus or its subjective positivity are thought to result from pairing that stimulus with other, positive or negative, stimuli (De Houwer et al., 2001).

A second learning procedure used in these experiments will be referred to as *evaluative statements (ES)*, a learning modality commonly characterized as propositional in nature. Unlike nonhuman animals that lack the capacity for language or higher-order symbolic logic (Penn, Holyoak, & Povinelli, 2008), humans have the unique ability to learn about their environment solely through language.¹ For example, when one person says to another, “The Falaknuma Palace is the most beautiful hotel I have ever stayed in,” the assumption is that the concept “Falaknuma Palace” is now linked to the positive trait “beautiful,” and an attitude may form based on the recipient’s cognitive response to the statement. In contrast with REP-based learning procedures, which apart from minor procedural variations are highly similar to each other, evaluative statements can be manifold. Classic work on attitudes used messages that varied in persuasive power to shift evaluations of varied attitude objects via high-level, relatively elaborative, reasoning processes (for reviews see Petty & Briñol, 2010; Wood, 2000). Other kinds of evaluative statements remain closer to REP-based learning in that they rely on mere signaling of upcoming repeated evaluative

¹ This should not be taken to suggest that language cannot play a role in repeated evaluative pairing paradigms involving human participants. However, both traditionally and in our project, such paradigms have used quite uninformative verbal instructions with almost sole focus on exposure to stimulus pairings.
pairings. For instance, Olsson and Phelps (2004; 2007) demonstrated that human participants need not be exposed to actual shock in order for fear conditioning to occur; it is sufficient to simply inform participants of upcoming stimulus–shock pairings to obtain a learning effect of similar magnitude to presenting actual shock. The power of language alone, in the absence of direct experience, to create learning is intriguing and uniquely human.

**Shifting implicit attitudes via repeated evaluative pairings and evaluative statements**

Olson and Fazio (2001) provided a clear demonstration of the effectiveness of repeated evaluative pairings (REP) in the domain of implicit attitude change. Olson and Fazio showed participants several hundred images, some of them in isolation and others in pairs. Critically, they embedded 40 US–CS pairs among the stimuli. Two initially neutral Pokémon characters served as CS, and valenced words (e.g., “awesome” vs. “terrible”) and images (e.g., of ice cream vs. a cockroach) served as US. At test, participants exhibited implicit preference in favor of the Pokémon character paired with positive US over the Pokémon character paired with negative US during the learning phase. The finding that REP influence implicit attitudes has been replicated numerous times (Gibson, 2008; Grumm et al., 2009; C. J. Mitchell et al., 2003; Olson & Fazio, 2006; Prestwich et al., 2009). The robustness of this result may not be surprising given the abundance of evidence for associative learning and its pride of place in many theories of attitude change (Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Strack & Deutsch, 2004). Such findings are, however, also compatible with a propositional theory of implicit evaluation (De Houwer, 2014), which posits that the implicit attitude created as a result of exposure to stimulus pairings is mediated by propositional, rather than associative, mental representations. Although the stimuli may be presented in physiological or non-linguistic form (e.g., shock or photographs), the assumption is that propositions about CS–US relations can also be formed on
the basis of exposure to stimulus pairings. When receiving repeated CS–US pairings, humans can infer the underlying relationship and represent that information in propositional form.

A separate body of work has provided ample evidence that it is possible to shift implicit attitudes by using mere *evaluative statements (ES)*. For instance, Gregg, Seibt, and Banaji (2006) created strong novel implicit attitudes toward initially neutral fictitious social groups using vignettes that described one of the social groups as positive and the other social group as negative (see also Ranganath & Nosek, 2008) and, in another condition, by instructing participants to suppose that group X has positive traits and group Y has negative traits. In these studies, change in implicit attitudes was created on the basis of information that was highly diagnostic of the targets (see Cone & Ferguson, 2015). De Houwer (2006) showed the presence of such learning using an even more minimal manipulation that is crucial to the present studies. He simply informed participants that nonwords (Experiment 1) or names from fictitious social groups (Experiment 2) would be paired with known positive or negative words in an upcoming evaluative conditioning task. Although participants were never exposed to any actual pairings between CS and US, novel implicit attitudes toward the initially neutral stimuli were found to have formed in line with this most minimal verbal instruction. Together, these studies demonstrate that ES are sufficient to shift implicit attitudes (see also Gast & De Houwer, 2013; Van Dessel, De Houwer, Gast, & Smith, 2015b). Such a result may be surprising for at least two reasons. First, implicit attitudes are often recalcitrant: A large-scale comparative investigation showed the failure of many learning interventions to produce any change in implicit attitudes, although some did have impact (Lai et al., 2014). Second, the finding that one-shot language-based attitude induction modalities have the power to shift implicit attitudes contradicts strict theories of associative learning positing that implicit attitudes form as a result of exposure to stimulus associations in one’s environment over
extended periods of time. The result obtained by De Houwer (2006) that even minimal propositional statements can shift implicit evaluation is, however, compatible with single-process theories suggesting that REP and ES are two ways of creating the same underlying propositional representation.

**Comparing REP and ES**

It is established that, compared to a neutral baseline, implicit attitudes can shift either on the basis of REP or ES. But do REP and ES give rise to implicit attitudes of comparable magnitude? And does the joint effect of REP and ES exceed their respective separate effects? The answer to these important questions is not known. In spite of an impressive number of studies exploring the effects of REP, on the one hand, and the effects of ES, on the other hand, no systematic comparison of the two learning modalities has been undertaken, even though such a comparison would provide evidence about their relative power to influence attitudes, thus informing and constraining theories of implicit evaluation. The conspicuous lack of a direct comparison might have to do with the fact that the overwhelming majority of studies on implicit attitude acquisition and change have been conducted within the confines of two major theories (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986), neither of which favor a comparative approach. As we discuss below, projects steeped in associative models (Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Strack & Deutsch, 2004) have mostly attempted to find a match between cognitive systems and learning modalities (ES change explicit attitudes vs. REP change implicit attitudes). On the other hand, work relying on propositional theories (De Houwer, 2009; 2014; Hughes, Barnes-Holmes, & De Houwer, 2011; C. J. Mitchell et al., 2009) has focused on demonstrating the—at least initially—surprising result that implicit attitude change via ES is possible. Neither theoretical perspective has had a stake in directly contrasting or combining REP and ES.
in the context of implicit evaluation even though both would gain from the results of such a test, whatever the outcome. Such a comparison as well as a test of the joint influence of the two learning modalities on implicit attitudes is the primary purpose of the present research.

In a sense, this project sought to do for the study of implicit social attitudes what Olsson and Phelps (2004; 2007) accomplished for fear conditioning, namely a direct comparison of REP and ES. Accordingly, in our studies, we induced the same implicit attitude toward the same attitude object within the same experiment with random assignment of participants to learning modality conditions, each of which involved either REP or ES. Some participants learned about an attitude object exclusively via exposure to target stimuli paired with valenced images, whereas other participants learned about the same attitude object using only verbal instructions and without any exposure to stimulus pairings.

Based on the extensive literature available we expected that both attitude induction modalities would create learning compared to baseline. The main focus here was to directly test the relative strength of the learning effects produced by each intervention. In terms of the comparison between REP and ES, the studies reported here had three possible outcomes: Both REP and ES could produce learning at comparable levels; REP might outperform ES; or ES might outperform REP. An additional condition involving joint presentation of both types of learning also allowed us to test whether the two together produce more robust attitude change than each one by itself. As discussed in more detail below, if the underlying representation is common to both interventions, then the two together need not be more robust than the single effect of the stronger of the two.

There are at least two lines of evidence that favor the superiority of REP. First, according to theories of associative learning and attitude change (Gawronski & Bodenhausen, 2006; Rydell
& McConnell, 2006; Strack & Deutsch, 2004), unlike explicit attitudes that can be updated very rapidly, implicit attitudes shift gradually over time. If this is the case, the REP condition, which involves multiple stimulus pairings, should be more effective than ES, which relies on a verbal instruction to expect particular pairings, with no actual pairings presented. In fact, certain kinds of ES, such as instructing participants to adopt an egalitarian mindset, do not change implicit race attitudes at all (Lai et al., 2014). Second, when REP and ES are contradictory, REP usually outperform ES in their effect on implicit attitudes (but see Peters & Gawronski, 2011; Zanon, De Houwer, Gast, & Smith, 2014). For instance, Gregg, Seibt and Banaji (2006; Study 3) found that informing participants that the REP to which they were exposed are invalid (an ES-based manipulation) does not have an impact on implicit evaluations. Similar results have been obtained using other paradigms involving supraliminally presented behavioral information vs. subliminally presented valenced primes (Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006a), affirmative vs. negative sentences (DeCoster, Banner, Smith, & Semin, 2006), and REP vs. ES containing causal information (Moran & Bar-Anan, 2013). These studies suggest that even in the absence of contradictory information, REP might be superior to ES.

On the other hand, single-process theories of implicit evaluation (De Houwer, 2009; 2014; Hughes et al., 2011; C. J. Mitchell et al., 2009) have advanced the idea that REP and ES should produce learning effects of similar magnitude. If, as posited by these theories, all human learning is mediated by the same kinds of propositional representations, neither learning modality should be inferior or superior to the other. In fact, when Gast and De Houwer (2013) induced novel implicit attitudes toward the same nonwords using ES (Experiment 2a) and REP (Experiment 2b), the magnitude of implicit attitudes was found to be highly similar across the two experiments. However, the comparison between REP and ES remains post hoc and incidental be-
cause in this project participants were not randomly assigned to conditions, and time on intervention was not fixed across studies. Nonetheless, it is possible that the effects of both attitude induction modalities might be comparable under more controlled conditions as well.

Finally, although unlikely given the research we have reviewed, it is conceivable that under the conditions created by the current experiments, ES will be superior to REP in producing change in implicit attitudes. As of now, no evidence suggesting this outcome exists and no theories would make such a prediction. Thus, the studies we conduct offer the first test that allows for the possibility of such a result to emerge. Unlike previous studies conducted under the dual-process view (e.g., Moran & Bar-Anan, 2013), our design made it possible for ES to exert their effect without the influence of countervailing information, and unlike previous studies conducted under the propositional view (Gast & De Houwer, 2013), we ensured, to the extent possible, equivalent exposure to REP and ES.

In addition to the question of relative effectiveness, we sought to explore whether attitude acquisition effects from REP and from ES are additive or redundant. In line with single-process theories of implicit evaluation (De Houwer, 2014; C. J. Mitchell et al., 2009), the effects of REP and ES on implicit attitudes should be comparable and redundant because both are mediated by the same propositional representations (e.g., “X is good and Y is bad”). Under this theory, one would expect the combination of both attitude induction modalities to produce a learning effect on par with their separable effects. On the other hand, dual-process theories of implicit evaluation might offer two different predictions depending on whether they are strictly associationist or allow for interactions between different learning modalities. Under strict associative learning theories (Rydell, McConnell, Mackie, & Strain, 2006a; Strack & Deutsch, 2004), ES should produce at most a negligible learning effect on implicit attitudes, which renders the question of addi-
tivity vs. redundancy moot. On the other hand, some dual-process theories (Gawronski & Bodenhausen, 2006) allow for interactions between the explicit and implicit system. Under such theories, the effects of both attitude induction modalities should be additive, with REP creating implicit attitudes by strengthening links between conceptual nodes in long-term memory and ES affecting implicit attitudes via a separate pathway mediated by explicit attitudes.

**Overview of the present studies**

Each of Studies 1–5 reported below consisted of a learning phase and a test phase. Based on the considerations outlined above, the learning phase included four between-participant conditions: (1) a *control condition* in which we measured participants’ implicit preferences at baseline; (2) a *REP condition* in which stimuli representing one social or nonsocial category were paired with positive images and stimuli representing another social or nonsocial category were paired with negative images over a sequence of trials; (3) an *ES condition* in which participants were informed that they would be exposed to stimulus pairings but in reality, the verbal instruction about stimulus pairings served as the sole attitude induction experience; and (4) a *combined (ES + REP) condition* in which verbal instructions were followed by actual stimulus pairings. After the learning phase, participants completed an Implicit Association Test measuring their implicit attitudes toward the two targets included in the learning phase and, finally, explicit attitudes toward the targets were measured.

Intrinsically valenced photographs (Study 4) and valenced line drawings (all other studies) served as unconditioned stimuli (US) that would produce the necessary evaluative changes in the REP condition. In order to examine the generalizability and potential moderators of our findings, attitude objects—represented by visual conditioned stimuli (CS)—were varied across studies. Study 1 used names from two initially neutral fictitious social groups; Study 2 used more
ecologically valid social stimuli (drawings of long and square faces); Study 3 used nonsocial targets (squares and rectangles); Study 4 used faces from preexisting social groups, young and elderly people; and Study 5 tested another preexisting social contrast (US vs. foreign) using photographs and drawings as CS. Finally, in Studies 6A–6C we used novel social stimuli (fictitious names and individuals with long and square faces) in order to eliminate possible confounds that may have accounted for some of the results observed in the first five studies, including time spent on intervention, levels of construal, and expectancy effects.

In these studies, some of the categories we used, such as fictitious groups and geometric shapes, might be expected to be relatively evaluatively neutral. However, as several previous studies have reported (Bargh et al., 1992; Glaser & Banaji, 1999), implicit measures can detect surprising preferences in one or another direction even when the stimuli seem to be evaluatively equal. For instance, as we show below, most participants exhibit a baseline implicit preference in favor of squares over rectangles (perhaps due to their greater symmetry) but the opposite preference for faces, rectangular over square (perhaps because they are thinner). Therefore, the conditions of greatest interest will be those where attitude objects are paired with attributes that are evaluatively opposite of the baseline attitude.

**Study 1**

Previous research has established that implicit attitudes can be shifted either using repeated evaluative pairings (REP) of stimuli (Gibson, 2008; Grumm et al., 2009; Olson & Fazio, 2001; 2006; Prestwich et al., 2009) or using evaluative statements (ES) about an attitude object (De Houwer, 2006; Gast & De Houwer, 2013; Gregg et al., 2006; Van Dessel, De Houwer, Gast, & Smith, 2015b). In Study 1, we sought to directly compare these two learning modalities in terms of their effectiveness in shifting implicit attitudes. To our knowledge, this is the first study
that directly contrasts the effectiveness of the two attitude induction modalities within the same experiment, using random assignment of participants to conditions, and fixing the time spent on each learning experience. The relative contribution of each learning modality would be difficult to gauge without such a test. Moreover, a third joint condition was included to measure the combined effects of the two learning modalities, i.e., to probe whether ES followed by REP create more robust learning of attitudes than either REP or ES in isolation. In Study 1, we used two attitudinally neutral fictitious social groups, the Laapians and the Niffians (Gregg, 2000), as target stimuli.

**Method**

**Participants and design.** 707 volunteers were recruited from the Project Implicit educational website (https://implicit.harvard.edu/implicit/) to participate in the study. Participants who did not complete the Implicit Association Test (IAT; Greenwald et al., 1998), which was the main dependent measure of the study, could not be included in subsequent analyses ($N = 22$). In line with the recommendations of Greenwald, Nosek, and Banaji (2003), participants whose response latencies were below 300 milliseconds on more than 10 percent of all IAT trials, suggesting inattentive responding, were also excluded from all analyses ($N = 10$). This resulted in a final sample size of $N = 675$ participants ($N = 446$ female, mean age = 39.24, $SD = 15.17$ years). For web-based research, involving no payment, these are excellent retention rates. Because obtaining no differences across learning modality conditions was a potentially interesting finding, we used a large sample to make sure that we had sufficient power to detect differences provided that they existed in the population. In fact, post-hoc power calculated on the basis of the effect size that we obtained was 1 to within machine precision. Participants were randomly assigned to one of four learning modalities (control vs. REP vs. ES vs. combined) and one of two congruency conditions.
learning congruent vs. incongruent with overall, not individually-based, prevailing attitudes), with both factors varied between participants. In the context of the present study, we refer to the Laapian–good condition as the congruent condition because in the control condition 98 participants showed an implicit preference in favor of Laapians over Niffians, and only 72 participants exhibited the opposite preference. This distribution does not differ from chance, \( p = .054 \) by binomial test, and congruency did not modulate the main effect of learning modality; therefore, the analyses reported below collapse across this factor. Assignment to conditions was fully random, i.e., the number of participants assigned to each condition was not fixed. 170 participants were assigned to the control condition, 154 to the REP condition, 159 to the ES condition, and 192 to the combined condition.

**Materials.** Cartoon drawings of intrinsically positive objects, including a beach, a flower, a heart, an ice cream cone, and a sun, and intrinsically negative objects, including an insect, a frowny face, a fleeing man, a snake, and an assassin, served as unconditioned stimuli (US). Names from two fictitious social groups, the Laapians and the Niffians (Gregg, 2000), served as conditioned stimuli (CS). The names were all pronounceable nonsense words consisting of three syllables that followed a certain phonological pattern. Laapians’ names ended with the syllable *lap* (e.g., Deebolap or Maasolap), whereas Niffians’ names ended with the syllable *nif* (e.g., Ibbonif or Yossanif). These stimuli had been pretested and were found to be attitudinally neutral on both explicit and implicit measures (Gregg, 2000).

**Procedure and measures**

---

2 The stimuli for Study 1 and all other studies are available for download from the Open Science Framework (OSF; [https://osf.io/jyk8c/](https://osf.io/jyk8c/)).
Overview of the procedure. The procedure consisted of a learning phase, followed by an implicit and an explicit measure of attitudes. In the learning phase, each participant was randomly assigned to one of four learning modalities (control vs. REP vs. ES vs. combined) in a between-participant design. We varied learning modality conditions between participants for two reasons. First, repeated administrations of the Implicit Association Test (IAT; Greenwald et al., 1998) are known to decrease the magnitude of the IAT effect (Nosek, Banaji, & Greenwald, 2002). Second, taking a pre-intervention IAT selectively reduces the effects of REP on implicit attitudes (Lai et al., 2014), which would have given an unfair advantage to the ES condition in the present studies. Moreover, each participant was also assigned to one of two congruency conditions, i.e., they learned either that Laapians are good and Niffians are bad (congruent condition) or that Niffians are good and Laapians are bad (incongruent condition). In order to make learning modalities fully comparable to each other, time spent on acquiring the attitude was fixed to 160 seconds across all learning conditions. Initial instructions, stimulus pairs, and evaluative statements appeared on the screen automatically for a fixed amount of time. Immediately after the learning episode, participants completed an IAT designed to measure attitudes toward the two target groups, followed by two feeling thermometer items tapping participants’ explicit attitudes toward the target groups. Finally, participants were thanked and debriefed.

Learning modality conditions

Repeated evaluative pairings (REP) condition. In this condition, the learning phase consisted of exposure to stimulus pairings. No verbal statements were used to explain the relationship between stimuli and no verbal labels were used to refer to the target groups or the US. Participants received general instructions explaining the nature of the learning task, followed by a
presentation of the full set of CS and US. Participants were then exposed to 37 trials in a standard visual–visual evaluative conditioning paradigm (Levey & Martin, 1975), i.e., both CS and US were presented in the visual modality. On each trial, one CS (a Laapian or Niffian name) and one US (a line drawing of an intrinsically positive or negative object) were presented simultaneously next to each other in the center of the screen for 2,500 milliseconds, followed by an inter-trial interval of 1,000 milliseconds, consisting of a blank screen. Within participants, target group and valence were perfectly matched with each other, i.e., a certain target group (Laapians or Niffians) was paired only with positive US and the other target group was paired only with negative US. The pairings of specific CS to specific US were randomized for each participant.

*Evaluative statements (ES) condition.* In this condition, learning occurred by presenting a verbal instruction; participants did not experience any actual stimulus pairings. Upon entering the study, participants received some general explanations describing the nature of the learning task. Crucially, participants were informed that they would see pairings of names with pictures such that one target group (e.g., Laapians) would always be paired with pictures of pleasant things and the other target group (e.g., Niffians) would always be paired with pictures of unpleasant things, followed by a presentation of the full set of CS and US. In fact, participants were not exposed to stimulus pairings.

*Combined condition.* Like in the ES condition, participants were informed that one target group would always be paired with pleasant things, whereas the other target group would always be paired with unpleasant things. Like in the REP condition and unlike in the ES condition, participants were exposed to CS–US pairings. In order to keep time spent on learning consistent

---

3 The full text of the instructions is available for download from OSF (https://osf.io/jyk8e/).
across conditions, participants in the combined condition were exposed to 20, rather than 37, conditioning trials.\(^4\)

**Control condition.** The structure of the control condition was identical to the combined condition; however, the target groups were never mentioned and participants were not exposed to any presentations of the CS. Rather, some participants were informed that they would see positive images paired with positive images and negative images paired with negative images, whereas other participants were informed that they would see positive and negative images paired with each other. Like in the combined condition, participants were subsequently exposed to pairings; however, instead of CS–US pairings, they saw 20 US–US pairings. Thus, no task-relevant learning could occur.

**Implicit attitudes.** Following the learning phase of the study, participants completed an Implicit Association Test (IAT; Greenwald et al., 1998) as the measure of implicit attitudes toward the two target groups. The IAT was chosen as an implicit measure because it tends to produce large effect sizes both in general and specifically in the context of attitude induction via evaluative conditioning (Hofmann et al., 2010; Nosek et al., 2007). The IAT consisted of five blocks with 20, 20, 40, 20, and 40 trials each. In block 1, participants sorted positively and negatively valenced words\(^5\) and in block 2, they sorted Laapian and Niffian names used in the learning phase. In block 3, the first critical block, both sorting tasks were combined. For participants in the control condition, the initial assignment of valence to target group was randomized. For participants who underwent prior learning, the first critical block of the IAT was always in line

---

\(^4\) In future studies, time can be allowed to vary with number of trials held constant, but there is no reason to assume that one trade-off is superior to the other. Because time was the crucial dimension to compare ES to REP, it was used by extension for this third condition as well. Moreover, as described in more detail in the context of Study 6A, increasing the number of evaluative conditioning trials from 20 to 37 does not create a stronger implicit attitude.

\(^5\) Positive items included *love, peace, joy, happy, sweet, glory,* and *success.* Negative items included *hate, war, devil, bomb, bitter, agony,* and *failure.*
with the prior attitude induction. For instance, if a participant learned that Laapians are good and Niffians are bad, they sorted Laapians with positive words and Niffians with negative words on the first critical block. This was necessary in order to make sure that the IAT remained purely a test of previously acquired attitudes rather than forcing participants to override prior learning by presenting counterattitudinal statements. In block 4, participants practiced the new assignment of target groups to response keys and finally, in block 5, they completed a combined task with the opposite assignment of target group to valence.

**Explicit attitudes.** Following the IAT, participants were asked to respond to two feeling thermometer items, each of which measured explicit attitudes toward one of the target groups. On the first item participants were instructed to indicate how warmly or coldly they felt toward Laapians. On the second item they were instructed to indicate how warmly or coldly they felt toward Niffians. Responses were provided on 10-point Likert scales, anchored by “extremely warmly” on the left-hand side and “extremely coldly” on the right-hand side.

In line with the research reviewed earlier, the primary interest of the present work is in the effects of ES and REP on implicit attitudes. We collected matching data on explicit attitudes in the hope that the results may provide some insight into dissociations between more and less automatic evaluations. Given their status in this research, explicit measures were always administered following the IAT, and are best viewed as manipulation checks. As such, data from these measures will be reported briefly for each experiment, with the focus on whether they are evaluatively in the same direction as the IAT score. Given the attitudinally neutral targets used in the present study, no dissociation between implicit and explicit attitudes is expected.
Results

**Manipulation check.** As in many such experiments, explicit and implicit attitudes were significantly correlated with each other, Pearson’s $r = .36 [.30; .43]$, $t(664) = 10.04$, $p < .001$, underscoring the soundness of the manipulation and experimental procedure.

**Implicit attitudes.** Participants’ implicit attitudes toward the two target groups were calculated using the scoring algorithm recommended by Greenwald, Nosek and Banaji (2003), with positive D scores indicating an implicit preference in line with the prior learning episode. Descriptive statistics (means and standard deviations) by condition are reported in Table 1 but will not be discussed in detail given that the coefficients in the regression analysis below express condition means and mean differences across conditions.

Table 1. IAT D score means and standard deviations (in parentheses) by congruency and learning condition (Control; REP = repeated evaluative pairings; ES = evaluative statements; Combined) for Studies 1–5. Positive D scores indicate implicit attitudes in line with the prior learning.

<table>
<thead>
<tr>
<th>Study</th>
<th>Congruent</th>
<th>Incongruent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>REP</td>
<td>ES</td>
</tr>
<tr>
<td>Study 1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Study 2</td>
<td>0.48</td>
<td>0.54</td>
</tr>
<tr>
<td>Study 3</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>Study 4</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>Study 5</td>
<td>0.75</td>
<td>0.78</td>
</tr>
</tbody>
</table>

In order to assess the relative strength of the attitude induction modalities, D scores were submitted to a linear regression, with learning modality (control vs. REP vs. ES vs. combined) as

---

6 Raw data and analysis scripts for all studies are available for download from OSF ([https://osf.io/jyk8c/](https://osf.io/jyk8c/)).

7 The mean levels of explicit attitudes for each experimental condition are available on OSF ([https://osf.io/jyk8c/](https://osf.io/jyk8c/)).
the sole predictor.\(^8\) Because we used dummy coding with the control condition as the reference category, the intercept corresponds to the mean of the control condition and the slope parameters correspond to learning effects compared to control. The intercept did not significantly differ from zero, \(b = 0.00 \ [\text{-0.07; 0.08}]\), \(t(671) = 0.11, p = .912\), indicating a neutral baseline in the control condition. More importantly, however, REP created a statistically significant learning effect compared to baseline, \(b = 0.43 \ [0.32; 0.54]\), \(t(671) = 7.73, p < .001\). ES produced an even more sizable learning effect, \(b = 0.50 \ [0.39; 0.61]\), \(t(671) = 9.09, p < .001\), and the strongest learning effect was observed in the combined condition, \(b = 0.55 \ [0.44; 0.65]\), \(t(671) = 10.46, p < .001\).

In order to directly investigate the relative effectiveness of the REP, ES, and combined conditions compared to each other, we drew 10,000 bootstrap samples for each mean difference and calculated 95-percent, 99-percent, and 99.9-percent bootstrap confidence intervals around the estimate of the difference. If the 95-percent confidence interval includes zero, we conclude that the two condition means are not significantly different. If the 95-percent, 99-percent, or 99.9-percent confidence interval does not include zero, we conclude that the condition means are significantly different from each other with \(p < .05\), \(p < .01\), or \(p < .001\), respectively. For the REP vs. ES contrast, we found a difference of \(b_{\text{diff}} = 0.07 \ [-0.04; 0.18]\) in favor of ES; however, it failed to reach statistical significance, \(p > .05\). Whereas the combined condition added significant value to the REP condition, \(b_{\text{diff}} = 0.12 \ [0.01; 0.23]\), \(p < .05\), the difference between the ES and combined conditions was not significant, \(b_{\text{diff}} = 0.05 \ [-0.05; 0.15]\), \(p > .05\).

\(^8\) In lieu of the analyses reported here, i.e., linear regressions combined with bootstrapping for additional contrasts, it would have been possible to conduct two-way ANOVAs, with REP (present vs. absent) and ES (present vs. absent) as the two predictors. We decided against this possibility given that the combined condition was not a perfect combination of the REP and ES conditions, among other reasons because time on intervention was fixed across the three learning conditions and thus the combined condition did not include the same number of stimulus pairings as the REP condition. However, the raw data files published on OSF enable other investigators to pursue this or any other alternative data analysis approach.
In terms of absolute strength, the implicit attitudes found in each of the three learning modality conditions were on par with some of the strongest implicit attitudes found in American adult populations toward real groups such as the implicit preference for young over elderly people or for White Americans over Black Americans (Nosek, Greenwald, & Banaji, 2005). In terms of their relative strength, REP and ES produced comparable learning effects; however, the combined condition added value only to the REP condition and not to the ES condition.

**Discussion**

Three learning modalities and a control condition were created with the aim of testing the relative effectiveness of two different methods of implicit attitude change involving two fictitious social groups. All three learning modalities proved to be effective even though one of them, ES, relied on a most minimal instruction. We found strong implicit attitudes on par with some of the most robust social group attitudes observed in previous work. This study reveals the ease with which implicit attitudes can be created, both in time taken to produce an effect (less than 3 minutes) and the minimal instruction that is needed to produce them, at least in the ES condition. Importantly, Study 1 replicates prior research showing that exposing participants to ES simply stating that the attitude object will be paired with positive or negative stimuli can result in strong implicit attitudes toward the target (De Houwer, 2006; Gast & De Houwer, 2013; Van Dessel, De Houwer, Gast, & Smith, 2015b).

That ES were not inferior to REP in shifting attitudes is a new result and surprising for a number of reasons. First, the REP condition was arguably more attentionally intense than the ES condition. Whereas in the former condition participants were presented with a new stimulus pair every 3.5 seconds, in the latter condition participants were merely asked to keep reading some instructions that remained stationary on the screen. Second, informing participants that they will
view particular pairings and then never presenting the pairings should not have given rise to any strong attitudinal effects, considering that this information is not in any way diagnostic of the targets (Cone & Ferguson, 2015). Moreover, based on these results it seems that learning via REP and learning via ES are, at least to some extent, redundant. REP did not offer any incremental value beyond ES, suggesting that the learning created by the combined intervention is subsumed under the learning created by the ES intervention alone. For these reasons, it is particularly striking that ES allowed the formation of any attitude at all, let alone as robust as the one observed. However, this first result may reflect possible artifacts of the experimental situation.

Moreover, it is at least theoretically possible that in spite of the exceedingly low p value, the result is a Type I error.

**Study 2**

In Study 1 we found that REP and ES both created learning effects. The learning effect produced by ES was on par with the learning effect produced by REP, and the combined condition only added value to REP and not to ES. The stimuli used in Study 1 offered the advantage of a neutral baseline, which is known to be difficult to achieve in the context of implicit social attitudes. By virtue of this attitudinal neutrality, we had complete control over participants’ learning, without any contaminating effects of prior exposure to the stimuli. However, in experimental psychology there is often a conflict between internal and external validity (Banaji & Crowder, 1989; 1991). In Study 2 we sought to tilt the balance toward the latter and used more ecologically valid social target stimuli—drawings of individuals with long and square faces. Based on pre-testing we expected that participants would have a fairly strong implicit preference in favor of Longfaces at baseline. Seemingly neutral stimuli are often found to be non-neutral using implicit measures of attitudes (Bargh et al., 1992; Fazio et al., 1986; Glaser & Banaji, 1999; Gregg et al.,
Thus, the incongruent condition (i.e., the condition in which participants’ attitudes are allowed to move away from, rather than closer to, the prevailing attitude) presents a cleaner test case for our hypotheses, considering that in the congruent learning condition ceiling effects might occur. In fact, to avoid this problem, many classic attitude induction studies (Cacioppo, Petty, & Morris, 1983; Petty & Cacioppo, 1979; Petty, Wells, & Brock, 1976) and studies on attitude induction via EC and REP (De Houwer, 2006; Lai et al., 2014; Van Dessel, De Houwer, Gast, & Smith, 2015b) contain only incongruent learning conditions. Accordingly, our analyses focus on the incongruent condition. Findings from the congruent condition are reported in a footnote and discussed in more detail only in the context of results collapsing across experiments. Moreover, if REP is to be a purely non-linguistic learning modality, Study 1 was problematic in that group members were presented as words. In Study 2, all stimuli—including the good–bad UC attributes and the long-faced and square-faced individuals serving as CS—were presented non-linguistically, in pictures.

Method

Participants. 1162 volunteers were recruited from the Project Implicit educational website to participate in the study. 119 participants were excluded from all analyses for failing to complete the Implicit Association Test and 9 participants were excluded because they had response latencies below 300 milliseconds on more than 10 percent of all IAT trials. This resulted in a final sample size of $N = 1034$ participants ($N = 584$ female, mean age = 33.48, $SD = 14.47$ years). This sample size, combined with the effect size obtained, yielded post-hoc power of $.99$ in the incongruent condition and $.86$ in the congruent condition. 314 participants were assigned to the control condition, 226 to the REP condition, 250 to the ES condition, and 244 to the combined condition. Moreover, each participant was also assigned to one of two congruency condi-
tions, i.e., they learned either that Longfaces are good and Squarefaces are bad (congruent condition, $N = 506$) or that Squarefaces are good and Longfaces are bad (incongruent condition, $N = 528$). We refer to the Longfaces–good condition as the congruent condition because in the control condition 210 participants showed an implicit preference in favor of Longfaces over Squarefaces, and only 104 participants exhibited the opposite preference, $p < .001$, by binomial test.

**Materials and procedure.** The learning procedure (including the US) and the implicit and explicit attitude measures paralleled the ones used in Study 1. However, unlike in Study 1, drawings of human faces served as CS. Faces were chosen as target stimuli because of their singular ability to represent a social being. Previous research has shown that human faces carry a wealth of social information that gives rise to instantaneous trait inferences (Ballew & Todorov, 2007; Olivola & Todorov, 2010; Oosterhof & Todorov, 2008; Todorov, Mandisodza, Goren, & Hall, 2005; Todorov, Said, Engell, & Oosterhof, 2008; Willis & Todorov, 2006). Neonates preferentially attend to faces (C. C. Goren, Sarty, & Wu, 1975), and the psychological dispositions inferred from faces by young children are remarkably similar to those inferred by adults (Cogsdill & Banaji, 2015; Cogsdill, Todorov, Spelke, & Banaji, 2014). However, face-to-trait inferences are dynamic and malleable (Hehman, Flake, & Freeman, 2015), suggesting that faces can be the targets of new learning.

In the present study we used two sets of faces that differed along a salient perceptual dimension (length-to-width ratio) for the purposes of attitude induction (Hill, Lewicki, Czyzewska, & Schuller, 1990; Lewicki, 1986). Each face was generated by morphing photographs of two young White men facing the camera, cropped at the neck and displaying a neutral facial expression. In the following step, the faces were subjected to a sketch effect in order to create the appearance of line drawings. Crucially, in the last step of preparation, we manipulated the length-
to-width ratio of the faces, yielding a set of long faces (length-to-width ratio = 2:1) and a set of square faces (length-to-width ratio = 4:3).

Results

**Manipulation check.** Explicit and implicit attitudes were found to be moderately correlated with each other, $r = .25 \ [0.20; 0.31]$, $t(1015) = 8.46$, $p < .001$, underscoring the soundness of our design and manipulation.

**Implicit attitudes.** Descriptive statistics by condition are reported in Table 1. In order to assess the relative strengths of the attitude induction modalities, participants’ D scores from the incongruent condition\(^9\) were submitted to a linear regression. In the incongruent condition, the intercept significantly differed from zero, $b = 0.09 \ [0.01; 0.16]$, $t(524) = 2.20$, $p = .028$, indicating a weak baseline preference for Squarefaces. REP did not produce a significant learning effect compared to baseline, $b = 0.08 \ [-0.04; 0.20]$, $t(524) = 1.35$, $p = .176$. In contrast, the ES condition, $b = 0.32 \ [0.20; 0.44]$, $t(524) = 5.21$, $p < .001$, and the combined condition, $b = 0.37 \ [0.25; 0.49]$, $t(524) = 6.18$, $p < .001$, produced robust learning effects. By bootstrap, the ES condition created significantly stronger learning than the REP condition, $b_{\text{diff}} = 0.23 \ [0.10; 0.37]$, $p < .001$. The combined condition outperformed the REP condition, $b_{\text{diff}} = 0.28 \ [0.16; 0.41]$, $p < .001$, but not the ES condition, $b_{\text{diff}} = 0.05 \ [-0.07; 0.16]$, $p > .05$.

Discussion

\(^9\) The learning effects created in the congruent condition were overall negligible. The intercept significantly differed from zero, $b = 0.48 \ [0.41; 0.55]$, $t(502) = 13.95$, $p < .001$, indicating a strong baseline preference for Longfaces. The fact that baseline preferences differed across congruency conditions is due to the well-documented order effect on the IAT (Greenwald & Nosek, 2001; Nosek et al., 2005; 2007). REP did not produce a significant learning effect compared to baseline, $b = 0.06 \ [-0.05; 0.16]$, $t(502) = 1.03$, $p = .302$. The learning effect produced by ES also failed to reach significance, $b = 0.09 \ [-0.02; 0.19]$, $t(502) = 1.67$, $p = .096$. The combined condition produced a small but statistically significant learning effect, $b = 0.19 \ [0.09; 0.30]$, $t(502) = 3.56$, $p < .001$. Using a bootstrapping method for additional contrasts, we found no significant difference between REP and ES, $b_{\text{diff}} = 0.03 \ [-0.07; 0.14]$, $p > .05$. The combined condition added significant value to both the REP condition, $b_{\text{diff}} = 0.13 \ [0.02; 0.25]$, $p < .05$, and the ES condition, $b_{\text{diff}} = 0.10 \ [0.00; 0.20]$, $p < .05$. 

42
With a new set of social group stimuli, Study 2 replicated the results of Study 1 to the extent that we observed a fairly strong implicit attitude toward a novel social attitude object in each learning modality condition, confirming that the attitude induction was, again, successful. Moreover, the pattern of results obtained in the incongruent condition, which allowed ample room for movement, was similar to the pattern of results observed in Study 1 in that the REP condition produced the weakest learning effect, and the learning effects produced by ES and the combined intervention were on par with each other. Unlike in Study 1, however, where the learning effects created by ES were descriptively but not inferentially stronger than those created by REP, here REP was clearly outperformed by ES. This is a novel result that was not predicted by any of the theoretical perspectives reviewed. Moreover, informing participants that they would be exposed to certain stimulus pairings created an implicit attitude that was nearly as strong as the attitude created by first informing participants of the stimulus pairings and then actually exposing them to the stimuli. Taken together, these results speak to the generalizability of the findings obtained in Study 1 to more ecologically valid targets where the baseline implicit preference is not neutral. Even minimal language-based manipulations involving nondiagnostic information seem to be robustly effective in shifting implicit attitudes and at least as powerful, if not more so, than repeated presentation of evaluative pairings.

**Study 3**

Studies 1 and 2 have demonstrated that ES are more effective than REP in shifting implicit attitudes. Moreover, REP did not seem to have any incremental effect over and above ES, suggesting that language provides solid cognitive scaffolding for social attitudes to be built. However, it remains to be directly investigated whether the social–nonsocial divide modulates the relative effectiveness of REP vs. ES in shifting implicit attitudes. Different bodies of litera-
ture from social psychology suggest various possible hypotheses regarding this issue. According to one long-standing view, theories of social cognition can be derived from general theories about the physical world (Banaji & Bhaskar, 2000; see also Contreras, Banaji, & Mitchell, 2012). This perspective would predict that there should not be much difference between shifting implicit social and implicit nonsocial attitudes. On the other hand, social groups tend to be more complex than categories of nonsocial objects (Cantor & Mischel, 1979), and knowledge about social targets often evokes more emotion than knowledge about nonsocial targets (Norris, Chen, Zhu, Small, & Cacioppo, 2004). Moreover, work from social neuroscience suggests that different neural processes might underlie social and nonsocial learning (Contreras et al., 2012; J. P. Mitchell, Heatherton, & Macrae, 2002). The latter set of findings would predict that the acquisition of social and nonsocial attitudes might differ from each other. Specifically, according to Dunbar (1996), language may have evolved for the purpose of gossip, i.e., strengthening social bonds by talking about shared acquaintances. If this is so, language-based learning mechanisms such as ES might be especially effective in creating attitudes toward social, rather than nonsocial, targets. Thus, in Study 3, we sought to test whether the separate and joint learning effects of REP and ES would remain the same for nonsocial targets (squares vs. rectangles) while keeping superficial visual features of the stimuli constant.

**Method**

**Participants.** 918 volunteers were recruited from the Project Implicit educational website to participate in the study. 50 participants were excluded from all analyses for failing to complete the Implicit Association Test and 5 participants were excluded because they had response latencies below 300 milliseconds on more than 10 percent of all IAT trials. This resulted in a final sample size of $N = 863$ participants ($N = 606$ female, mean age = 37.87, $SD = 14.17$ years). This
sample size, combined with the effect size obtained, yielded post-hoc power of 1 within machine precision for the incongruent condition and .97 for the congruent condition. 236 participants were assigned to the control condition, 199 to the REP condition, 209 to the ES condition, and 219 to the combined condition. Moreover, each participant was also assigned to one of two congruency conditions, i.e., they learned either that squares are good and rectangles are bad (congruent condition, \( N = 428 \)) or that rectangles are good and squares are bad (incongruent condition, \( N = 435 \)). We refer to the squares–good condition as the congruent condition because in the control condition 164 participants showed an implicit preference in favor of squares over rectangles, and only 72 participants exhibited the opposite preference, \( p < .001 \), by binomial test.

**Materials and procedure.** The learning procedure (including the US) and the implicit and explicit attitude measures paralleled the ones used in Studies 1 and 2, with squares and rectangles serving as CS. The squares and rectangles were created by turning the pixels that formed the face stimuli in Study 2 into random noise and cutting them into form using square-shaped and rectangle-shaped cropping tools. This way, the CS used in Study 3 were similar to the CS used in Study 2 in terms of superficial visual features such as hue, saturation, and brightness. Pretesting had revealed a fairly high level of mean implicit preference in favor of squares over rectangles. In an attempt to reduce this implicit preference, we used sharp edges for squares and round edges for rectangles. Moreover, in order to keep the verbal labels maximally similar to the labels used in Study 2, squares were referred to as “square shapes” and rectangles were referred to as “long shapes” throughout the instructions and the words “square” and “long” were used as labels on the IAT.

**Results**
**Manipulation check.** Explicit and implicit attitudes were found to be moderately correlated with each other, $r = .32 \, [.25; .37]$, $t(848) = 9.68$, $p < .001$, underscoring the soundness of our design and manipulation.

**Implicit attitudes.** Descriptive statistics by condition are reported in Table 1. In order to assess learning effects in the incongruent condition\(^{10}\), a linear regression was fit to the data, with learning modality as the sole predictor and the control condition as the reference category. The intercept, i.e., the mean of the control condition, did not differ from zero, $b = 0.03 \, [-0.06; 0.11]$, $t(431) = 0.66$, $p = .508$, indicating no baseline preference. Each learning modality produced a strong learning effect. In the REP condition, $b = 0.51 \, [0.38; 0.63]$, $t(431) = 8.03$, $p < .001$; in the ES condition, $b = 0.52 \, [0.39; 0.65]$, $t(431) = 8.05$, $p < .001$; and in the combined condition, $b = 0.52 \, [0.39; 0.65]$, $t(431) = 7.99$, $p < .001$. We found no differences across learning conditions: For the REP vs. ES contrast, $b_{\text{diff}} = 0.01 \, [-0.11; 0.14]$, $p > .05$; for REP vs. combined, $b_{\text{diff}} = 0.01 \, [-0.12; 0.14]$, $p > .05$; and for ES vs. combined, $b_{\text{diff}} = 0.00 \, [-0.14; 0.13]$, $p > .05$, all by bootstrap.

**Discussion**

In Study 3 we tested the effectiveness of REP vs. ES in the context of nonsocial attitudes. The pattern of results obtained in Study 3 had both similarities and differences compared to the previous nonsocial studies. In line with Studies 1 and 2, we observed strong learning effects. Moreover, ES produced robustly strong implicit attitudes and REP did not offer any incremental value over and above ES. Unlike in Study 2, however, the learning effects produced by REP in

\(^{10}\)The learning effects created in the congruent condition were overall negligible. The intercept significantly differed from zero, $b = 0.49 \, [0.42; 0.57]$, $t(424) = 12.96$, $p < .001$, indicating a strong baseline preference for squares. The difference in baseline preference across congruency conditions is due to an IAT order effect (Greenwald & Nosek, 2001; Nosek et al., 2005; 2007). As in Study 2, REP did not produce a significant learning effect compared to baseline, $b = 0.00 \, [-0.11; 0.11]$, $t(424) = 0.05$, $p = .961$. The learning effects produced by ES, $b = 0.15 \, [0.05; 0.26]$, $t(424) = 2.85$, $p = .005$, and the combined condition, $b = 0.19 \, [0.09; 0.30]$, $t(424) = 3.64$, $p < .001$, were small but significant. The ES condition, $b_{\text{diff}} = 0.15 \, [0.03; 0.27]$, $p < .05$, and the combined condition, $b_{\text{diff}} = 0.19 \, [0.08; 0.30]$, $p < .01$, clearly outperformed REP. However, we found no significant difference between ES and combined, $b_{\text{diff}} = 0.03 \, [-0.06; 0.13]$, $p > .05$, all three contrasts by bootstrap.
isolation were on par with the learning effects produced by ES. Overall, learning effects were somewhat larger than in Studies 1 and 2, suggesting that attitudes toward nonsocial stimuli might be more malleable than attitudes toward social stimuli, especially in the face of stimulus pairings experienced in the environment. Crucially, however, Study 3 further underscores the generalizability of our findings about the power of ES-based learning in shifting implicit attitudes.

**Study 4**

In Studies 1 and 2 we demonstrated that ES are not any less effective than REP in shifting implicit attitudes toward novel social groups created within the context of the experiment. Using such *ad hoc* social groups offers the advantage of complete experimental control over participants’ learning; however, such social groups may elicit “nonattitudes” (Pierce & Rose, 1974), thus compromising the external validity of the findings. Moreover, prior research suggests that REP and ES might differ in terms of their effectiveness in shifting real-world attitudes. Several studies have found that REP can shift implicit attitudes toward real-world social groups like the elderly or African Americans (Karpinski & Hilton, 2001; Lai et al., 2014; Olson & Fazio, 2006). However, this might not be the case for ES. In a recent study (Van Dessel, De Houwer, Gast, & Smith, 2015b), approach/avoidance instructions shifted implicit attitudes toward fictitious social groups (Niffites and Luupites) but not toward White and Black Americans. This finding suggests that our results on the power of ES might not generalize to existing social groups.

However, a potential limitation of the study conducted by Van Dessel et al. (2015b) is that they tested the difference between existing and novel targets using only White vs. Black Americans as target groups. Results obtained with these two groups may or may not generalize to other existing social groups. Implicit attitudes toward Black Americans are recalcitrant (Lai et al., 2014), and the attempt to change them might even result in ironic reactance effects (Eagly &
Chaiken, 1993). Therefore, rather than using White–Black as the target contrast, we probed whether the results obtained in previous studies would generalize to shifting attitudes toward two other social contrasts subject to high levels of implicit bias in contemporary American society. In order to provide a robust test of the relative and combined effects of REP and ES, we selected two pairs of target categories, one of them a biologically and evolutionary relevant social category and the other one an arbitrary-set category (Pratto, Sidanius, & Levin, 2006): young vs. elderly and American vs. foreign. Negative explicit (Cuddy & Fiske, 2004) and implicit (B. R. Levy & Banaji, 2002) attitudes toward the elderly are highly prevalent. Similarly, Americans express high levels of explicit (Fiske, Cuddy, Glick, & Xu, 2002) and implicit (Carter, Ferguson, & Has-sin, 2011) preference for their ingroup over foreigners. More specifically, high levels of implicit bias measured toward the elderly (Nosek et al., 2007) and foreigners (Cunningham, Nezlek, & Banaji, 2004) have been amply documented in the context of IAT research, making these two contrasts an ideal test case for a comparative investigation of REP and ES in real-world settings.

**Method**

**Participants.** 881 volunteers were recruited from the Project Implicit educational website to participate in the study. 29 participants were excluded from all analyses for failing to complete the Implicit Association Test and 4 participants were excluded because they had response latencies below 300 milliseconds on more than 10 percent of all IAT trials. This resulted in a final sample size of \( N = 848 \) participants (\( N = 600 \) female, mean age = 33.85, \( SD = 14.39 \) years). This sample size, combined with the effect size obtained, yielded post-hoc power of 1 within machine precision for the incongruent condition and, given the small effect size, .51 for the congruent condition. 217 participants were assigned to the control condition, 221 to the REP condition, 199 to the ES condition, and 211 to the combined condition. Moreover, each participant was also as-
signed to one of two congruency conditions, i.e., they learned either that young people are good and elderly people are bad (congruent condition, \( N = 419 \)) or that elderly people are good and young people are bad (incongruent condition, \( N = 429 \)). We refer to the young–good condition as the congruent condition on the basis of previous work that has revealed robust implicit bias in favor of young over elderly people in American society (B. R. Levy & Banaji, 2002; Nosek et al., 2007). In line with this previous work, 209 participants in the control condition showed an implicit preference in favor of young over the elderly, and only 8 participants exhibited the opposite preference, \( p < .001 \), by binomial test.

**Materials and procedure.** The learning procedure and the implicit and explicit attitude measures paralleled the ones used in previous studies. In this study we used faces of young and elderly people as CS and a novel set of photographs as US. Conditioned stimuli were black-and-white photographs of faces of young and elderly people (White, both female and male), with external facial features removed. These stimuli were obtained from the Project Implicit website (http://implicit.harvard.edu). The US were taken from the Open Affective Standardized Image Set (OASIS; Kurdi, Lozano, & Banaji, 2017), an open-access collection of color photographs with valence and arousal norms. The US were selected to be highly positive or negative and moderately arousing. The positive set, consisting of OASIS images I256, I335, I345, I456, and I462, had an overall mean valence of 6.29 and a mean arousal of 4.22 (both measured on a 7-point scale). The negative set, consisting of OASIS images I119, I227, I236, I276, and I382, had an overall mean valence of 1.74 and a mean arousal of 4.06. This novel set of images was used for two reasons. First, because REP did not outperform ES in any of the previous studies and unexpectedly so, we sought to verify that the superiority of ES was not due to the particular US included in the studies. Second, because the US used in Studies 1 through 3 were line drawings, we
expected that they might contaminate the results of this study by virtue of their stereotypic association with the category “young” rather than the category “elderly.” The images that we selected for Study 4 were not stereotypically associated with either category.

Results

Manipulation check. Unlike in previous studies, explicit and implicit attitudes were found to be only weakly correlated with each other, \( r = .07 \) [.00; .14], \( t(834) = 1.99, p = .046 \). This is often the case when explicit measures are subject to social desirability pressures (Hofmann et al., 2005; Nosek, 2005; 2007).

Implicit attitudes. Descriptive statistics by condition are reported in Table 1. In order to assess learning effects in the incongruent condition\(^{11}\), a linear regression was fit to the data, with learning modality as the sole predictor and the control condition as the reference category. The intercept was significantly different from zero, \( b = -0.61 \) [-0.69; -0.53], \( t(425) = -14.88, p < .001 \), which indicates a strong baseline preference for young over elderly, replicating previous research (B. R. Levy & Banaji, 2002). All three conditions, including REP, \( b = 0.26 \) [0.15; 0.38], \( t(425) = 4.52, p < .001 \), ES, \( b = 0.42 \) [0.30; 0.54], \( t(425) = 6.96, p < .001 \), and combined, \( b = 0.48 \) [0.37; 0.60], \( t(425) = 8.10, p < .001 \), produced considerable learning effects compared to control. Compared to the REP condition, the learning effects created by the ES condition, \( b_{\text{diff}} = 0.15 \) [0.03; 0.27], \( p < .05 \), and by the combined condition, \( b_{\text{diff}} = 0.22 \) [0.09; 0.34], \( p < .01 \), were more robust.

---

\(^{11}\) Learning effects in the congruent condition were negligible. The intercept significantly differed from zero, \( b = 0.61 \) [0.54; 0.68], \( t(415) = 17.11, p < .001 \), revealing a strong baseline preference in favor of young over elderly. As in Studies 2 and 3, REP did not produce a significant learning effect, \( b = 0.05 \) [-0.05; 0.15], \( t(415) = 0.05, p = .960 \). The learning effects produced by ES, \( b = 0.10 \) [0.00; 0.20], \( t(415) = 2.04, p = .041 \), and the combined condition, \( b = 0.11 \) [0.01; 0.20], \( t(415) = 2.14, p = .033 \), were fairly small but significant and of comparable magnitude. None of the three remaining contrasts were significant, with REP vs. ES at \( b_{\text{diff}} = 0.05 \) [-0.04; 0.15], \( p > .05 \); REP vs. combined at \( b_{\text{diff}} = 0.05 \) [-0.04; 0.15], \( p > .05 \), and ES vs. combined, \( b_{\text{diff}} = 0.00 \) [-0.09; 0.09], \( p > .05 \), all by bootstrap.
and comparable with each other in terms of magnitude, $b_{\text{dif}} = 0.06 [-0.06; 0.19]$, $p > .05$, all three contrasts by bootstrap.

**Discussion**

In Study 4, we tested the effectiveness of REP and ES in shifting implicit attitudes in the context of a real-world contrast, young vs. elderly people. Consistent with previous work (B. R. Levy & Banaji, 2002; Nosek et al., 2007), we obtained strong implicit preference in favor of young over elderly at baseline. The learning effects observed in the incongruent condition, designed to move participants away from the prevailing implicit preference in favor of young over elderly, were robust. In line with Studies 1 and 3 (and unlike in Study 2), REP produced a significant learning effect in isolation, replicating previous work showing that implicit attitudes toward the elderly can be changed via REP (Karpinski & Hilton, 2001). However, like in Study 2, ES was more effective than REP and the learning effect observed in the combined condition was not superior to the learning effect observed in the ES condition. In other words, Study 4 seems to suggest that learning involving existing and ad hoc social groups might be less different from each other than previously assumed. The failure to find learning effects for mere instructions in previous work (Van Dessel, De Houwer, Gast, & Smith, 2015b) might have been the result of the particular contrast used (i.e., White vs. Black Americans). Thus, the pattern of results obtained in Studies 1 and 2 does not seem to be confined to novel social groups created for the purpose of the experiment; rather, they seem to generalize to existing social groups subject to high levels of preexisting implicit bias.

**Study 5**

As described in the introduction to Study 4, previous research has failed to find effects for ES in the context of real-world social groups (Van Dessel, De Houwer, Gast, & Smith,
2015b). Hence we sought to probe the robustness of the findings of Study 4 by replicating the same experiment using another real-world contrast subject to strong preexisting explicit (Fiske et al., 2002) and implicit (Carter et al., 2011) attitudes—American vs. foreign.

Method

Participants. 1035 volunteers were recruited from the Project Implicit educational website to participate in the study. 59 participants were excluded from all analyses for failing to complete the Implicit Association Test and 18 participants were excluded because they had response latencies below 300 milliseconds on more than 10 percent of all IAT trials. This resulted in a final sample size of $N = 958$ participants ($N = 672$ female, mean age = 30.79, $SD = 14.09$ years). This sample size, combined with the effect size obtained, yielded post-hoc power of > .99 for the incongruent condition and .90 for the congruent condition. Participation was restricted to volunteers with both American citizenship and U.S. residence. 264 participants were assigned to the control condition, 206 to the REP condition, 213 to the ES condition, and 275 to the combined condition. Moreover, each participant was also assigned to one of two congruency conditions, i.e., they learned either that American is good and foreign is bad (congruent condition, $N = 478$) or that foreign is good and American is bad (incongruent condition, $N = 480$). We refer to the American–good condition as the congruent condition because in the control condition 254 participants showed an implicit preference in favor of American over the foreign, and only 10 participants exhibited the opposite preference, $p < .001$, by binomial test.

Materials and procedure. The learning procedure and the implicit and explicit attitude measures paralleled the ones used in previous studies. In this study we used drawings and photographs as CS and the line drawings used in Studies 1 through 3 as US. CS were adapted from Devos and Banaji (2005) and Devos, Gavin, and Quintana (2010). For the American category,
the stimuli included a 1-dollar bill, the Great Seal of the United States, the U.S. flag, the Golden Gate Bridge, and a map of the continental United States. For the foreign category, the stimuli included a 100-hryven bill, the coat of arms of Flanders, the flag of Djibouti, the Great Sphinx of Giza, and a map of Luxembourg rotated by 90 degrees to the left.

Results

Manipulation check. Explicit and implicit attitudes were found to be moderately correlated with each other, $r = .40 [0.34, 0.46]$, $t(944) = 13.54, p < .001$, underscoring the soundness of our design and manipulation.

Implicit attitudes. Descriptive statistics by condition are reported in Table 1. In order to assess learning effects in the incongruent condition$^{12}$, a linear regression was fit to the data, with learning modality as the sole predictor and the control condition as the reference category. The intercept was significantly different from zero, $b = -0.63 [-0.71; -0.56]$, $t(476) = -16.07, p < .001$, which indicates a strong baseline preference for American over foreign, replicating previous work (Nosek, 2005; Nosek et al., 2007). Unlike in Studies 1, 3 and 4, and like in Study 2, the REP condition did not produce a significant learning effect, $b = 0.06 [-0.05; 0.18]$, $t(476) = 1.04$, $p = .299$. However, as in all previous studies, the ES condition, $b = 0.34 [0.22; 0.46]$, $t(476) = 5.65, p < .001$ and the combined condition, $b = 0.32 [0.22; 0.43]$, $t(476) = 5.89, p < .001$, created strong learning effects compared to baseline. Accordingly, we found statistically significant differences between the REP and ES conditions, $b_{\text{diff}} = 0.28 [0.16; 0.40], p < .001$, as well as the

---

$^{12}$ Learning effects in the congruent condition were negligible. The intercept significantly differed from zero, $b = 0.75 [0.68; 0.81]$, $t(474) = 23.44, p < .001$, indicating a strong baseline preference for American over foreign. As in Studies 2, 3 and 4, REP did not produce a significant learning effect, $b = 0.03 [-0.06; 0.12]$, $t(474) = 0.66, p = .508$. The learning effects produced by ES, $b = 0.14 [0.05; 0.23]$, $t(474) = 3.03, p = .003$, and the combined condition, $b = 0.13 [0.05; 0.22]$, $t(474) = 3.00, p = .003$, were fairly small but statistically significant and of similar magnitude. Accordingly, both ES, $b_{\text{diff}} = 0.11 [0.01; 0.21], p < .05$, and combined, $b_{\text{diff}} = 0.10 [0.01; 0.20], p < .05$, created a stronger learning effect than REP. We found no difference between ES and combined, $b_{\text{diff}} = -0.01 [-0.09; 0.07], p > .05$, all by bootstrap.
REP and combined conditions, $b_{diff} = 0.26 \ [0.14; 0.38]$, $p < .001$. However, no significant difference was found between ES and combined, $b_{diff} = -0.02 \ [-0.14; 0.11]$, $p > .05$, all contrasts by bootstrap.

**Discussion**

In Study 5 we tested whether the results obtained in Study 4 would generalize to another pre-existing social contrast. We replicated the major results from Study 4, with one notable exception. REP did not produce significant learning effects, suggesting that stimulus pairings experienced in the environment may not always be sufficient to move implicit attitudes away from a strong baseline. This is a useful result to have obtained because it represents a unique contrast to the strength of ES. Indeed, ES and the combined intervention produced learning effects significantly above baseline and on par with each other.

**Study 6A**

With the ES condition producing learning effects at least on par with the REP condition and the combined condition not adding any value to the ES condition, Studies 1–5 yielded some surprising yet consistent findings. Therefore, we conducted three follow-up studies in which we sought to test to what extent these findings are robust to procedural variations involving time on intervention (Study 6A), levels of construal (Study 6B), and expectancy effects (Study 6C), thus eliminating potential confounds. In Study 6A, we probed whether fixing time on intervention across the learning conditions was instrumental to the superiority of the ES manipulation, or the same effects would be obtained if learning unfolded in a self-paced manner. Moreover, we investigated whether the combined condition would produce stronger learning than the ES condition if it included the same number of conditioning trials as the REP condition.

**Method**
In Study 6A, conducted with 873 participants on Project Implicit and involving the same stimuli (i.e., line drawings of individuals with long and square faces) and the same four learning conditions (i.e., control, REP, ES, and combined) as Study 2, we did not equate the time spent on intervention across learning conditions. Rather, learning was self-paced in the ES condition, and the REP and combined conditions featured the same number of conditioning trials (37).

Results and discussion

In the incongruent (i.e., Squarefaces–good) condition, the intercept, i.e., the mean of the control condition, did not differ from zero, \( b = -0.10 [-0.20; 0.00], t(429) = -1.90, p = .058 \), indicating no baseline preference. Each learning modality produced a learning effect compared to control. In the REP condition, \( b = 0.22 [0.08; 0.36], t(429) = 3.13, p = .002 \); in the ES condition, \( b = 0.46 [0.33; 0.61], t(429) = 6.52, p < .001 \); and in the combined condition, \( b = 0.59 [0.44; 0.74], t(429) = 7.77, p < .001 \). The ES condition, \( b_{\text{diff}} = 0.25 [0.11; 0.38], p < .001 \), and the combined condition, \( b_{\text{diff}} = 0.37 [0.22; 0.51], p < .001 \), each outperformed REP, whereas the ES and combined conditions were on par with each other, \( b_{\text{diff}} = 0.12 [-0.03; 0.26], p > .05 \). Thus, we fully replicated the results of Study 2, suggesting that, at least in the present paradigm, time spent on intervention and number of conditioning trials may not be crucial in determining the magnitude of learning effects.

Study 6B

13 Raw data for Studies 6A–6C, permitting further analyses not reported here, are available for download from OSF (https://osf.io/jyk8c/).

14 The learning effects created in the congruent condition were overall smaller than in the incongruent condition. The intercept significantly differed from zero, \( b = 0.37 [0.28; 0.46], t(436) = 8.18, p < .001 \), indicating a strong baseline preference for Longfaces. Both the REP condition, \( b = 0.23 [0.10; 0.35], t(436) = 3.64, p < .001 \), and the combined condition, \( b = 0.28 [0.16; 0.41], t(436) = 4.38, p < .001 \), produced significant learning effects compared to control, whereas the ES condition did not, \( b = 0.12 [-0.01; 0.25], t(436) = 1.87, p = .061 \). Using a bootstrapping method for additional contrasts, we found no significant difference between REP and ES, \( b_{\text{diff}} = -0.10 [-0.22; 0.01], p > .05 \). The combined condition added significant value only to the ES condition, \( b_{\text{diff}} = 0.16 [0.04; 0.29], p < .01 \), but not to the REP condition, \( b_{\text{diff}} = 0.06 [-0.05; 0.16], p > .05 \), all by bootstrap.
The differences across the REP and ES conditions may have emerged due to the fact that the instructions provided for each activated different levels of construal (Trope & Liberman, 2010). Whereas instructions in the ES condition explicitly referred to social groups (e.g., “you will learn that a certain group of people is associated with pleasant things or unpleasant things”), thus encouraging high-level construal, instructions in the REP condition referred to the stimuli themselves (e.g., “you will see two types of drawings”), thus encouraging low-level construal. Even though this difference mirrors the way in which the two types of learning unfold under more ecologically realistic conditions, we probed whether removing it from the procedure would eliminate the superiority of the ES condition.

**Method**

In Study 6B, conducted with 603 participants on Project Implicit, we used the same stimuli as in Study 1 (Laapians vs. Niffians) but modified the instructions\(^\text{15}\) such that in the REP+ condition they explicitly referred to social groups and attributes (“you will see two types of names, representing two groups, and two types of drawings, representing two traits”), encouraging high-level construal, whereas in the ES- condition, they referred to exclusively to stimuli (“you will learn that certain names are associated with pleasant drawings or unpleasant drawings”), encouraging low-level construal.

**Results and discussion**

Even though the present study sought to tilt the balance in favor of the REP condition over the ES condition by facilitating abstract construal of the stimuli in the former and concrete construal of the stimuli in the latter, the ES- condition still produced a significantly stronger learning effect than the REP+ condition, \(t(599.9) = 2.39, p = .017, \text{Cohen’s } d = 0.19\). Thus, diff-

---

\(^{15}\) The full set of instructions is available for download from OSF (https://osf.io/jyk8c/).
ferent levels of construal created by the two sets of instructions are unlikely to account for the superiority of the ES condition observed in the previous studies.

**Study 6C**

The evaluative statements used in Studies 1–5 generated the expectation that the stimulus pairs would be shown to the participant later, whereas no such unfulfilled expectation was created in the REP condition. This feature of the ES condition raises the possibility that its superiority over the REP condition may be due to a Zeigarnik effect (Rajagopal, Raju, & Unnava, 2006; Savitsky, Medvec, & Gilovich, 1997; Zeigarnik, 1927), i.e., a memory advantage as a result of the interrupted nature of the task. In Study 6C, we sought to eliminate this potential confound.

**Method**

Study 6C was conducted on Project Implicit with 429 participants and using the same stimuli as Study 1 (Laapians vs. Niffians). In this study, we tested whether removing the expectation of upcoming stimulus pairings would modulate the learning effect created by the ES condition. Thus, participants were assigned to one of two conditions. The first learning condition was identical to the ES condition of Study 1 (expectation condition), whereas the other one did not create the expectation of upcoming stimulus pairings but rather simply informed participants that one group was good and the other group was bad (no expectation condition).

**Results and discussion**

In spite of the high-powered design, we found no difference across the expectation condition and the no expectation condition in terms of the strength of the implicit attitudes created, $t(426.94) = 1.21, p = .226, \text{Cohen’s } d = 0.03$. Thus, overall, Studies 6A–6C demonstrate that time

---

16 The full set of instructions is available on OSF (https://osf.io/jyk8c/).
spent on intervention, levels of construal, or a Zeigarnik effect cannot account for the results of Studies 1–5 reported above.

**Combined Analyses Across Experiments**

Because Studies 1–5 shared the same basic design, we aggregated the results across these studies meta-analytically (see Figure 1). First, we calculated Cohen’s $d$ effect size measures for each learning modality condition (compared to the control condition of the given study). Second, in order to arrive at an estimate of the uncertainty around the effect sizes aggregated across studies, we adjusted D scores for mean differences across studies and obtained 95-percent confidence intervals around the effect size estimates for each learning modality compared to control on the basis on 10,000 bootstrap samples.\(^{17}\) Finally, we investigated whether effect sizes are significantly different from each other by using an indirect testing approach. We drew 10,000 bootstrap samples for each pairwise comparison, i.e., REP vs. ES, REP vs. combined, and ES vs. combined, separately in the congruent and incongruent conditions.\(^{18}\) Next, we used the indirect testing approach described in more detail in the Results section of Study 1 to establish significant differences at the $p < .05$, $p < .01$, or $p < .001$ levels.

\(^{17}\) Due to the larger number of observations, this procedure yielded more stable estimates than bootstrapping Cohen’s $d$ values.

\(^{18}\) For the purpose of this analysis, all participants from Study 1 were included in the incongruent condition because of the neutral baseline, which allowed ample room for movement in both the Laapian–good and the Niffian–good condition.
Figure 1. Meta-analytic results from Studies 1–5. The x-axis shows learning modality conditions (REP = repeated evaluative pairings; ES = evaluative statements; Combined) and the y-axis shows standardized effect sizes (Cohen’s $d$s) comparing each learning modality to the control condition. Statistically significant differences between effect sizes (adjusted for mean D score differences across studies) are marked with asterisks, * = $p < .05$, ** = $p < .01$, and *** = $p < .001$, based on 10,000 stratified bootstrap samples.

Overall, learning effects were considerably stronger in the incongruent condition, Cohen’s $d = 0.71$ [63; 80], than in the congruent condition, $d = 0.27$ [17; 38], $p < .05$ by bootstrap. However, in spite of the difference in average effect sizes, the patterns of data by learning modality were similar across the congruent vs. incongruent conditions.

REP produced the smallest learning effect compared to baseline both in the congruent condition, $d = 0.13$ [0.02; 0.27], and in the incongruent condition, $d = 0.53$ [0.42; 0.65]. Moreover, in the incongruent condition, effect sizes for REP were variable across studies, ranging from 0.16 in Study 2 to 1.04 in Study 3. Overall, the learning effects produced by ES were larger and
less variable than the learning effects produced by REP, with $d = 0.27 \, [0.14; 0.40]$ in the congruent condition and $d = 0.84 \, [0.72; 0.96]$ in the incongruent condition, $ps < .01$ by bootstrap for the REP–ES contrasts. The combined condition also produced reliably large learning effects, with $d = 0.39 \, [0.27; 0.52]$ in the congruent condition and $d = 0.83 \, [0.72; 0.95]$ in the incongruent condition. The combined condition produced stronger learning effects than REP both in the congruent and the incongruent conditions, $ps < .001$ by bootstrap, but the ES–combined contrast did not reach significance in either case, $ps > .05$ by bootstrap, suggesting that even in these tests with greater power to detect differences, the combined condition reflects the contribution of ES with no additional boost from REP.

Finally, we probed whether these findings were mediated by the type of attitude object (novel social, including Studies 1 and 2; nonsocial, including Study 3; and existing social, including Studies 4 and 5). We found no differences across novel and existing social groups, with the ES and combined conditions outperforming the REP condition in both cases, $ps < .05$ by bootstrap and the combined condition offering no added value over the ES condition, $p > .05$. For nonsocial targets, the pattern of results was slightly different, with the only significant difference emerging between the REP and combined conditions, $p < .05$ by bootstrap.

**General Discussion**

In this project we conducted six experiments in order to answer two, as yet unexplored, questions about the acquisition and change of implicit attitudes. Building on large numbers of previous studies demonstrating that both repeated evaluative pairings (REP) and evaluative statements (ES) can lead to changes in implicit evaluation, we investigated whether the two learning modalities would create implicit attitudes on par with each other. Additionally, our ex-
periments were designed to probe whether the learning effects created by REP and ES are additive or redundant.

Regarding the first question of relative effectiveness, we made two contradictory predictions based on existing literature. On the basis of association formation theories positing that implicit attitudes should shift gradually over time (Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Strack & Deutsch, 2004) and empirical demonstrations of the superiority of REP in determining implicit attitudes when REP and ES are contradictory (DeCoster et al., 2006; Gregg et al., 2006; Moran & Bar-Anan, 2013; Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006a), we expected that REP might create stronger implicit attitudes than ES. Conversely, propositional theories (De Houwer, 2009; 2014; Hughes et al., 2011; C. J. Mitchell et al., 2009) and some empirical findings (Gast & De Houwer, 2013) suggested that REP and ES should produce learning effects of similar magnitude.

In fact, focusing on the incongruent conditions that allowed ample opportunity for attitudinal shift, we found a robust pattern of data that we had not *a priori* predicted based on either theoretical perspective. Unexpectedly and despite high-powered designs, REP created significant learning effects in only three out of five experiments, and the effect sizes produced by REP displayed considerable variability across studies. Even more surprisingly, ES resulted in consistently large shifts in implicit attitudes that were never inferior to and, in three out of five studies, even significantly exceeded the attitude change created by REP. These findings were unmediated by the domain of learning (social vs. nonsocial) and the status of the targets (well-known vs. novel). Moreover, they remained robust in the face of procedural variations designed to remove potential confounds, including time on intervention, levels of construal, and expectancy effects (Studies 6A–6C).
This pattern of results is surprising for a number of reasons. First, unlike in previous experiments where information about the attitude objects was highly diagnostic (Cone & Ferguson, 2015; Mann & Ferguson, 2015), participants in the present studies were merely informed that the targets would be paired with pleasant or unpleasant drawings, which is hardly on par with learning that a person is a child molester, the manipulation used by Cone and Ferguson (2015). Second, with new stimuli appearing every few seconds, if anything, the REP condition should have been more attention-grabbing, and thus more powerful, than the ES condition. Finally, our findings are difficult to reconcile with association formation theories of implicit attitudes, which have dominated the field ever since the beginnings of implicit social cognition research (Bargh, 1989; Fazio et al., 1986; Greenwald & Banaji, 1995) well into the present (Gawronski & Bodenhausen, 2006; Rydell, McConnell, Mackie, & Strain, 2006a; Strack & Deutsch, 2004).

For the second issue concerning the combined effects of REP and ES, we had envisaged three possible outcomes. Strict association formation theories (Rydell, McConnell, Mackie, & Strain, 2006a; Strack & Deutsch, 2004) predicted that ES should, at best, have minimal effects on implicit attitudes, rendering this question irrelevant; dual-process theories allowing for interactions between the explicit and implicit systems (Gawronski & Bodenhausen, 2006) predicted that the combination of REP and ES should result in additive, or at least subadditive, effects of each isolated learning modality; and propositional models (De Houwer, 2014; C. J. Mitchell et al., 2009) predicted that the combination of both attitude induction modalities should produce learning on par with their separable effects. Regarding this second question, the prediction made by propositional models was clearly borne out: The ES and combined modalities were statistically indistinguishable from each other across all five studies, and the combined condition did not produce larger learning effects than ES even when (a) REP created sizable learning in isolation.
(Studies 1, 3, and 4) and (b) when the combined condition included the same number of stimulus pairings as the REP condition (Study 6A). Ceiling effects cannot account for this finding; it emerged consistently irrespective of the strength of the prevailing attitude, including in cases where learning occurred in the opposite direction from a strong initial evaluation (e.g., for participants learning that elderly people or foreigners are good).

These findings are considerably easier to reconcile with propositional than with association formation theories of implicit attitudes. According to single-process theories (De Houwer, 2009; 2014; C. J. Mitchell et al., 2009), both REP and ES give rise to the same mental representation (e.g., “Squares are good and rectangles are bad”), and given the redundancy observed between the REP and the ES conditions, our data are clearly in line with this prediction. Even though the prediction of equal effects across REP and ES that we derived from single-process theories was not borne out, one might argue that the superiority of ES should, after all, not be that surprising from their perspective. In order for REP and ES to be able give rise to propositional representations about the valence of the attitude objects, participants need to make inferences about the stimuli to which they are exposed. However, the number of inferential steps is not the same across the REP and the ES conditions. In the former case, (1) participants are exposed to stimulus pairings from which (2) they deduce a rule (“Squares are always paired with pleasant; rectangles are always paired with unpleasant”), from which (3) they can infer the intended proposition (“Squares are good and rectangles are bad”). The ES condition, by contrast, does not require participants to undertake step (2), because the rule has already been provided to them. This difference could explain the lower average effectiveness of the REP modality and the
greater variability in its success.\textsuperscript{19} Even though our data do not conclusively demonstrate the validity of this reasoning, prior work has shown that the ability to correctly recall the instructions is a precondition for successful instructed conditioning effects (Gast & De Houwer, 2013; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a). The lack of conscious recollection, of course, precludes any further inferential steps.

The pattern of data observed in our studies is harder to reconcile with association formation theories of implicit attitudes. Strict association formation theories (Rydell & McConnell, 2006; Strack & Deutsch, 2004) posit that explicit and implicit attitudes are subserved by different learning mechanisms. Whereas explicit (i.e., propositional) attitudes can be acquired in one-shot learning episodes, implicit (i.e., associative) attitudes emerge from protracted exposure to contingencies in the environment. Contrary to these theories, our data show that not only are one-shot language-based learning episodes capable of creating implicit attitudes (De Houwer, 2006; Gast & De Houwer, 2013; Gregg et al., 2006; Van Dessel, De Houwer, Gast, & Smith, 2015b), these implicit attitudes are at least on par with, and usually even stronger than, implicit attitudes created by exposure to stimulus pairings. Thus, these data seem to militate against distinguishing between explicit and implicit attitudes on the basis of learning processes. Association formation theories that allow for interactions between the explicit and the implicit system, such as the associative–propositional evaluation (APE; Gawronski & Bodenhausen, 2006) model, also posit different underlying processes for explicit and implicit evaluation. Under this model, REP shift implicit attitudes directly, whereas ES first create explicit attitudes, and those explicit atti-

\textsuperscript{19} It is possible that the variability in the effectiveness of REP may have to do with the ease with which participants are able to spontaneously generate verbal labels for the two groups of stimuli. For instance, it is conceivable that REP did not create significant attitude change in Study 5 because participants failed to impose a verbal label on the quite disparate stimuli representing the foreign category. We thank an anonymous reviewer for pointing out this possibility. In a similar vein, it has been shown that providing verbal labels for social groups greatly facilitates social inference in four-year-old children (Baron, Dunham, Banaji, & Carey, 2014).
tudes, in turn, influence implicit evaluation. Although this reasoning is not fully incompatible with our results, three particular aspects of the data seem to pose somewhat of an explanatory challenge for it. First, if REP and ES create implicit attitudes via separate cognitive pathways, why are their effects redundant? Second, how does the APE model explain the fact that overall, ES outperformed REP, even though for the former, an additional cognitive step is posited to be intervening between learning and evaluation? Finally, how is it possible for implicit attitudes to change without corresponding changes in explicit attitudes (Study 4; see also Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a)?

It should be noted, however, that the results reported here do not directly speak to the content or format of the mental representations created by each of the two learning interventions under consideration. That is, although the redundancy of the two interventions suggests that ES and REP may give rise to the same mental representations, our studies are agnostic about the question of whether (a) both strengthen unqualified links between conceptual nodes in long-term memory (e.g., between elderly and good), (b) both form the basis of propositional representations with compositional structure and referential meaning (e.g., “Elderly people are good.”), or (c) some other possibility. Examples of one-trial associative learning abound both in the domain of animal learning (Dufort, Guttman, & Kimble, 1954; Jarvik & Essman, 1960) and human learning (Öhman, Eriksson, & Olofsson, 1975; Rock & Heimer, 1959). Thus, even though our results seem to militate against current associative theories from social cognition that posit that the updating of implicit attitudes can only happen in a piecemeal manner via incremental strengthening of unqualified links between conceptual nodes, they are by no means inconsistent with a potential theory of implicit attitudes that would allow for the possibility of the immediate
creation or updating of associations based on information presented in either verbal or nonverbal form. To date, however, no such theory of implicit evaluation has been proposed.

Somewhat unexpectedly, our results were not modulated by the novelty of the attitude objects. ES was more effective than REP and the two methods were redundant for both novel and existing social groups. Thus, it seems plausible that Van Dessel et al. (2015b) might have failed to obtain implicit attitude change based on approach/avoidance instructions because of the particular outgroup that they selected for investigation (i.e., African Americans). Due to the long history of racial animosity in the United States (Nosek, Banaji, & Jost, 2009), interventions that might mitigate bias against other outgroups might easily backfire when used in this particular context. More generally, our results show that ES-based interventions might be more effective than previously thought. Moreover, they caution against treating all outgroups as alike and suggest that different interventions might be differentially effective for different implicit outgroup biases. For instance, the hierarchy of interventions established by Lai et al. (2014) might change if one were to examine a different target group.

Even though our studies unequivocally demonstrate the effectiveness of ES and suggest that learning on the basis of stimulus pairings and verbal information may share the same underlying cognitive mechanism, they do not offer an immediate explanation for why ES are so powerful or through what (shared) mental processes they exert their effect. It is possible that the power of ES may have to do with the explanation outlined above in relation with propositional theories of implicit evaluation. Unlike repeated evaluative pairings, which are experienced over time and might leave room for interpretational ambiguities, ES are compact and, at least in our studies, express unequivocal rules (such as “squares = good” and “rectangles = bad”). Alternatively, the power of ES might not have been due to their compact propositional nature but rather
due to the fact that upon reading the instructions, participants mentally simulated (Markman, Klein, & Suhr, 2012) the upcoming stimulus pairings, thus making the learning experience more associative in nature. This hypothesis does not explain, however, (a) why simulated stimulus pairings were more effective in shifting implicit attitudes than experienced stimulus pairings and (b) why we observed the same level of learning across the self-paced (Study 6A) and the fixed-time (Study 2) ES condition, even though if this account is accurate, participants spent almost three times longer simulating stimulus pairings in the latter than in the former.

In summary, this project has demonstrated that, in opposition to predictions, ES are more effective than REP in shifting implicit attitudes and the effects of both learning modalities are fully redundant. These findings have a number of theoretical and possible practical implications. Crucially, they call into question the idea of a match between more conscious modes of thought as using language-based learning on the one hand and less conscious modes of thought being more sensitive to experience-based learning on the other hand. They also show that whatever the driving force behind REP-based learning, the effects of REP in shifting implicit attitudes are subsumed under the effects of ES-based learning. Finally, we hope that our results might inform future work probing the mechanisms behind successful debiasing interventions, both by drawing attention to the potential of language in mitigating implicit bias and by highlighting the importance of taking into account subtle differences among outgroups.

At the same time, one might wonder to what extent these results might have turned out differently given alternative interpretations of some key concepts and concomitant variations in procedures. Importantly, stimulus pairings in the world and instructions expressed through language seem to be fundamentally incommensurable, i.e., finding a common metric for the amount of information conveyed by each is far from a trivial task. In our studies, we overcame this prob-
lem by fixing time spent on intervention across learning conditions. Future studies might fix the strength of implicit attitude to be achieved and measure the amount of learning necessary to reach that prespecified goal. In addition, given the strong psychometric properties of the IAT both in general (Bar-Anan & Nosek, 2014) and specifically in the context of attitude induction via evaluative conditioning (Hofmann et al., 2010; Nosek et al., 2007), all studies reported here used the IAT as their dependent measure. However, considering the substantial amount of non-overlapping variance observed among implicit attitude measures (Bar-Anan & Nosek, 2014), future work may wish to seek corroborating evidence for these results using a different implicit task. Moreover, attitude strength as measured immediately after learning is not the only possible measure of effectiveness, and other operationalizations might lead to substantive conclusions different from the ones reached here. For instance, one would expect effective interventions to bring about durable learning effects, and there is reason to suspect that attitudes created by ES might decay faster than attitudes created by REP (Hughes & Barnes-Holmes, 2011). Finally, an effective intervention should result in attitudes that are resistant to counterinformation and can be activated effortlessly even in the absence of available cognitive resources. How REP and ES fare against each other along these dimensions remains to be investigated.
Paper 2 • Attitude Change via Repeated Evaluative Pairings versus Evaluative Statements: Shared and Unique Features
Abstract

When tested immediately, evaluative statements (ES; verbal information about upcoming categories and their positive/negative attributes) surprisingly shift implicit (IAT) attitudes more effectively than repeated evaluative pairings (REP; actual pairing of category members with positive/negative attributes). The present project (total $N = 5,317$) explored the shared and unique features of these two attitude change modalities by probing (a) commonalities visible in the extent to which propositional inferences created by ES infiltrate REP learning and (b) differences visible in performance of ES and REP learning over time. In REP, the number of stimulus pairings (varied parametrically from 4 to 24) produced no effect (Study 1), but verbally describing stimulus pairings as diagnostic vs. nondiagnostic did modulate learning (Study 2), suggesting that even REP give rise to some form of propositional representation. On the other hand, learning from ES decayed quickly, whereas learning from REP remained stable over time both within an immediate session of testing (Study 3) and following a 15-minute delay (Study 4), revealing a difference between these two forms of learning. Beyond their theoretical import, these findings may inform interventions designed to produce short- and long-term change in implicit attitudes.
Introduction

Humans are confronted with a fundamental and consequential type of decision numerous times a day: They must decide whether to approach or to avoid each of the myriad stimuli that they encounter in their environment. According to an influential idea as old as the concept itself, choices about approach and avoidance are driven by attitudes, that is, evaluations of stimuli along a positive–negative valence dimension: Positive evaluations result in approach behaviors and negative evaluations result in avoidance behaviors (Allport, 1935; Cacioppo, Harkins, & Petty, 1981; Eagly & Chaiken, 1998; Lewin, 1935). If decisions about approach and avoidance are to adaptively guide behavior, they must be made not only accurately but also quickly and effectively. As such, approach–avoidance decisions may, at least in part, be driven by implicit attitudes, or evaluations that are activated automatically upon encountering a stimulus (Bargh et al., 1992; Devine, 1989; Fazio et al., 1986; Greenwald & Banaji, 1995). Such automatically activated evaluations are usually measured via response interference tasks, as opposed to explicit attitudes, which are measured via self-report.20

Given their role in shaping affect, cognition, and behavior, the question of how implicit evaluations shift in the face of novel information is both of basic theoretical interest and of relevance to the more applied endeavor of producing enduring shifts in implicit evaluations of existing categories, such as social groups. In attempting to characterize implicit attitude change, the present work takes as its starting point two fundamentally different approaches that have been used to capture a wide range of phenomena in human learning, including concept learning

20 In line with the Fazio and colleagues (Fazio & Olson, 2003), we see the distinction between explicit and implicit attitudes as located primarily at the level of measures and do not make a priori assumptions about the nature of the mental representations or learning processes underlying implicit evaluations. In fact, we believe that the view advocated by De Houwer (2007; 2018), according to which evaluative conditioning should be conceptualized as an effect rather than a process, has significantly advanced evaluative conditioning research by allowing for the possibility that implicit measures may reflect propositional processes of attitude change. This work continues in that tradition.
(Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Kruschke, 1992), causal learning (Dickinson, 2001; Gopnik et al., 2004), and language acquisition (Griffiths & Kalish, 2010; Rumelhart & McClelland, 1987): the associative approach and the inferential approach.21

The associative approach relies on the premise that human learning arises from the incremental updating of associative representations in long-term memory as a result of repeated experiences with the environment. The associative approach encompasses powerful and diverse ideas such as neural networks (Rumelhart et al., 1986), the Rescorla–Wagner model of Pavlovian conditioning (Rescorla & Wagner, 1972), and temporal difference learning (Sutton & Barto, 1998). Even though specific instantiations of this approach obviously differ in their details of implementation, they are unified by a few core ideas, including (a) relatively slow and iterative updating that progressively reduces the difference between an existing representation and a target representation based on prediction errors; (b) learning as an informationally encapsulated process that preferentially relies on certain kinds of input while ignoring others (see Fodor, 1983); and (c) long-term memory as a collection of associative strengths.

By contrast, the inferential approach relies on the radically different premise that human learning arises from a process of theory testing in which the learner entertains multiple competing hypotheses about some problem and revises her beliefs about the plausibility of each hypothesis upon encountering new data. Like the associative approach, the inferential approach encompasses powerful and diverse ideas such as the child as scientist (Gopnik, 1996), Bayesian models of human cognition (Tenenbaum, Kemp, Griffiths, & Goodman, 2011), and the pedagogical

21 We do not argue that the two approaches are mutually exclusive. For instance, a recent proposal posits that Pavlovian conditioning operates not over observable cues and outcomes but rather over latent causes inferred by the learner (Gershman, Norman, & Niv, 2015). Moreover, given that neural nets are universal function approximators, they can be used to implement symbolic, e.g., Bayesian, approaches to learning. However, these observations do not invalidate the fact that the two approaches have traditionally been seen as competing in providing a description and explanation of human learning.
view of learning and teaching (Csibra & Gergely, 2009). Again, in spite of important differences in implementation, a few core ideas seem to unify the inferential approach, including (a) the possibility of drawing quick inferences from sparse data; (b) learning as an informationally promiscuous process that integrates information from all relevant sources (see Fodor, 1983); and (c) long-term memory as a store of symbolic representations.

Ever since its inception, the very concept of implicit evaluations has been intimately linked with the idea of associative learning and representation. Early investigations of implicit race attitudes drew heavily from the idea of spreading activation in associative networks (Fazio et al., 1986) and relied on methods that had been implemented to study conceptual associations in nonsocial category representation (Neely, 1976). In fact, one of the most widely applied measures of implicit evaluations, also used in the present studies, has the term “association” in its name (Greenwald et al., 1998). As such, it should not be surprising that the study of implicit social cognition has traditionally been dominated by associative theories (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004). These theories may be regarded as applications of the associative approach to human learning to the more specific problem of implicit attitude acquisition and change. Even though these theories differ from each other in numerous details, they converge on a generally consistent view that implicit attitudes should shift (a) slowly and incrementally, (b) preferentially as a result of direct experience with the environment, and (c) by creating conceptual associations in long-term memory.

Specifically, under associative theories, if implicit attitudes toward a social group, such as the Laapians (Gregg et al., 2006), are to be shifted in a positive direction, members of this group should be paired with intrinsically positive stimuli for a sufficient (large) number of trials,
as a result of which the links between the conceptual nodes LAAPIAN and GOOD will be strengthened in long-term memory. Associative theories describe implicit attitude change as slow and incremental: Smith and DeCoster (2000) mention a memory system that “[…] build[s] up knowledge slowly and incrementally […]” (p. 111); Strack and Deutsch (2004) talk about a simple memory system that “[…] slowly forms enduring, nonpropositional representations […] over many learning trials” (pp. 223–224); and, finally, according to Rydell and McConnell (2006), “[…] implicit attitudes form and change through the use of slow-learning, associative reasoning” (p. 995; see also McConnell & Rydell, 2014).

The idea of informational encapsulation is also reflected in associative theories of implicit evaluation. According to Smith and DeCoster (2000), “[…] the two [associative vs. propositional] processing modes tap separate databases that represent knowledge in different formats […]”, with the associative mode drawing “[…] solely on patterns built up over time in the slow learning memory system” (p. 127). The same authors rule out the possibility that a newly learned rule may be able to affect associative processing. Similarly, McConnell and Rydell (2014) posit “two dissociable systems of knowledge” (p. 204): a slow-learning system relying on associative information and a fast-learning system relying on verbal information (Rydell & McConnell, 2006). In the same vein, Strack and Deutsch (2004) hypothesize that whereas “[…] explicit measures tap into people’s knowledge or beliefs, implicit measures tap into their associative structures” (p. 239).

Finally, and related to both the idea of slow learning and informational encapsulation, associative theories propose that implicit and explicit evaluations are subserved by separate memory systems. For instance, Smith and DeCoster (2000) argue that whereas implicit attitudes are “[…] represented by associations built up in a connectionist distributed memory […], explicit
beliefs are symbolically represented” (p. 122). Strack and Deutsch (2004) propose a different kind of memory dissociation, with implicit evaluations relying on long-term memory and explicit evaluations relying on a temporary memory store. However, in spite of such differences, all associative theories reviewed here share the basic premise that implicit attitudes arise from conceptual associations stored in long-term memory.

More recently, propositional theories (De Houwer, 2009; 2014; 2018; C. J. Mitchell et al., 2009), which can be conceptualized as applying the inferential approach to human learning to implicit evaluation, have begun to challenge the primacy of the associative view. Propositional accounts are fundamentally different from associative accounts in the learning processes and mental representations posited to underlie implicit evaluations. Specifically, under the propositional view, implicit and explicit attitudes are hypothesized to differ from each other solely in terms of the level of automaticity with which evaluative representations are activated, with implicit measures such as the IAT (Greenwald et al., 1998) forcing participants to respond relatively automatically and explicit measures imposing no such requirement.

When it comes to learning, the propositional account, unlike its associative counterparts, posits that implicit evaluations can shift (a) quickly, without the need for incremental updating, (b) in an informationally promiscuous manner, taking into account all available evidence (including both stimulus pairings encountered in the environment and verbal information), and (c) by creating symbolic and propositional representations in long-term memory. Specifically, in a clear deviation from associative accounts, De Houwer (2009) suggests that “[…] the generation and evaluation of propositions will [not] always be slow; if necessary, it can be done expeditiously” (p. 10), and such propositions may, in turn, be activated automatically on implicit measures of evaluation such as the IAT used in the present studies. Moreover, according to De Houwer
propositions about events can be formed not only on the basis of the repeated experience of those events but also as the result of a single instruction or inference concerning those events” (p. 344). This idea of informational promiscuity is in opposition to the core tenet of associative theories according to which implicit attitudes should be exclusively, or at least preferentially, responsive to stimulus pairings experienced in the environment. Finally, De Houwer (2018) also does away with the idea of a separate associative memory store driving responding on implicit measures. In fact, he proposes that even in situations where the procedure itself may be described as associative, including repeated presentations of stimulus pairings, such pairings “[…] can influence liking only after a proposition has been formed about the relation between the stimuli” (p. 3).

What kind of information is most effective in shifting implicit evaluations?

Associative and propositional theories of implicit evaluation both predict that repeated evaluative pairings (REP) of category members with intrinsically valenced stimuli should shift implicit attitudes. Indeed, evidence in line with this prediction has been produced by multiple labs and involving many specific instantiations of the REP intervention (Gibson, 2008; Grumm et al., 2009; Hughes & Barnes-Holmes, 2011; C. J. Mitchell et al., 2003; Olson & Fazio, 2001; 2002; 2006; Prestwich et al., 2009; for a meta-analysis see Hofmann et al., 2010). However, associative and propositional theories differ in their predictions about the effectiveness of evaluative statements (ES), i.e., informing participants about upcoming stimulus pairings without ever presenting such stimulus pairings. Surprisingly, and contrary to the informational encapsulation idea central to many associative theories, a series of recent studies have demonstrated significant changes in implicit evaluation as a result of purely verbal instructions, without any direct experience involving intrinsically valenced stimuli (De Houwer, 2006; Gast & De Houwer, 2013; Van
Dessel, De Houwer, Gast, & Smith, 2015b; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a; Van Dessel, Gawronski, Smith, & De Houwer, 2017a; Van Dessel, Mertens, Smith, & De Houwer, 2017c).

If REP and ES both have the ability to shift implicit attitudes, such results raise the question of relative effectiveness, which informs both the theoretical debate regarding the nature of the learning processes from which implicit evaluations emerge and efforts aimed at creating change in real-world evaluations. In line with such theoretical and practical interest, recent work has investigated whether stimulus pairings experienced in the environment or mere verbal instructions produce stronger evaluative learning (Kurdi & Banaji, 2017). In the REP condition of Kurdi and Banaji’s studies, participants were exposed to pairings of members of each target category (conditioned stimuli; CS) with intrinsically positive or negative images (unconditioned stimuli; US). In the ES condition, participants were informed about upcoming stimulus pairings without actual exposure. Evaluative statements were consistently more effective in shifting implicit attitudes than repeated evaluative pairings, even in the face of procedural changes intended to enhance learning in the REP condition or weaken learning in the ES condition. Such results are difficult to explain within the framework of associative theories, which propose that implicit evaluations should rely exclusively (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) or at least preferentially (McConnell & Rydell, 2014) on stimulus pairings experienced in the environment.

By what kind(s) of process is evaluative learning produced?

Beyond the relative effectiveness of REP vs. ES in shifting implicit evaluations, Kurdi and Banaji (2017) also explored the learning effects produced by a combination of the ES and REP interventions. They found that a third learning condition in which descriptions of stimulus
pairings were followed by the presentation of actual stimulus pairings (ES + REP) never significantly outperformed the ES condition in isolation, suggesting that direct experience with stimulus pairings did not add any value to purely verbal instructions. This result is unsurprising from the perspective of propositional theories: If both REP and ES give rise to the same propositional inference about the attitude object, their effects on implicit evaluations should be redundant. However, it is difficult to accommodate this result under associative theories given that these theories posit that implicit evaluations should not be influenced by purely verbal manipulations (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004).

Kurdi and Banaji (2017) concluded that, in spite of reasonable expectations to the contrary, the data were more closely aligned with propositional theories of implicit evaluation. Specifically, their findings strongly suggest that in humans, higher-level reasoning can insert itself into learning even from repeated presentations of pairings between categories and attributes. In other words, even stringent conditions of (supraliminal) associative learning may involve inferential processes that gave rise to propositional representations. Such a conclusion is all the more surprising given that learning from repeated evaluative pairings has often been characterized as mediated by low-level stimulus-driven processes (Baeyens, Eelen, Crombez, & van den Bergh, 1992; Baeyens, Vansteenwegen, & Hermans, 2009; Gawronski & Bodenhausen, 2006; Gawronski, Balas, & Creighton, 2013; Levey & Martin, 1975; Martin & Levey, 1978; Rydell & McConnell, 2006; Strack & Deutsch, 2004).

In light of this unexpected yet robust finding, replicated six times by Kurdi and Banaji (2017), the first part of the present paper reports two further tests of the process by which repeated evaluative pairings lead to changes in implicit evaluation. In Study 1, we investigated the effects of the number of stimulus pairings presented during learning. In Study 2, we probed wheth-
er a purely verbal manipulation (describing the stimulus pairings as diagnostic vs. nondiagnostic of the underlying nature of the target categories) significantly modulated learning effects from repeated evaluative pairings. We chose to study the effects of these two potential moderators of evaluative learning from REP for three reasons. First, whether learning is \((a)\) fast or slow and \((b)\) informationally promiscuous (i.e., drawing on all available data) or informationally encapsulated (i.e., restricted to a specific kind of informational input; see Fodor, 1983) is of basic theoretical interest. Findings regarding these features of the learning process stand to constrain any future theory of implicit attitude acquisition. Second, empirical support for both sets of predictions (quick and informationally promiscuous vs. slow and informationally encapsulated learning) seemed conceivable in light of existing evidence. Third, current associative and propositional theories of implicit evaluation differ in terms of the degree to which they are compatible with different patterns of empirical data.

**Number of stimulus pairings presented during learning**

Existing empirical evidence regarding the effect of number of CS–US pairings on evaluative learning is mixed. A number of studies have investigated this issue using explicit measures of attitude (Baeyens et al., 1992; Levey & Martin, 1987; Stuart, Shimp, & Engle, 1987). Baeyens and colleagues (1992) have found a quadratic effect of number of stimulus pairings for positively valenced USs and a linear effect for negatively valenced USs. By contrast, Levey and Martin (1987) concluded that number of stimulus pairings was inconsequential; however, no statistical evidence in favor of this proposition was put forth. Finally, Stuart and colleagues (1987) found nonsignificant linear trends using some measures of attitude but no effect using others. However, whether the effects obtained using explicit measures of evaluation (to the extent that such effects
are, indeed, reliable) would generalize to implicit attitudes is an open empirical question and in the present research we will have the opportunity to test it.

Evidence obtained using implicit measures of attitude has been similarly inconclusive. Rydell, McConnell, Strain, Claypool, and Hugenberg (2006b) found that the number of counter-attitudinal statements linearly predicted the amount of change in implicit attitudes toward a novel social target in an impression formation task. However, it should be noted that this finding was obtained using a learning intervention involving (a) verbal, rather than pictorial, stimuli, (b) a single novel target rather than a novel group, and (c) positive and negative feedback rather than passive exposure to stimuli. As such, the results obtained by Rydell et al. (2006b) may not readily generalize to the REP manipulation implemented here. More relevant for the present purposes, Hu, Gawronski, and Balas (2016) observed no statistically significant difference in implicit attitudes as a result of eight vs. 24 stimulus pairings and the meta-analysis by Hofmann et al. (2010) found no significant influence of the number of stimulus pairings presented on evaluative conditioning effects. However, (a) the meta-analysis did not examine explicit and implicit attitudes separately, (b) by definition, meta-analysis can provide only correlational, but not causal, evidence, and (c) lack of statistical significance in a frequentist framework should not be accepted as positive evidence for the absence of an effect (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wilson & Miller, 1964).

When it comes to current associative and propositional theories of implicit evaluation, both sets of theories predict that repeated evaluative pairings should shift implicit attitudes. However, each set of theories makes unique predictions about how such learning should unfold as the number of stimulus pairings increases. As reviewed above, associative theories of implicit evaluation share the core idea, also present in more general associative approaches to human
learning, that implicit attitudes should shift slowly and incrementally. If this is, indeed, the case, the number of stimulus pairings presented in the REP condition should modulate learning effects: A relatively small number of stimulus pairings may not produce any change, and increasing numbers of stimulus pairings should result in increasingly powerful learning effects. By contrast, in line with more general ideas about humans’ ability to make quick and accurate inferences even from noisy and sparse observations, the propositional account allows for the possibility that asymptotic performance may be reached after only a few stimulus pairings. As such, repeated presentations of redundant evaluative information may not result in additional learning.

**Diagnosticity**

Prior work has demonstrated that a single piece of extremely negative information deemed highly diagnostic of a person’s character (e.g., being a child molester or someone who enjoys torturing animals) is sufficiently powerful to reverse previously positive implicit attitudes formed as a result of a hundred behavioral statements (Cone & Ferguson, 2015). But can a purely verbal manipulation of diagnosticity (describing stimulus pairings as reflecting the underlying true nature of the target groups vs. having been randomly generated by the computer) have an effect on learning from repeated evaluative pairings? Existing empirical evidence on the issue is contradictory.

Some studies have provided evidence for the idea that implicit attitude acquisition from repeated evaluative pairings is informationally promiscuous in that it takes into account all available information, regardless of its format. As such, describing stimulus pairings as invalid (Peters & Gawronski, 2011) or as representing an opposition, rather than equivalency, relationship (Zanon et al., 2014) has been found to significantly modulate evaluative learning as reflected by implicit attitude measures. Moreover, verbally describing stimulus pairings as unrelated, predictive,
or causal also appears to modulate implicit attitude acquisition (Hughes, Ye, Van Dessel, & De Houwer, 2018). These results seem to suggest that, just like validity or type of relationship, diagnosticity should influence learning effects from repeated evaluative pairings.

However, in other studies, implicit attitude acquisition from repeated evaluative pairings appears to be relatively informationally encapsulated. For instance, Rydell, McConnell, Mackie, and Strain (2006a) found that implicit attitudes toward a novel target uniquely reflected subliminally presented REP, with no effect of supraliminally presented ES of the opposite valence (but see Heycke, Gehrmann, Haaf, & Stahl, 2018). Similarly, implicit attitudes do not consistently reflect verbal information about the nature of the relationship between upcoming stimulus pairings in a REP paradigm (Hu et al., 2016; Moran & Bar-Anan, 2013), and even when they do, CS–US contingencies can still have an effect over and above relational information (Moran, Bar-Anan, & Nosek, 2016). This latter set of findings would suggest that REP-based evaluative learning effects should not be modulated by verbal information on diagnosticity.

These two sets of results align with the predictions of propositional vs. associative theories, respectively. Propositional theories suggest that because verbal and nonverbal manipulations influence implicit evaluations via the same process of propositional inference, they should interact in producing evaluative learning. By contrast, associative theories posit that implicit attitude acquisition is relatively informationally encapsulated and, as such, should preferentially rely on stimulus pairings encountered on the environment, with little or no effect of verbal manipulations.

**What kind of information is most effective in durably shifting implicit evaluations?**

To gain further insight into the nature of the evaluative learning processes arising from interventions based on stimulus pairings vs. mere verbal instructions, the second set of experi-
ments reported here probed the temporal stability of implicit attitudes created via REP vs. ES. Specifically, Study 3 used a new analytic strategy to examine whether implicit attitude strength decayed differentially as a function of learning conditions during a single session of the Implicit Association Test (IAT; Greenwald et al., 1998), administered immediately after learning. Study 4 investigated the effects of time more directly by imposing a 15-minute delay between the learning and test phases of the experiment.

Why might time differentially affect implicit attitudes acquired via REP vs. ES and what are the theoretical implications of such potential differences? The associative perspective on implicit evaluation posits that implicit attitudes should shift preferentially in response to stimulus pairings experienced in the environment. Although the results of Kurdi and Banaji (2017) seem to contradict this idea, it could be argued that the immediate effects of learning manipulations observed there should be characterized as arising from temporary shifts in concept accessibility rather than as genuine conceptual change. As such, the present studies may offer a more stringent test of the relative effectiveness of evaluative learning based on REP vs. ES.

By contrast, under the propositional perspective, implicit and explicit evaluations are generally hypothesized to rely on shared memory representations. As such, studies on differences in episodic memory for personally experienced vs. verbally described events may be instructive regarding the effects that should emerge on implicit measures of evaluation. For instance, in a study conducted by Toglia, Shlechter, and Chevalier (1992), some participants directly experienced a staged event, whereas others were exposed to a verbal description of the same event. When surprise tests of recall and recognition for details of the event were administered imme-

22 However, as a reviewer of this work pointed out, propositional theories of implicit evaluation do not explicitly make this prediction. We concur with the reviewer’s judgment on this point.
ately, they revealed superior performance in the verbal description condition; however, following a delay, performance deteriorated sharply in the verbal description condition but remained unaffected in the direct experience condition (Baggett, 1979; Baggett & Ehrenfeucht, 1983; Beentjes & van der Voort, 1991; see also Larsen & Plunkett, 1987). To the extent that this effect generalizes from explicit measures of episodic memory to implicit measures of evaluative learning, REP should produce more durable changes in implicit evaluation than ES, despite the initial superiority of the latter (for some initial evidence see Hughes & Barnes-Holmes, 2011).

To summarize, the associative and propositional views on implicit evaluation both predict that learning from repeated evaluative pairings should be superior to learning from evaluative statements provided that the test of learning is administered following a delay. The associative perspective makes this prediction on the basis of a posited match between the format of the information (stimulus pairings) and the format of the memory representation (conceptual associations). By contrast, the propositional perspective allows for reliance on work on episodic memory to make the same prediction; however, this prediction is made for a completely different reason. Specifically, a crucial difference between REP and ES is that the former requires participants to generate inferences about the valence of each target category. Propositions generated by participants, as opposed to propositions provided to them in verbal form, have been hypothesized to result in superior retention on explicit memory tasks (Baggett, 1979; for related findings see Greenwald & Banaji, 1989; MacLeod & Bodner, 2017; Slamecka & Graf, 1978).

As such, the associative and propositional perspectives differ in their predictions for the combined ES + REP condition: If the effect is based on preferential reliance of implicit evaluations on environmental associations, then the ES + REP condition should behave the same way as REP isolation given that both include the REP manipulation; however, if the effect is based on
a difference between experimenter-provided vs. participant-generated inferences, then the ES + REP condition should behave the same way as ES in isolation given that the combined condition involves no participant-generated inferences.

Finally, beyond their theoretical implications, Studies 3 and 4 may also shed new light on recent attempts to create durable change in implicit attitudes (Devine, Forscher, Austin, & Cox, 2012; Forscher et al., 2017; Lai et al., 2016; Mann & Ferguson, 2015; 2017). Even though temporary malleability in implicit attitudes within a single experimental session has been demonstrated numerous times and using myriad different manipulations (Blair, 2002; Lai et al., 2014), ranging from social influence (Lowery, Hardin, & Sinclair, 2001) to context effects (Wittenbrink, Judd, & Park, 2001) and from exposure to counterstereotypic exemplars (Dasgupta & Greenwald, 2001) to implementation intentions (Stewart & Payne, 2008), successful demonstrations of long-term change in implicit evaluation have been conspicuously absent from the literature (for some notable exceptions in the context of implicit evaluations of single individuals see Mann & Ferguson, 2015; 2017). For instance, when Lai et al. (2016) measured the effectiveness of the most immediately successful manipulations following a 24-hour delay, none of them had any appreciable impact on implicit attitudes toward African Americans. Given that the studies reported here used a timeframe of minutes rather than hours or days, they have the potential to offer some insight into the time course of decay in implicit attitudes before learning effects dissipate entirely.

**Overview of the present project**

Implicit attitudes have been robustly demonstrated to shift in response to both repeated evaluative pairings of category members with positive and negative attributes (Gibson, 2008; Grumm et al., 2009; Hughes & Barnes-Holmes, 2011; Kurdi & Banaji, 2017; C. J. Mitchell et
al., 2003; Olson & Fazio, 2001; 2002; 2006; Prestwich et al., 2009) and evaluative statements, i.e., purely verbal manipulations merely describing upcoming stimulus pairings without actual exposure (De Houwer, 2006; Gast & De Houwer, 2013; Kurdi & Banaji, 2017). In an attempt to improve understanding of the processes by which evaluative learning is achieved in response to repeated evaluative pairings, the present project investigated two possible moderators of this learning effect: In Study 1, we varied the number of stimulus pairings to which participants were exposed and, in Study 2, we manipulated whether stimulus pairings were described to participants as randomly generated or reflecting the true underlying character of the target categories. Moreover, in Studies 3 and 4, we probed the temporal stability of learning from REP and ES, as well as from a combined ES + REP condition.

These manipulations were chosen for a number of reasons. First, they provide evidence on basic features of the learning processes underlying implicit attitude acquisition, including its temporal unfolding, informational inputs, and stability over time. As such, the results of the present work have the ability constrain any future theory of implicit evaluation. Second, existing empirical evidence on the effects of the chosen manipulations was mixed (Studies 1–2) or scarce (Studies 3–4). Third, although no single study or even set of studies may be sufficient to conclusively arbitrate between existing associative vs. propositional theories of implicit evaluation, certain patterns of data are nonetheless easier or more difficult to reconcile with different theoretical traditions. Fourth, the results of the current work stand to inform efforts that are aimed at designing interventions to produce durable change in implicit attitudes, with or without a desire to advance basic theory.

Study 1
Prior work has demonstrated that repeated evaluative pairings of category members with
valenced images can produce shifts in implicit attitudes toward the categories involved (Gibson,
2008; Grumm et al., 2009; Hughes & Barnes-Holmes, 2011; Kurdi & Banaji, 2017; C. J. Mitch-
ell et al., 2003; Olson & Fazio, 2001; 2002; Prestwich et al., 2009). The present study probed a
key moderator of this effect, investigating whether the number of stimulus pairings presented to
participants modulates the strength of the implicit attitudes created. On the one hand, updating
may unfold slowly as a result of incrementally reducing the difference between a target represen-
tation and the information provided by each stimulus pairing. Alternatively, implicit evaluations
may be updated via quick conceptual change, especially in the face of information that provides
a clear learning signal.

Method

Participants and design. 1,243 volunteers were recruited from the Project Implicit edu-
cational website (http://implicit.harvard.edu) to participate in the study. In line with the standard
scoring procedure (Greenwald et al., 2003), participants who did not complete the IAT (Green-
wald et al., 1998) and participants with a response latency of 300 ms or less on at least 10% of
IAT trials (N = 43) were excluded from all further analyses. Moreover, consistent with estab-
lished practice in research on evaluative learning (Van Dessel, De Houwer, Gast, & Smith,
2015b; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a; Van Dessel, Gawronski,
Smith, & De Houwer, 2017a), participants who provided an inaccurate response on a manipula-
tion check probing explicit recollection of the learning phase (N = 55) were also eliminated from
consideration. Participant exclusions resulted in a final sample size of N = 1,145 (N = 693 female
and N = 417 male participants, mean age = 39.56, SD = 17.05). This sample size provides .80
power to detect a small effect of r = .08.
All participants underwent an attitude induction procedure using repeated evaluative pairings (REP) of category members with positive and negative images. In a between-subjects design, each participant was randomly assigned to be exposed to a certain number of stimulus pairings, varied parametrically from four to 24. The median number of participants in each cell was 53 ($SD = 6.45$). There was no evidence for differential attrition across number of stimulus pairing conditions, as indicated by a lack of correlation between number of stimulus pairings and number of participants in each cell of the design, $\rho = -.036, p = .875$.

**Materials.** Line drawings depicting the faces of young White men were used as conditioned stimuli (CS). In order to enable easy categorization, the faces were manipulated to vary along a salient perceptual dimension: *Long faces* had a length-to-width ratio of 2:1, whereas *square faces* had a length-to-width ratio of 4:3. Faces were selected to serve as CS because they communicate a wealth of social information (Cogsdill et al., 2014; Willis & Todorov, 2006) and have a demonstrated ability to serve as the targets of evaluative learning (Hehman et al., 2015).

Positive unconditioned stimuli (US) included line drawings of a flower, a heart, a cone of ice cream, a sun, and a beach, and negative US included line drawings of a frowny face, a fleeing man, a snake, a terrorist, and an ant. Both the CS and US have been used successfully in the context of attitude induction in previous work (Kurdi & Banaji, 2017).

**Procedure and measures.** The study was administered entirely online and consisted of (a) a *learning phase* in which implicit attitudes were induced toward the target categories via exposure to REP, with the number of stimulus pairings varied between participants, and (b) a *test phase* in which implicit attitudes and explicit attitudes toward the targets as well as memory for stimulus pairings were assessed.

---

23 All stimuli are available for download from the Open Science Framework (OSF; https://osf.io/serq4/).
Learning phase. Participants were informed that in the course of the experiment they would see two types of faces and two types of drawings and were instructed to learn the general relationship between a certain type of face and a certain type of drawing without focusing on particular exemplars.\textsuperscript{24} Crucially, the instructions did not use verbal labels to refer to the groups and did not mention stimulus valence. Next, participants were introduced to the full set of conditioned and unconditioned stimuli. Finally, participants were exposed to a standard evaluative conditioning paradigm (Levey & Martin, 1975; Martin & Levey, 1978). Each trial consisted of a CS (a line drawing of a long face or a square face) and a US (a line drawing of an intrinsically positive or negative object) presented side-by-side on the screen for 2,500 ms. Each trial was followed by an intertrial interval of 1,000 ms. Depending on their random assignment to condition, participants received different numbers of stimulus pairings, ranging from four to 24. Previous work revealed an overall implicit preference for long faces over square faces at baseline (Kurdi & Banaji, 2017). In order to allow for learning to unfold without ceiling effects, the attitude induction manipulation was designed to move participants away from this baseline, i.e., square faces were always paired with positive stimuli and long faces were always paired with negative stimuli.\textsuperscript{25}

Test phase. In the test phase, participants completed (a) an IAT (Greenwald et al., 1998) measuring their implicit attitudes toward long faces and square faces, followed by (b) two feeling thermometer measures used to construct a relative measure of explicit preferences, and finally (c) a manipulation check probing participants’ recollection of their previous learning.

\textsuperscript{24} The verbatim text of the instructions is available for download from OSF (https://osf.io/serq4/).
\textsuperscript{25} For exploratory purposes, the opposite pairings were also included in the study but are not discussed here. Interested readers may conduct their own analyses using the raw data made available on OSF (https://osf.io/serq4/).
Implicit attitudes. Participants completed a standard five-block IAT (Greenwald et al., 1998) as a measure of implicit attitudes toward the two target groups (individuals with square vs. long faces). In the first practice block (20 trials), participants sorted positively and negatively valenced words. In the second practice block (20 trials), participants sorted the images of square-faced and long-faced individuals that were used as CS in the learning phase. In the first critical block (40 trials), positive words and square-faced individuals were mapped onto one response key, whereas negative words and long-faced individuals were mapped onto the other response key. In the third practice block (20 trials), participants learned the new assignment of target groups to response keys. Finally, in the second critical block (40 trials), the combined task was repeated with the opposite assignment of target groups to valence (long-faced/good and square-faced/bad). Given our interest in relative attitude strengths across different numbers of conditioning trials without comparison to an absolute standard, the order of critical blocks was kept constant across participants, with the congruent (square faces–good/long faces–bad) block always presented first. Performance on the IAT was assessed using the scoring algorithm recommended by Greenwald, Nosek, and Banaji (2003), with positive D scores indicating implicit evaluations in line with the attitude induction. Using the split-half correlation method recommended by Kurdi et al. (2018), an estimate of $R = .77$ was obtained for the internal consistency of the IAT, which is comparable to but somewhat lower than the estimate reported by Bar-Anan and Nosek (Bar-Anan & Nosek, 2014).

Explicit attitudes. Participants responded to two feeling thermometer items, one of them asking how warmly they felt toward square-faced individuals and the other one asking how

---

26 Positive words used as stimuli on the IAT included love, peace, joy, happy, sweet, glory, and success, and negative words included hate, war, devil, bomb, bitter, agony, and failure.
warmly they felt toward long-faced individuals. Responses were provided on 10-point scales, anchored by “extremely warmly” and “extremely coldly.” Given the focus of the current work on implicit attitudes, explicit attitudes are not discussed any further.27

Manipulation check. Given that contingency awareness is a major moderator of attitude change via repeated evaluative pairings (Hofmann et al., 2010), participants were asked to complete a manipulation check item, which instructed them to recall what they had learned at the beginning of the study. Response options included “square faces are good and long faces are bad” (accurate response), “long faces are good and square faces are bad” (inaccurate response), and “nothing that would have indicated whether the groups are good and bad” (intermediate response).28

Results29

Number of stimulus pairings. The focal question of Study 1 was whether the number of stimulus pairings presented to participants would modulate the strength of the implicit attitudes created as a result of a learning intervention involving repeated evaluative pairings of category members with valenced images. This question was probed by fitting a linear regression to the data, with IAT D score as the response variable and number of stimulus pairings presented as the sole predictor (see Figure 2).
Figure 2. Implicit attitude strength (measured using IAT D scores) as a function of the number of conditioning trials presented (Study 1). Positive scores imply implicit attitudes in line with the learning manipulation. Error bars show 95-percent confidence intervals comparing each effect size to zero. The dashed line shows neutrality.

We obtained a significant intercept, $b_0 = .31 [.25, .37], t(1122) = 10.56, p < .001$, indicating considerable deviation from neutrality among participants who had been exposed to as few as four CS–US pairings. Crucially, evidence for linear change in attitude strength as a result of increasing the number of stimulus pairings would be provided by a significant slope parameter, which we failed to obtain, $b_1 = .00 [-.00, .01], t(1122) = 1.63, p = .102$. Visual inspection revealed no obvious nonlinearity in the data that could have accounted for this result.

**Summary of supplementary analyses.** In Supplement 1 ([https://osf.io/serq4/](https://osf.io/serq4/)), we report additional analyses of the same data. To summarize, we demonstrate that (a) the evaluative
learning effects reported here are significantly modulated by participants’ conscious recollection of the learning manipulation (see Hofmann et al., 2010); (b) the focal result reported above, suggesting no effect of number of stimulus pairings, emerges in a Bayesian linear regression using uninformative priors; (c) in a model including all participants without exclusions based on contingency awareness, only contingency awareness but not number of stimulus pairings emerges as a significant predictor of implicit attitude strength; (d) a significant difference in implicit attitudes emerges between the incongruent learning condition reported here (square faces paired with positive and long faces paired with negative stimuli) and the congruent learning condition (long faces paired with positive and square faces paired with negative stimuli) and this difference is consistent across number of stimulus pairings; and (e) the effect of number of stimulus exposures is not significant in the congruent learning condition, thus mirroring the results reported here. Overall, these supplementary analyses demonstrate the robustness of the focal result of no effect of number of stimulus pairings to (i) participant exclusions, (ii) stimulus effects, and (iii) analytic frameworks.

Discussion

The present study probed the effects of the number of stimulus pairings on the strength of implicit attitudes created toward novel targets. Experiencing four stimulus pairings was sufficient to induce novel implicit attitudes, with no difference observed between implicit evaluations that emerged as a result of 24 compared to four CS–US pairings (for a similar result with two between-participant conditions, see Hu et al., 2016). Based on a Bayesian linear regression reported in more detail in Supplement 1 (https://osf.io/serq4/), the most likely posterior value of the slope parameter is zero, thus providing positive evidence in favor of attitude stability across number of stimulus pairings, rather than incremental updating. Unlike the number of stimulus pairings, par-
Participants’ conscious recollection of the content of the learning manipulation was strongly predictive of the strength of implicit attitudes (see also Gast & De Houwer, 2012; Van Dessel, De Houwer, & Gast, 2015a; Van Dessel, De Houwer, Roets, & Gast, 2016b).

Overall, these results seem easier to reconcile with an approach to human learning that allows for quick inferences from novel data than with an approach that characterizes learning as a process of incremental updating. At the same time, it should be noted that even though the associative approach to human learning generally relies on the idea of incremental updating, the same approach can also allow for high learning rates and, as such, accommodate episodes of single-shot associative learning, e.g., fear conditioning from a single CS–US pairing (Drew, Denny, & Hen, 2010). However, the implementations of this approach in the context of implicit evaluation (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) all posit a slow-learning process, which should not be able to produce change of the magnitude observed here within as few as four learning trials, corresponding to only 10 seconds of learning.

**Study 2**

Following a test of the unfolding of implicit attitude acquisition from an increasing number repeated evaluative pairings, Study 2 probed another key moderator of this learning effect. Specifically, we investigated whether implicit evaluations are updated in an informationally encapsulated manner, arising preferentially from stimulus pairings experienced in the environment, or in an informationally promiscuous manner, combining all relevant sources of evidence, including pairings and verbal information. Existing evidence on this is issue is contradictory, which provides further impetus for another test: Some studies have found that verbal descriptions of the nature of stimulus pairings significantly modulate evaluative learning from REP (Moran et
al., 2016; Moran, Bar-Anan, & Nosek, 2015; Peters & Gawronski, 2011; Zanon et al., 2014), whereas in others, similar manipulations produced no effect (Hu et al., 2016; Moran & Bar-Anan, 2013).

Recent work has demonstrated that diagnosticity, i.e., whether a piece of information is perceived as being reflective of an attitude object’s true underlying character, can have sizeable effects on implicit evaluations acquired on the basis of purely verbal manipulations (Cone & Ferguson, 2015). Inspired by this work, Study 2 investigated whether manipulations of diagnosticity can influence the acquisition of implicit attitudes from repeated evaluative pairings: In one condition, participants were informed that stimulus pairings were highly diagnostic (i.e., revealing the true nature of the target categories), whereas in another condition they were told that stimulus pairings were randomly generated and thus completely nondiagnostic. Given that the time at which verbal information about stimulus pairings is provided seems to be a crucial moderator of its effectiveness in influencing implicit attitudes (Hu et al., 2016; Peters & Gawronski, 2011; Zanon et al., 2014), we also varied whether participants received diagnosticity information before or after exposure to REP.

Method

Participants and design. 2,097 volunteers were recruited from the Project Implicit educational website (http://implicit.harvard.edu) to participate in the study.\textsuperscript{30} Consistent with standard scoring procedures (Greenwald et al., 2003), participants who did not complete the IAT and participants with a response latency of 300 ms or less on at least 10% of trials ($N = 246$) were excluded from all further analyses. Moreover, in line with standard exclusion rules used in the

\textsuperscript{30} The initial data collection included 833 participants. In response to a reviewer’s concern about the strength of the effect given the sample size, an additional round of data collection, involving 1,264 new participants, was conducted. Given that the new data showed a similar pattern for the crucial interaction, $\chi^2(9) = 4.25$, $p = .894$, the analyses reported in the paper are based on the data from the full sample.
field in general (Van Dessel, De Houwer, Gast, & Smith, 2015b; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a; Van Dessel, Gawronski, Smith, & De Houwer, 2017a) and in Study 1 in particular, participants who provided an inaccurate response on a manipulation check probing explicit recollection of the learning phase (N = 483) were also eliminated from consideration. This resulted in a final sample size of N = 1,365 (N = 896 female and N = 428 male participants, mean age = 38.32, SD = 15.39). This sample size provides .80 power to detect a small effect of r = .08.

All participants underwent an attitude induction procedure using repeated evaluative pairings (REP) of category members with positive and negative images. In a 2×2 between-subjects design, each participant was randomly assigned to a diagnosticity condition (stimulus pairings described as diagnostic vs. nondiagnostic) as well as to a timing condition (diagnosticity information received before vs. after exposure to stimulus pairings). Similar to Study 1, we failed to obtain evidence for differential attrition across conditions, χ²(3) = 7.27, p = .064.

Materials. Names from two fictitious social groups served as conditioned stimuli (CS). Names were pronounceable nonsense words constructed two conform to one of two phonological patterns: Laapian names ended on the syllable lap (e.g., Caalap) and Niffian names ended on the syllable nif (Gregg et al., 2006). The unconditioned stimuli (US) were the same as in Study 1. Both the CS and US have been used successfully for attitude induction in previous work (Gregg et al., 2006; Kurdi & Banaji, 2017).31

Procedure and measures. The study was administered entirely online and consisted of (a) a learning phase in which implicit attitudes were induced toward the target categories via exposure to REP, with a diagnosticity manipulation implemented either before or after the presen-

31 All stimuli are available for download from OSF (https://osf.io/serq4/).
tation of stimulus pairings, and (b) a test phase in which implicit attitudes and explicit attitudes toward the targets as well as memory for stimulus pairings were assessed.

**Learning phase.** The learning phase was identical to the learning phase of Study 1, with three exceptions. First, as mentioned above, Laapian and Niffian names, rather than individuals with long and square faces, served as CS. Because of a slight baseline preference in favor of Laapians (Kurdi & Banaji, 2017), Niffians were always paired with positive stimuli and Laapi-ans were always paired with negative stimuli to allow for learning to unfold without ceiling ef-fects. Second, the number of stimulus pairings presented to participants was not varied but was rather held constant at 20. Finally, and crucially, the diagnosticity of stimulus pairings was ma-nipulated using a simple verbal instruction. Participants in the *diagnostic condition* were in-formed that the pairings had been created to teach them something fundamental about the nature of the groups, whereas participants in the *nondiagnostic condition* were informed that the pair-ings had been randomly generated by the computer. Depending on their assignment to timing condition, this information was communicated to participants either as part of the initial instruc-tions and before the repeated evaluative pairings were displayed (*before condition*) or at the end of the learning phase, after the repeated evaluative pairings had already been presented (*after condition*).

**Test phase.** The test phase was identical to the test phase of Study 1, with the exception that the target categories were Laapians and Niffians rather than individuals with long faces and square faces. Participants completed (a) an IAT (Greenwald et al., 1998) measuring their implicit attitudes toward the targets (internal consistency $R = .77$), followed by (b) two feeling thermom-eter measures used to construct a relative measure of explicit preferences, and finally (c) a manip-

---

32 The verbatim text of the instructions is available for download from OSF ([https://osf.io/serq4/](https://osf.io/serq4/)).
ulation check probing participants’ recollection of the stimulus pairings presented in the learning phase.

Results

Diagnosticity and timing. The focal question of Study 2 was whether the diagnosticity of stimulus pairings, manipulated via minimal verbal information presented either before or after exposure to stimulus pairings, would influence the magnitude of implicit attitudes. The effects of the diagnosticity and timing conditions were probed by fitting a linear regression to the data, with IAT D score as the response variable, and timing condition (0 = after and 1 = before), diagnosticity condition (0 = nondiagnostic and 1 = diagnostic), as well as their interaction, entered as predictors (for a visual display of the results, see Figure 3).
Figure 3. Implicit attitude strength (measured using IAT D scores) as a function of learning conditions (Study 2). The left pane shows the information before condition, whereas the right pane shows the information after condition. Positive scores imply implicit attitudes in line with the learning manipulation. Error bars show 95-percent confidence intervals comparing each effect size to zero and differences significant at $p < .005$ are marked **.

We obtained a significant intercept, $b_0 = .26 [.21, .32]$, $t(1326) = 10.11, p < .001$, suggesting considerable deviation from neutrality in the reference category, i.e., among the group of participants who received information about the stimulus pairings being nondiagnostic after being exposed to them. However, the main effects of timing condition, $b_1 = .04 [-.04, .11], t(1326) = .93, p = .350$, and of diagnosticity condition, $b_2 = .02 [-.06, .09], t(1326) = 0.41, p = .682$, were both nonsignificant. Thus, the present data provide no evidence for an effect of (a) timing (before vs. after stimulus presentation) among participants in the diagnostic information condition or (b) diagnosticity information (stimulus pairings diagnostic vs. nondiagnostic) when it was communicated to participants after the stimulus pairings have already been presented.

Crucially, the Timing × Diagnosticity interaction emerged as significant, $b_3 = -.12 [-.23, -.01], t(1326) = -2.19, p = .029$, indicating a larger effect of diagnosticity in the information before condition than in the information after condition. Taken together, these results suggest that whereas the diagnosticity manipulation had considerable influence on the extent of evaluative learning when it was implemented before exposure to stimulus pairings, it did not produce any reliable effect after stimulus pairings had already been presented. The same impression can also be confirmed by fitting separate linear regressions to the data in the before vs. after conditions, with diagnosticity as the predictor. In the before condition, a significant effect of diagnosticity emerged, $t(669) = 2.70, p = .007$; however, in the after condition, no evidence of the same effect could be obtained, $t(657) = 0.41, p = .682$.

**Summary of supplementary analyses.** In Supplement 1 ([https://osf.io/serq4/](https://osf.io/serq4/)), we report additional analyses of the same data. To summarize, we demonstrate that (a) the evaluative
learning effects reported here are significantly modulated by participants’ conscious recollection of the learning manipulation (see Hofmann et al., 2010 and current Study 1); (b) the focal result reported above, yielding a significant Timing × Diagnosticity interaction in predicting implicit attitudes, also emerges in a Bayesian linear regression using uninformative priors; and (c) in a model including all participants without exclusions based on contingency awareness, only contingency awareness emerges as significant predictors of the strength of implicit attitudes, whereas the Timing × Diagnosticity interaction is reduced to nonsignificance. Overall, these supplementary analyses (i) demonstrate the robustness of the focal Timing × Diagnosticity interaction to analytic frameworks and (ii) suggest that the effects of the experimental manipulation on implicit evaluations may not be independent of the effects of contingency awareness.

Discussion

Study 2 probed whether (a) implicit attitude acquisition preferentially relies on stimulus pairings encountered in the environment, without taking into account verbal information accompanying such stimulus pairings, or (b) verbal information about the diagnosticity of such stimulus pairings influences the inferences made by participants and thus the strength of implicit attitudes toward the categories. We found that implicit attitudes were stronger in the condition in which stimulus pairings were described as diagnostic than in the condition in which they were described as randomly generated. In combination with previous work (Moran et al., 2015; 2016; Peters & Gawronski, 2011; Zanon et al., 2014), this finding provides evidence that implicit attitude acquisition from REP is not impervious to the influence of verbal instructions. Moreover, it demonstrates that in addition to validity (Moran et al., 2015; Peters & Gawronski, 2011) and type of relationship such as opposition (Zanon et al., 2014) or causation (Moran et al., 2015; 2016), information about diagnosticity can also modulate learning from valenced stimulus pairings.
However, the effects of the diagnosticity manipulation are subject to at least two qualifications. First, implicit attitudes significantly differed from neutrality even in the condition in which stimulus pairings were described as nondiagnostic before actual exposure. This result is in line with previous findings demonstrating that even purely verbal information presented to participants as fully nondiagnostic can influence implicit evaluations (Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a). Such findings may be construed as evidence that REP-based interventions shift implicit attitudes via a mix of associative and propositional processes (Gawronski & Bodenhausen, 2018). Alternatively, they may be parsimoniously accounted for within a propositional framework. Specifically, human communication ubiquitously relies on an assumption of relevance (Grice, 1975) and a single manipulation of diagnosticity may not be sufficient to fully override this assumption. Moreover, as pointed out by De Houwer (2018), the time constraints imposed by implicit measures may result in “quick and dirty” propositional reasoning in the course of which not all relevant pieces of information may be fully integrated with each other. Indeed, in a recent study using a process dissociation framework, Hütter and De Houwer (2017) provided evidence for automatic (memory-independent) effects of purely verbal manipulations on evaluative learning, suggesting that associative processes need not underlie the influence of nondiagnostic information on implicit evaluations.

Second, the manipulation of diagnosticity was ineffectual in the condition in which it was implemented after exposure to stimulus pairings (see also Hu et al., 2016; Peters & Gawronski, 2011; Zanon et al., 2014). This finding was not, a priori, predicted by propositional theories, although it is compatible with a Bayesian model of evaluative conditioning assuming a relatively strong prior expectation of stimulus pairings being diagnostic. Further implications of the latter
qualification on the effect of the diagnosticity manipulation, including potential computational models that may be able to accommodate them, are addressed in the General Discussion.

**Study 3**

The effect of the passage of time on the retention of information is among the earliest and most widely studied phenomena in the psychology of learning and memory (Bouton, 1993; Rubin & Wenzel, 1996). In the initial investigations probing the relative effectiveness of repeated evaluative pairings vs. evaluative statements in shifting implicit evaluations, attitude change was assessed immediately after the learning phase (Kurdi & Banaji, 2017). However, in the context of implicit attitude acquisition, the rate of decay in learning effects as a result of different interventions has far-reaching practical and theoretical implications.

Attempts at creating long-term change in implicit evaluations are rare and usually unsuccessful (for some notable exceptions see Mann & Ferguson, 2015; 2017). Failures to achieve durable change may, at least in part, be attributable to the fact that most studies to date have used long delays of days or at least hours between learning and test (Devine et al., 2012; Lai et al., 2016; Mann & Ferguson, 2017) during which (a) forgetting may have occurred and (b) participants may have spontaneously encountered sources of interference. In order to offer some insight into the decay of learning effects before the return of implicit attitudes to their baseline levels, the present studies used shorter delays than are customary in this literature, probing the effects of time within a single IAT session (Study 3) and imposing a delay of 15 minutes, rather than hours or days, between the learning and test phases of the experiment (Study 4).

Crucially, rates of decay in implicit attitudes induced via repeated evaluative pairings (REP) vs. evaluative statements (ES) also have the potential to illuminate the processes by which each intervention produces evaluative learning. As discussed above, associative theories predict
that REP should create more enduring change in implicit evaluations than ES because of a proposed fit between stimulus associations experienced in the environment and conceptual associations stored in long-term memory.

Conversely, existing propositional accounts of implicit evaluation do not make a strong prediction regarding the memory effects that should emerge. However, to the extent that they posit explicit and implicit evaluations to be subserved by the same memory systems, propositional accounts seem to be compatible with the idea that findings regarding episodic memory should be applicable to the present study. Specifically, episodic memory for verbally described events has been demonstrated to be initially superior to episodic memory for actually experienced events (Baggett, 1979; Baggett & Ehrenfeucht, 1983; Beentjes & van der Voort, 1991; Larsen & Plunkett, 1987; Toglia et al., 1992); however, this pattern is reversed if a delay is imposed between the event and the test of memory. If the same finding extends to the effects of evaluative learning, as reflected by implicit attitude measures, then the effects of ES should dissipate quickly, whereas learning in the REP condition should remain relatively stable.

As such, more durable learning effects of REP compared to ES are generally compatible with both associative and propositional theories. However, the predictions of both perspectives diverge when it comes to a combined ES + REP intervention. Specifically, if the superiority of REP over ES is due to the presence of stimulus pairings, as argued by associative theories, the ES + REP combined condition should be indistinguishable from the REP condition given that both involve exposure to stimulus pairings. By contrast, experienced events, such as the REP condition of the present studies, may result in superior long-term retention because they require participants to generate inferences about the event rather than merely encoding information already provided to them in propositional form (Baggett, 1979). Should this be the case, then the
combined ES + REP condition should exhibit the same patterns of decay as the ES condition given that participants in the combined condition, like in the ES condition, need not generate any inferences about the attitude objects. Study 3 provides an initial test of these predictions by conducting a reanalysis of the data reported by Kurdi and Banaji (2017), focusing on the effects of time on implicit attitude strength as a function of different learning interventions (REP, ES, and ES + REP combined).

Method

Participants and design. A final sample of 2,201 volunteers from the Project Implicit educational website (http://implicit.harvard.edu) participated in the study. This sample size provides .80 power to detect a small effect of $r = .06$. Exclusion criteria and demographic characteristics are described in Kurdi and Banaji (2017). In a between-subjects design, participants were randomly assigned to one of four learning conditions: control (involving exposure to only US–US pairings), repeated evaluative pairings (REP; exposure to pairings of category members with positive and negative images), evaluative statements (ES; exposure to verbal information about upcoming stimulus pairings without actual stimulus presentations), and combined (ES + REP; verbal information followed by exposure to stimulus pairings). Unlike in Studies 1 and 2, we observed differential attrition across learning conditions, $\chi^2(3) = 16.68, p = .001$. Specifically, differential attrition affected primarily the ES condition, which included $N = 498$ participants (as opposed to the expected $N$ of 550 under $H_0$).

Materials. Conditioned stimuli included Laapian and Niffian names (Study 3A; $N = 329$), individuals with long faces and square faces (Study 3B; $N = 528$), squares and rectangles (Study 3C; $N = 435$), faces of young and elderly individuals (Study 3D; $N = 429$), and photographs of American and foreign symbols (Study 3E; $N = 480$). The line drawings used in Studies
1 and 2 served as unconditioned stimuli, with the exception of Study 3D in which valenced photographs drawn from the Open Affective Standardized Image Set (OASIS; Kurdi et al., 2017) were used.33

**Procedure and measures.** The study was administered entirely online and consisted of *(a) a learning phase* in which implicit attitudes toward the target categories were induced using one of three learning manipulations, described in detail below, and *(b) a test phase* in which implicit and explicit attitudes toward the targets were assessed.34

**Learning phase.** Participants were randomly assigned to one of four conditions: a control condition and three learning conditions (REP, ES, ES + REP combined), with time on task kept constant across all four conditions. In the *control condition*, participants were exposed to US–US pairings to control for stimulus exposure. In the *REP condition*, attitudes were induced via exposure to pairings of members of one target category with positive images and members of the other target category with negative images. Crucially, in this condition, no verbal labels were used for the groups and no mention was made of stimulus valence. In the *ES condition*, participants were informed that they would see pairings of members of one category with positive images and members of the other category with negative images (De Houwer, 2006). In fact, no stimulus pairings were presented. Finally, in the *combined ES + REP condition*, verbal descriptions of stimulus pairings were followed by exposure to actual stimulus pairings.35 In order to avoid ceiling effects, the attitude induction procedure was designed to move participants away from the prevalent baseline. That is, participants were taught that *(a) Niffians are good and Laapians are

33 All stimuli are available for download from OSF (https://osf.io/serq4/).
34 Study 3 was conducted before Studies 1, 2, and 4 and thus did not contain a manipulation check item.
35 Because time on task was kept constant across conditions, participants in the control and combined conditions were exposed to 20 stimulus pairings, whereas participants in the REP condition were exposed to 37 stimulus pairings. However, in line with the results of Study 1, the pattern of results did not change when the combined condition involved the same number of stimulus pairings as the REP condition (Kurdi & Banaji, 2017).
bad (Study 3A), (b) square faces are good and long faces are bad (Study 3B), (c) rectangles are good and squares are bad (Study 3C), (d) elderly are good and young are bad (Study 3D), and (e) foreign is good and American is bad (Study 3E). Further procedural details are described in Kurdi and Banaji (2017).

**Test phase.** The test phase was identical to the test phases of Studies 1 and 2, with the exception that no manipulation check probing for contingency awareness was administered to participants. Participants completed (a) an IAT (Greenwald et al., 1998) measuring their implicit attitudes toward the target categories, followed by (b) two feeling thermometer measures used to construct a relative measure of explicit preferences.

**Analytic strategy.** Traditional analyses of these data relying on D scores (Greenwald et al., 2003) to measure IAT performance were reported by Kurdi and Banaji (2017). Both the original scoring algorithm producing a simple mean difference (Greenwald et al., 1998) and the improved scoring algorithm producing a standardized mean difference across critical blocks (Greenwald et al., 2003) treat each IAT trial as equivalent, without taking into account the temporal aspect of IAT data. The focal question of the present study was whether IAT D scores unfolded differently over time as a result of the different learning manipulations administered to participants. We investigated this question using a *moving time window method*, i.e., we recalculated the IAT D score for overlapping subsets of ten trials, resulting in *mini D scores*. The first mini D score was calculated from trials 1 through 10 of both the consistent and inconsistent IAT blocks, the second mini D score was calculated from trials 2 through 11, and so on. Because both

---

36 For exploratory purposes, the opposite associations were also reinforced for some participants. Data from those participants are not discussed here but are available for download from OSF (https://osf.io/serq4/).
critical blocks consisted of 40 trials, 31 mini D scores were calculated for each participant and then submitted to statistical analyses.\textsuperscript{37}

Results

**Descriptive statistics.** The temporal unfolding of mini D scores by learning condition is visually displayed in Figure 4. Adjusting for mean differences across studies, D scores in the control condition approached neutrality as the IAT session progressed, decreasing in absolute value from -0.37 (first mini D score involving trials 1 through 10) to -0.22 (last mini D score involving trials 31 through 40). D scores in the REP condition remained relatively stable, with an initial mini D score of -0.05 and a final mini D score of -0.01. The decline of mini D scores in the ES and combined conditions was considerably steeper, shifting from .20 to .09 in ES and from .23 to .12 in combined.

\textsuperscript{37} Because this method is new, we sought to ascertain its robustness to the size of the time window chosen. Therefore, we recalculated mini D scores based on 5, rather than ten, trials each and repeated all statistical analyses reported below. We obtained inferentially equivalent results, suggesting that the method is relatively robust to the number of trials chosen.
Figure 4. Implicit attitude strength (measured using IAT mini D scores) as a function of learning conditions and the passage of time within IAT sessions (Study 3). Positive scores imply implicit attitudes in line with the learning manipulation. The x-axis marks the passage of time within the IAT session, with each number showing the first trial taken into account for calculating each 10-trial mini D score. Light gray dots show the control condition, white squares show the repeated evaluative pairings (REP) condition, black diamonds show the evaluative statements (ES) condition, and dark gray triangles show the ES + REP combined condition.

**Model fitting.** In order to formally test whether the time course of the decay in evaluative learning effects significantly differed across learning conditions, we fit a linear mixed-effects model (Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012) to the data using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in the R statistical computing environment, with mini D scores as the response variable. Model fitting proceeded in a stepwise fashion, with a predictor added in each step and a likelihood ratio test conducted to probe whether the new predictor significantly improved model fit. After starting with a null model that contained
only random intercepts for participants to account for the fact that mini D scores were nested within participants, random intercepts for target categories were added to adjust for mean differences in implicit attitudes, $\chi^2(1) = 817.62, p < .001$. Subsequently, we entered a fixed effect for time as a continuous variable coded as the first trial included in the given mini D score, $\chi^2(1) = 36.46, p < .001$. In the next step, a fixed effect for learning condition (control vs. REP vs. ES vs. combined) was added, $\chi^2(3) = 253.75, p < .001$. Finally, as a test the focal question of the present study, we entered a Time × Learning Condition interaction, $\chi^2(3) = 759.58, p < .001$. Taken together, these model fitting steps suggest that, as expected, evaluative learning decays differentially depending on learning conditions even within a single IAT session. The nature of this interaction is discussed under Model interpretation below.

Table 2. Regression table showing the best-fitting mixed-effects model from Study 3. Random effects include random intercepts for participants and targets. Fixed effects include time (with the mini D score calculated from trials 1–10 coded as 0), learning condition (control, REP, ES, and ES + REP combined, with the control condition serving as the reference category), and a Time × Learning Condition interaction. SD = standard deviation, SE = standard error, df = degrees of freedom.

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants (intercept)</td>
<td>0.284</td>
<td>0.533</td>
</tr>
<tr>
<td>Target (intercept)</td>
<td>0.186</td>
<td>0.431</td>
</tr>
<tr>
<td>Residual</td>
<td>0.157</td>
<td>0.396</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Estimate</th>
<th>SE</th>
<th>df</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.335</td>
<td>0.194</td>
<td>5</td>
<td>-1.73</td>
<td>0.15</td>
</tr>
<tr>
<td>Time</td>
<td>0.005</td>
<td>0.000</td>
<td>66027</td>
<td>16.47</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Learning condition (REP)</td>
<td>0.341</td>
<td>0.033</td>
<td>2414</td>
<td>10.32</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Learning condition (ES)</td>
<td>0.591</td>
<td>0.033</td>
<td>2414</td>
<td>17.78</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Learning condition (Combined)</td>
<td>0.621</td>
<td>0.032</td>
<td>2414</td>
<td>19.43</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Time × Learning Condition (REP)</td>
<td>-0.004</td>
<td>0.000</td>
<td>66027</td>
<td>-9.28</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Time × Learning condition (ES)</td>
<td>-0.011</td>
<td>0.000</td>
<td>66027</td>
<td>-23.35</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Time × Learning Condition (Combined)</td>
<td>-0.011</td>
<td>0.000</td>
<td>66027</td>
<td>-22.77</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>
**Model interpretation.** The regression coefficients from the best-fitting mixed-effects model, including random intercepts for participants and targets, as well as a Time × Learning Condition interaction, are presented in Table 2. Dummy coding with the control condition as the reference category was used to represent the learning condition variable and time was entered as a continuous variable scaled such that 0 represents the first mini D score (and, correspondingly, 30 represents the last mini D score).

We observed a significant and positive main effect of time, suggesting that in the control condition, mini D scores approached neutrality as time progressed. In addition, we observed main effects for each learning condition compared to control, which is in line with previous work but irrelevant from the perspective of the present study. Crucially, we obtained significant Time × Learning Condition interactions, suggesting that the effects of time within the IAT session changed depending on the manner in which implicit attitudes had been acquired. Whereas in the REP condition implicit attitudes were predicted to stay quite stable throughout the IAT session (predicted change per trial: $d_{pred} = 0.001$), in the ES ($d_{pred} = -0.006$) and combined ($d_{pred} = -0.006$) conditions, implicit attitudes were predicted to approach neutrality at a considerably higher pace. According to the regression model, D scores in the REP condition should change (away from neutrality) by 0.026 D score units throughout the entire IAT session. By contrast, the overall predicted decrease toward neutrality was 0.178 units in the ES condition and 0.157 units in the combined condition.

**Summary of supplementary analyses.** A Bayesian reanalysis of the same data, reported in Supplement 1 (https://osf.io/serq4/), confirms the main inference reported above: Implicit attitudes measured in the in the ES and combined conditions approached neutrality at a considerably faster rate than implicit attitudes measured in the REP condition.
Discussion

Study 3 provides initial evidence that, in line with predictions derived from associative theories of implicit evaluation and from studies on episodic memory for personally experienced vs. verbally described events, the passage of time differentially affects implicit attitudes acquired from REP vs. ES. Prior work has robustly demonstrated that when implicit evaluations are assessed immediately following the learning intervention, the learning effects produced by ES are at least on par with, and often superior to, the learning effects produced by REP (Kurdi & Banaji, 2017). However, when the passage of time within the IAT session is taken into account, the interpretation of the same data changes substantially: Whereas attitudes created via ES were found to decay considerably even within a single IAT session, attitudes created via REP remained stable, suggesting that any differences in favor of the ES condition observed in previous work were driven by early trials of the IAT and dissipated quickly over time.

Furthermore, the combined condition exhibited the same rapid decay as the ES condition, indicating that exposure to stimulus pairings (as opposed to purely verbal information) may, in and of itself, not produce stronger evaluative learning unless participants are required to make inferences about those stimulus pairings (Toglia et al., 1992). This result, unlike the superior long-term effects of REP compared to ES, is difficult to reconcile with associative theories of implicit evaluation. Specifically, differences in the behavior of the ES + REP vs. REP condition are not easily explained by an account suggesting that the primary determinant of the durability of evaluative learning is whether the relevant manipulation involves exposure to actual stimulus pairings. Rather, the REP condition may create more durable representations due to participant-generated propositions being more memorable than experimenter-provided ones, possibly be-
cause the additional cognitive effort that they require may result in deeper processing of evaluative information (Craik & Lockhart, 1972).

**Study 4**

Study 3 provided initial evidence for the differential decay of evaluative learning effects as a function of the learning manipulation used: When implicit attitude strength was measured dynamically within a single IAT session, learning from stimulus pairings remained stable, whereas learning produced as a result of a purely verbal manipulation or a combination of verbal instructions and stimulus pairings exhibited precipitous decay. Taken together, these results suggest that the stability in implicit attitudes in the REP condition is most likely due to the memory benefit created by participant-generated inferences (Baggett, 1979; Craik & Lockhart, 1972; Toglia et al., 1992). In Study 4, we created another test of the same idea by imposing a 15-minute delay between the learning and test phases of the experiment. If the findings obtained in Study 3 extend to this relatively longer time scale, the advantage of the ES and ES + REP conditions over the REP condition, as observed when implicit attitudes are measured immediately after learning (Kurdi & Banaji, 2017), should be reduced or possibly even eliminated.

**Method**

**Participants and design.** A sample of 506 undergraduates from the study pool of a private university in the Northeastern United States and a sample of 189 participants from the Digital Lab for the Social Sciences (DLABSS; http://dlabss.harvard.edu/) were recruited for the study, the former in exchange for partial course credit and the latter as volunteers. In line with standard scoring procedures (Greenwald et al., 2003), participants who did not complete the IAT ($N = 22$) as well as participants with a response latency of 300 ms or less on at least 10% of trials ($N = 9$) were excluded from all further analyses. Moreover, consistent with Studies 1 and 2 and
standard exclusion rules used in the field in general (Van Dessel, De Houwer, Gast, & Smith, 2015b; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016a; Van Dessel, Gawronski, Smith, & De Houwer, 2017a), participants who provided an inaccurate response on a manipulation check probing explicit recollection of the learning phase \(N = 56\) were also eliminated from consideration. This resulted in a combined final sample size of \(N = 606\). Because the effects of learning manipulations did not differ across the undergraduate and DLABSS participants (see below), analyses collapse across the two samples. The combined sample size provides .80 power to detect a small effect of \(r = .11\).

Like in Study 3, participants were randomly assigned to one of four learning conditions in a between-subjects design: control (involving exposure to only US–US pairings), repeated evaluative pairings (REP; exposure to pairings of category members with positive and negative images), evaluative statements (ES; exposure to verbal information about upcoming stimulus pairings without actual stimulus presentations), and combined (ES + REP; verbal information followed by exposure to stimulus pairings). However, unlike in Study 3, a 15-minute delay was imposed between the learning phase and the test phase of the experiment. Unlike in Studies 1 and 2, but similar to Study 3, we observed differential attrition across learning conditions, \(\chi^2(3) = 45.34, p < .001\). However, unlike in Study 3, differential attrition affected primarily the control condition, which included \(N = 87\) participants (as opposed to the expected \(N = 152\) under \(H_0\)).

Materials. Conditioned stimuli included Laapian and Niffian names (undergraduate participants) and individuals with long faces and square faces (DLABSS participants). Because of different baseline attitudes across the two stimulus sets and samples, mixed-effects models (see
below) included random intercepts for participant pools. The line drawings used in Studies 1–3 and in Kurdi and Banaji (2017) served as unconditioned stimuli for all participants.

**Procedure and measures.** The study was administered entirely online and consisted of (a) a *learning phase* in which implicit attitudes toward the target categories were induced using one of three learning manipulations mentioned above (REP, ES, or ES + REP combined), with some additional participants measured at baseline in the control condition, and (b) a *test phase* in which implicit and explicit attitudes toward the targets as well as memory for stimulus pairings were assessed. Similar to previous studies, internal consistency of the IAT was acceptable, $R = .76$. Both the learning phase and the test phase were identical to Study 3, with two exceptions: (1) following the explicit attitude measures, the test phase included a measure of memory for stimulus pairings and, crucially, (2) a delay of 15 minutes was imposed between the learning and test phases of the experiment. During this delay, participants were allowed to spend their time in any way they wished. Once the delay was over, participants received an email with a link to the second part of the study.

**Results**

**Learning conditions.** Study 4 examined the relative effectiveness of different learning manipulations in shifting implicit attitudes, as assessed following a 15-minute delay between the learning and test phases of the experiment. Mean implicit attitudes by learning condition (control, REP, ES, and combined) are displayed in Figure 5. Descriptively, Figure 5 suggests that implicit attitudes in all three learning conditions remained different from control in spite of the delay. Moreover, unlike in previous work involving immediate testing of the effects of evaluative

---

38 All stimuli are available for download from OSF ([https://osf.io/serq4/](https://osf.io/serq4/)).
learning, implicit attitudes in the REP and combined conditions seem descriptively higher than in the ES condition; however, this impression may not be confirmed using inferential testing.

![Implicit attitudes by learning condition](image)

**Figure 5.** Implicit attitude strength (measured using IAT D scores) as a function of learning conditions (Study 4), including control, repeated evaluative pairings (REP), evaluative statements (ES), and ES + REP combined. Positive scores imply implicit attitudes in line with the learning manipulation. Error bars show 95-percent confidence intervals comparing each effect size to zero and differences significant at $p < .001$ are marked ***.

To formally test whether implicit attitudes significantly differed across learning conditions, we fit a linear mixed-effects model (Baayen et al., 2008; Judd et al., 2012) to the data using the lme4 package (Bates et al., 2015) in the R statistical computing environment, with IAT D scores as the response variable. As in Study 3, model fitting proceeded in a stepwise fashion. After starting with a null model that contained only random intercepts for participant pools to account for different baseline attitudes, a fixed effect for learning conditions was entered and found
to produce significant improvement in model fit, $\chi^2(3) = 54.54, p < .001$. Finally, we added random slopes for learning conditions across participant pools, accounting for the possibility that the effect of learning manipulations may have differed across undergraduate and DLABSS participants; however, we obtained no evidence for such differences, $\chi^2(9) = 3.42, p = .945$. Therefore, below we interpret the best-fitting model, which included random intercepts for participant pools and a fixed effect for learning conditions.

As in Study 3, the control condition served as the reference category. Therefore, the intercept represents mean implicit attitude strength in the control condition and slope parameters represent differences between the control condition and each learning condition. Participants in the control condition exhibited a mean implicit attitude level of $b_0 = .07 [-.18, .32], t(2.90) = 0.60, p = .589$, suggesting a neutral baseline. Compared to the control condition, each learning condition produced significant learning, $b_1 = .40 [.28, .52], t(601.00) = 6.46, p < .001$ in the REP condition, $b_2 = .34 [.22, .47], t(600.60) = 5.34, p < .001$ in the ES condition, and $b_3 = .44 [.32, .56], t(600.70) = 7.25, p < .001$ in the combined condition. This suggests that each of the three learning manipulations was sufficiently powerful to persist over a 15-minute delay between the learning and test phases of the experiment.

Given that potential differences across learning conditions are also of interest, we conducted indirect testing using bootstrap samples (see Kurdi & Banaji, 2017) to probe for such differences. In order to control for the familywise error rate, a Bonferroni-corrected alpha level of $\alpha = .05/6 = .008$ was used. We obtained no significant differences between learning conditions, including between REP and ES $b_{REP-ES} = -.11, 99.2$-percent CI: $[-.26, .05]$ $^{39}, p > .008$.

---

$^{39}$ Given that a Bonferroni correction was used, we report the confidence interval corresponding to the Bonferroni-corrected alpha level of .008 rather than the more usual 95-percent confidence interval.
REP and combined $b_{REP\text{-}COMB} = .02$, 99.2-percent CI: [-.13, .17], $p > .008$, or between ES and combined $b_{ES\text{-}COMB} = .12$, 99.2-percent CI: [-.02, .26], $p > .008$.\(^4\) Thus, unlike in previous experiments in which implicit attitudes were measured immediately, neither the ES nor the ES + REP condition outperformed REP in isolation, indicating that evaluative learning involving verbal manipulations decays faster than learning from stimulus pairings alone.

**Summary of supplementary analyses.** In Supplement 1 ([https://osf.io/serq4/](https://osf.io/serq4/)), we report additional analyses of the same data. To summarize, we demonstrate that (a) the evaluative learning effects reported here are significantly modulated by participants’ conscious recollection of the learning manipulation (see Hofmann et al., 2010 and current Studies 1 and 2); (b) the focal result reported above, yielding significant learning effects in all learning conditions compared to control but no differences across learning conditions, also emerges in a Bayesian linear regression using uninformative priors; and (c) in a model including all participants without exclusions based on contingency awareness, both contingency awareness and the learning condition variable emerge as significant predictors of implicit attitude strength. Overall, these supplementary analyses demonstrate the robustness of the focal result of overall learning but no difference across learning conditions to (i) participant exclusions and (ii) analytic frameworks.

**Discussion**

Study 4 constituted a second test of the idea that implicit evaluations created using different learning manipulations may decay differentially. In this study, unlike in previous work comparing the relative effectiveness of REP and ES in producing attitude change (Kurdi & Banaji, 2017) and most studies probing the effects of experimentally induced evaluative learning, implicit attitudes were assessed following a 15-minute delay rather than immediately. Three find-

\(^4\) The differences between the control condition and each learning condition survived Bonferroni correction.
ings from this study are noteworthy. First, implicit attitudes were significantly different from baseline in all three learning conditions, suggesting that the evaluative learning reported in previous work by Kurdi and Banaji (2017) was sufficiently powerful to persist over a 15-minute delay. Second, whereas immediate measurement of attitudes produced a robust advantage of the learning conditions involving verbal manipulations (ES and ES + REP), the 15-minute delay between learning and test used in the present study eliminated this advantage such that implicit attitudes across all three conditions were equivalent. This result demonstrates that, in line with the findings from Study 3, as well studies on episodic memory for personally experienced vs. verbally described events (Baggett, 1979; Baggett & Ehrenfeucht, 1983; Beentjes & van der Voort, 1991; Larsen & Plunkett, 1987; Toglia et al., 1992), the memory advantage of verbal instructions over direct experience with stimulus pairings can be subject to fast decay. Third, again in line with Study 3, the ES and ES + REP conditions behaved similarly. Given that both the REP and the ES + REP conditions involve actual experience with stimulus pairings, this result indicates that REP may create more stable implicit attitudes due to the self-generated nature of the propositional inference that is required for evaluative learning to be successful in this condition (Craik & Lockhart, 1972; Toglia et al., 1992) rather than because stimulus pairings are inherently more effective in shifting implicit evaluations than verbal instructions.

General Discussion

In recent work, Kurdi and Banaji (2017) provided evidence that evaluative statements (ES) merely informing participants of upcoming stimulus pairings without actual exposure are more effective in shifting implicit attitudes than repeated evaluative pairings (REP) of category members with valenced images. Moreover, a combined condition of ES + REP did not result in more pronounced attitude change than ES in isolation, suggesting that exposure to stimulus pair-
ings did not produce added value. Taken together, these findings indicated that, surprisingly, REP and ES may shift implicit attitudes via the same learning process, although ES seems to be more effective on an immediate test of implicit attitudes. To further explore the nature of these effects and the learning processes giving rise to them, the present work implemented tests of (a) the unfolding of learning as a result of increasing the number of stimulus exposures (Study 1), (b) the interaction between stimulus pairings and verbal information in producing implicit attitude change (Study 2), and (c) the temporal stability of implicit attitude change (Studies 3–4). In addition to their theoretical relevance, these studies also have implications for the more applied issue of producing enduring change in preexisting implicit attitudes.

In Study 1, the number of stimulus pairings presented in the REP condition did not modulate the strength of implicit attitudes created. Four stimulus pairings, presented over a duration of only 10 seconds, were sufficient to produce attitude change and were found to be just as effective as 24 stimulus pairings. These results suggest that no protracted experience with stimulus pairings is necessary for implicit attitude change to occur. Moreover, they seem difficult to reconcile with the idea of a slow and incremental learning process.

Study 2 provides evidence that, much like other kinds of verbal information specifying the nature of the relationship between the stimuli presented (Moran et al., 2015; 2016; Peters & Gawronski, 2011; Zanon et al., 2014), manipulations of diagnosticity (describing stimulus pairings as randomly generated vs. revealing the underlying true nature of the target categories) can influence the magnitude of learning from stimulus pairings experienced in the environment. In combination with previous work, this finding shows that such learning is informationally promiscuous. Moreover, it demonstrates, for the first time, that manipulations of diagnosticity can
be used to modulate the effects of evaluative learning not only from verbal statements (Cone & Ferguson, 2015) but also from stimulus pairings.

Third, across Studies 3–4, we have shown that implicit attitude change created by exposure to stimulus pairings is more durable than implicit attitude change created by verbal instructions merely signaling upcoming stimulus pairings. Moreover, the combined ES + REP condition exhibited the same precipitous decay in evaluative learning effects as the condition using ES alone. As such, this finding suggests that whatever feature of the REP manipulation leads to superior retention over time, this feature is eliminated by the presence of verbal instructions specifying the nature of stimulus pairings to which the participant will be exposed.

**Theoretical implications**

We believe that the present results are primarily of interest because they have the ability to constrain any future theory of implicit attitude acquisition, be they associative, propositional, an associative/propositional hybrid, or of any other flavor. Specifically, based on the findings of the recent work by Kurdi and Banaji (2017), combined with the current studies, any theory of implicit evaluation must be able to explain why and how (a) a combined intervention of ES + REP produces equivalent learning to ES alone; (b) stimulus pairings presented beyond the initial four do not produce any further learning in REP; (c) learning in REP is integrated with verbal information on the nature of stimulus pairings; and (d) REP in isolation, but not a combination of ES + REP, produces more durable evaluative learning than ES alone. Although these results seem generally more compatible with a quick, inferential, and informationally promiscuous learning process that stores propositions in long-term memory (Csibra & Gergely, 2009; Gopnik, 1996; Tenenbaum et al., 2011) than with a slow, stimulus-driven, and informationally encapsulated that stores conceptual associations in long-term memory (Rescorla & Wagner, 1972; Ru-
(Melhart et al., 1986; Sutton & Barto, 1998), we remain open to the possibility that a fundamentally different kind of account may provide a better explanation of the same data in the future.

When it comes to currently available associative theories of implicit evaluation (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004), these theories seem to be difficult to reconcile with numerous aspects of the present results. First, these theories posit that implicit evaluations should shift as a result of slow and incremental updating processes. However, in Study 1, learning emerged and stabilized quickly. Second, these theories suggest that implicit evaluations should rely exclusively, or at least preferentially, on associative rather than verbal interventions. And yet, in Study 2, implicit evaluations emerged from a combination of both kinds of information. Finally, the fact that stimulus pairings gave rise to more enduring attitude change than verbal statements (Studies 3–4) seems to be well accounted for by associative theories; however, such theories do not appear to be well equipped to explain why the benefit of stimulus pairings is eliminated when they are preceded by verbal information. As such, the present results also seem to call into question models of evaluative conditioning as a purely stimulus-driven and bottom-up process that bypasses high-level reasoning (Baeyens et al., 1992; 2009; Gawronski et al., 2013; Gawronski & Bodenhausen, 2006; Levey & Martin, 1975; Martin & Levey, 1978; Rydell & McConnell, 2006; Strack & Deutsch, 2004). At the same time, new associative theories of implicit evaluation may be able to account for the findings of Study 1 to the extent that, unlike currently available associative theories, they allow for single-shot associative learning. However, it is more difficult to see how purely associative theories would be able to deal with findings of interaction between verbal information and stimulus pairings in giving rise to implicit attitudes (Kurdi & Banaji, 2017 and current Studies 2–4).
Overall, current propositional and symbolic theories of implicit evaluation (De Houwer, 2009; 2014; 2018; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009) seem to be better suited to account for the present results. Specifically, although these theories do not require that the number of stimulus pairings never modulate the strength of learning from REP, they allow for the possibility of quick propositional inferences following limited exposure (Study 1). Second, these theories posit that propositional learning should productively combine all available evidence, whether it be presented in pictorial or verbal format, in shifting implicit attitudes; the results of Study 2 are largely in line with this idea. Finally, even though propositional theories did not a priori predict the results of Studies 3–4, they suggest that REP should shift implicit attitudes via propositional processes and, as such, can easily incorporate past findings on the memory advantage of self-generated propositions (Craik & Lockhart, 1972; Toglia et al., 1992). At the same time, as currently formulated, propositional theories do not seem to be well-equipped to explain why diagnosticity information should affect implicit attitude acquisition from REP only if presented before, but not if presented after, exposure to stimulus pairings (Study 2). Below we provide a potential explanation of this finding within the broad framework of inferential approaches to human learning.

Although the current findings are difficult to explain under associative theories of implicit evaluation, they may be compatible with more recent versions of a hybrid model of implicit evaluation, specifically the Associative–Propositional Evaluation model (APE; Gawronski & Bodenhausen, 2006; 2011; 2018). Indeed, since its inception, the APE model has recognized that implicit evaluations may be updated as a result of associative processes, propositional processes, or a combination of both. Moreover, after initially introducing REP as “[…] [t]he prototypical case for implicit attitude changes resulting from changes in associative structure” (Gawronski &
Bodenhausen, 2006), more recent versions of the APE model allow for the possibility that REP-based learning may be “[…] driven by associative learning under some conditions, but by propositional learning under other conditions” (Gawronski & Bodenhausen, 2011). Gawronski and Bodenhausen (2018) go even further in theorizing that “[…] repeated pairings of a CS and a US can influence mental representations via associative learning, propositional learning, or both” (p. 2). However, given that the propositional perspective is (a) able to accommodate the bulk of present findings using one, rather than two, learning processes and (b) more restricted in its predictions than the APE model, we believe that propositional models are preferable as the most parsimonious and well-specified account of the present results.

Furthermore, it should be noted that some may wish to characterize the theories that we refer to as associative in the present work (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) as hybrid theories given that these theories incorporate certain limited interactions between the posited explicit and implicit systems. In fact, Smith and DeCoster (2000) allow for one kind of explicit–implicit interaction in explaining that repeated use of a step-by-step rule-based process may give rise to associative representations (see also Strack & Deutsch, 2004). This stipulation does not apply to the present work given that none of the present studies involved the kind of protracted practice described by Smith and DeCoster. Strack and Deutsch (2004) further allow for the possibility that elements of propositions represented in the explicit system may be retrieved from the implicit system. This, again, seems irrelevant to the present experiments.

Finally, at the end of their chapter, McConnell and Rydell (2014) speculate that the explicit and implicit systems may interact with each other (a) by taking the same information as input, (b) sequentially, with one system taking the output from the other system as its input, or
by relying on past information processing involving both systems. We see this speculation as fundamentally inconsistent with the rest of McConnell and Rydell’s systems of evaluation model, which describes the explicit and implicit systems of evaluation as “dissociable systems of knowledge” (p. 204), with “[…] each system becom[ing] more fully engaged with information to which it is most sensitive and neglect[ing] (at least in part) information to which it is less sensitive” (p. 213). However, to the extent that McConnell and Rydell wish to posit the same kind of intricate pattern of explicit–implicit interactions as the APE model described above, then it is our view that their theory will face similar challenges in terms of its falsifiability.

Methodological implications

In addition to these substantive findings, we also introduced a new measure of performance on the IAT (Greenwald et al., 1998). This new measure is based on the improved scoring algorithm proposed by Greenwald, Nosek, and Banaji (2003) but, unlike the improved scoring algorithm, takes into account the temporal unfolding of response latencies within the IAT session. This new scoring method can be used to obtain mini D scores that index changes in participants’ IAT performance as the critical blocks progress. Given that mini D scores rely on data obtained within a single IAT session, they can be used to probe the temporal stability of implicit attitudes without the need for recruiting participants for studies that extend across multiple measurement occasions. Importantly, the results that we obtained using mini D scores were (a) robust to the size of the moving average window chosen and (b) consistent with the findings of an experiment that relied on a multi-session design (Study 4), thus providing evidence for the face validity of this method. We hope that future work will (a) subject the validity of mini D scores to further empirical scrutiny and (b) more generally probe whether the results reported in the present paper generalize to measures of implicit evaluation other than the IAT.
A call for computational modeling

Further, the present project suggests two additional areas for exploration that we believe may be crucial to understanding how implicit attitudes change. First, current associative (McConnell & Rydell, 2014; Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004), propositional/symbolic (De Houwer, 2009; 2014; 2018; De Houwer & Hughes, 2016), and hybrid (Gawronski & Bodenhausen, 2006; 2011; 2018) accounts of implicit evaluation are best understood as general conceptual frameworks rather than as theories sufficiently specific to make quantitative predictions about the computations that the human mind should perform in response to different kinds of evaluative information. Given that current theories of implicit evaluation are formulated in a purely verbal manner, without making reference to potential computational instantiations of the learning processes that they posit, they are difficult to falsify.

Even though all computational models are wrong (Box, 1976) and, quite possibly, even stupid (Smaldino, 2017), they provide the tools necessary for addressing questions left open by the current experiments as well as by this body of work as a whole. Such open questions include (a) how exactly verbal information and actually experienced stimulus pairings interact with each other in producing implicit attitude change (Studies 1 and 2), (b) why more confirming instances of group A paired with positive stimuli and group B paired with negative stimuli sometimes do not lead to more belief revision (Study 1), and (c) why diagnosticity information presented before but not after exposure to stimulus pairings has an effect on evaluative learning (Study 2).

A possible answer to these open questions may be found within the framework of hierarchical Bayesian models of cognition (Tenenbaum et al., 2011), which provide a computational account of human reasoning under conditions of uncertainty. Bayesian computational modeling
encompasses both \((a)\) the selection of a generative model that can explain data encountered in the world and \((b)\) the estimation of the parameters of that model, both via updating prior beliefs in the face of new information. In this particular case, Bayesian modeling may facilitate understanding how different kinds of information, such as verbal information about stimulus pairings and information provided by the stimulus pairings themselves, are productively combined with each other in order to arrive at the appropriate understanding of the causal processes through which the stimulus pairings were generated.

More specifically, the verbal instructions provided prior to exposure to stimulus pairings can be understood as influencing participants’ prior beliefs about the upcoming task by constraining the hypothesis space that they should reasonably entertain. For instance, in the REP condition of the current studies, participants were informed that they would see two types of targets (CS) and two types of drawings (US) and were instructed to learn the general relationship between a certain type of target and a certain type of drawing. Given this prior information, once participants are able to infer that the two types of images are instances of the general categories “good” and “bad,” there are only two potential hypotheses remaining: \((H_1)\) group A always paired with positive images and group B always paired with negative images or \((H_2)\) group A always paired with negative images and group B always paired with positive images. At this point, a single US–CS pairing already provides conclusive evidence to help decide between \(H_1\) and \(H_2\). This may be the reason why number of stimulus pairings did not produce an effect in Study 1: Because the hypothesis space was heavily constrained by initial instructions, four stimulus pairings were sufficient for participants to make the valence inference and form posterior beliefs decisively in favor of either \(H_1\) or \(H_2\).
Such a Bayesian perspective may also be able to account for why the diagnosticity manipulation affected evaluative learning when verbal information was provided to participants before, rather than after, exposure to stimulus pairings. If participants receive prior information that they are about to learn something fundamental about the nature of the target categories, this implies a highly structured hypothesis space in which group A and B are each paired with a unique set of USs. By contrast, if participants believe that the stimulus pairings are randomly generated, that would imply that any CS may be paired with any US. Given that the hypothesis space is considerably more constrained in the former case, the same amount of data should create more learning than in the latter case. However, when participants do not receive prior instructions about diagnosticity, they may reasonably assume that the stimulus pairings provided by the experimenter are relevant and informative (Grice, 1975), i.e., the hypothesis space that they consider may be quite limited. For instance, participants may be unlikely to spontaneously entertain the possibility that, in a learning experiment, stimuli will be paired with each other in a fully randomized manner. If this is the case and, therefore, participants’ learning from stimulus pairings is effective, post-hoc information about diagnosticity may not eliminate the robust learning effects already in place.

Regardless of whether these specific conjectures are accurate or not, future work involving Bayesian modeling will be able to provide a computational account of how information presented in different formats (e.g., verbally vs. pictorially) is integrated in updating implicit evaluations. As such, Bayesian modeling may be seen as a computationally tractable instantiation of propositional and symbolic accounts of implicit evaluation (De Houwer, 2009; 2014; 2018; De Houwer & Hughes, 2016), which posit that implicit attitudes are updated in an inferentially promiscuous, rather than in a purely bottom-up and stimulus-driven, manner. Models of causal rea-
soning implemented in a Bayesian framework (Gopnik et al., 2004; Lu, Rojas, Beckers, & Yuille, 2015) may serve as a useful starting point for implementing computational models of the propositional processes resulting in the updating of (implicit) evaluations.

**Implications for long-term change**

Although the present studies were conducted within a relatively constrained time span between learning and test, our findings have two kinds of implications for long-term change in implicit attitudes: (a) general implications for how the issue of long-term change should be conceptualized and (b) specific implications for interventions that might be successful in shifting implicit attitudes in a durable way.

Implicit attitudes are often characterized as recalcitrant and resistant to change (Bargh, 1999; Gregg et al., 2006; Rydell & McConnell, 2006). Accordingly, attempts at creating long-term change in implicit attitudes toward social groups are often unsuccessful (Forscher et al., 2017; Lai et al., 2016). However, at the same time, a recent analysis of data from more than 4 million participants from the Project Implicit educational website (http://implicit.harvard.edu/) has demonstrated sizable long-term shifts toward neutrality in implicit attitudes toward race, skin tone, and sexual orientation (Charlesworth & Banaji, 2019). If the effects of targeted interventions created in the lab dissipate within a few hours, how is it possible for implicit attitudes to change in the long run? The present project may provide some pointers toward a potential explanation.

The results of Study 4 suggest that the effects of interventions like repeated evaluative pairings and evaluative statements are sufficiently robust to survive a 15-minute delay between the learning and test phases of the experiment. As such, this project demonstrates that changes in implicit attitudes created via such interventions are not fully ephemeral and do not depend on the
information being actively rehearsed between learning and testing. Therefore, it seems that studies probing the effects of relatively minimal manipulations after relatively long delays (Forscher et al., 2017; Lai et al., 2016) may have been unsuccessful in creating long-term change not because the interventions themselves are ineffective. Rather, participants encounter members of social groups in evaluatively consequential contexts, as well as evaluatively relevant verbal information about them, thousands of times every day. Such information may have acted as spontaneous sources of interference and may have eliminated the effects of one-shot interventions.

Thus, if progress toward understanding long-term shifts in implicit evaluations is of interest, future studies may want to address this issue in one of two ways. First, they might examine the potential for long-term change through administering interventions that seem effective in the medium term (such as within a timeframe of hours) multiple times a day to counteract forgetting and interference. Second, a recent theoretical piece by Payne, Vuletich, and Lundberg (2017) suggests that, in addition to exploring implicit attitude change at the individual level, investigating change in the aggregate may also be a fruitful avenue of research. As such, one might use archival data to correlate changes or differences in potential antecedents to implicit attitudes across times or geographic areas with aggregate changes or differences across the same times or areas.

Finally, the present project has implications for the kinds of interventions whose effects should be relatively durable in creating implicit attitude change. Specifically, we found that repeated evaluative pairings resulted in more robust shifts in implicit attitudes over time than evaluative statements. However, this was only true when participants were required to make extensive inferences about the stimulus pairings themselves, without those inferences being provided to them in verbal form. In line with research exploring episodic memory for personally experi-
enced vs. merely verbally described events (Baggett, 1979; Baggett & Ehrenfeucht, 1983; Beentjes & van der Voort, 1991; Larsen & Plunkett, 1987; Toglia et al., 1992), this finding suggests that evaluative learning may be made more memorable by engaging participants in deeper processing of the stimuli. In addition, exposure to category members paired with positive images may create less reactance than verbal messages with the obvious intent to push participants toward change (De Houwer & Hughes, 2016).

However, long-term effectiveness of primarily nonverbal manipulations presupposes that participants have the ability to appropriately categorize stimuli and to detect spatiotemporal contingencies between them. This assumption may not always be warranted, as evidenced by high degrees of variability in initial learning effects from repeated evaluative pairings (Kurdi & Banaji, 2017). Thus, future attempts at shifting implicit attitudes may be particularly effective if interventions are designed to be sufficiently challenging to require participants’ attention and active engagement but not as challenging as to make it impossible for participants to make the intended inferences.
Paper 3 • Model-Free and Model-Based Learning Processes in the Updating of Explicit and Implicit Evaluations
Abstract

Evaluating stimuli along a good–bad dimension is a fundamental computation performed by the human mind. In recent decades, research has documented both dissociations and associations between explicit (self-reported) and implicit (indirectly measured) forms of evaluations. However, it is unclear whether such dissociations arise from relatively more superficial differences in measurement techniques or from deeper differences in the processes by which explicit and implicit evaluations are acquired and represented. The current project (total \( N = 2,354 \)) relies on the computationally well-specified distinction between model-based and model-free reinforcement learning to investigate the unique and shared aspects of explicit and implicit evaluations. Study 1 used a revaluation procedure to reveal that whereas explicit evaluations of novel targets are updated via both model-free and model-based processes, implicit evaluations depend on the former but are impervious to the latter. Studies 2–3 demonstrated the robustness of this effect to (a) the number of stimulus exposures in the revaluation phase and (b) the deterministic vs. probabilistic nature of initial reinforcement. These findings provide a novel framework, going beyond traditional dual-process and single-process accounts, to highlight the context-sensitivity and long-term recalcitrance of implicit evaluations as well as variations in their relationship with their explicit counterparts. These results also suggest novel avenues for designing theoretically guided interventions to produce change in implicit evaluations.
Introduction

The human mind continuously assigns subjective value to information encountered in the environment (Allport, 1935). Such evaluations of humans, abstract concepts, and physical objects are crucial to structuring thinking, feeling, and behavior. A wealth of research conducted over the past 30 years has shown that evaluations can be revealed not only via traditional self-report measures (explicit evaluations) but also via more indirect measures of response interference (implicit evaluations; Greenwald et al., 1998). One such measure, the Implicit Association Test (IAT; Greenwald et al., 1998) indexes relative evaluations of two targets (e.g., social groups, individuals, or objects) using a comparison of response latencies across two speeded sorting tasks: a first sorting task in which one of the targets shares a response key with positive items and the other target shares a response key with negative items, and a second sorting task in which the mapping of targets to valences is reversed. For instance, an implicit evaluation is inferred based on the speed and accuracy to associate the concept FLOWER (e.g., tulip, daisy) with pleasant attributes (e.g., angel, success) while associating the concept INSECT (e.g., bug, fly) with negative attributes (e.g., devil, failure) versus the opposite pairing of FLOWER with negative attributes and INSECT with positive attributes.

Implicit evaluations have been shown to predict behavior above and beyond their explicit counterparts in a range of consequential settings, including intergroup relations, consumer choice, psychopathology, and close relationships (Greenwald et al., 2009; Kurdi et al., 2018). For instance, implicit evaluations of African Americans measured at the level of geographic areas predict police brutality (Hehman, Flake, & Calanchini, 2017); implicit evaluations of products predict product usage and brand recognition (Maison, Greenwald, & Bruin, 2004); implicit evaluations of the self and self-injury predict suicidal behavior (Nock & Banaji, 2007); and implicit
evaluations of one’s romantic partner predict long-term relationship success (McNulty, Olson, Meltzer, & Shaffer, 2013). As such, understanding the processes by which implicit evaluations emerge and are updated is not only of theoretical interest across several areas of psychology but also of considerable practical and societal importance.

Implicit and explicit evaluations differ from each other in terms of the method by which they are measured. Implicit evaluations are usually indexed by tasks that bypass effortful control, with typical measures involving speeded responses to preselected pairs of stimuli. By contrast, explicit evaluations are usually measured using self-report (e.g., via responses to Likert items). Dominant dual-process theories of evaluation posit that, beyond differences in measurement, explicit and implicit evaluations also differ from each other in more profound ways. Crucially, explicit and implicit evaluations are commonly hypothesized to originate from fundamentally different learning processes. Specifically, the learning processes giving rise to explicit evaluations are posited to be flexible and rule-governed and to rely on propositional information, whereas the learning processes giving rise to implicit evaluations are posited to be slow and gradual and to rely on associative regularities encountered in the environment (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004).

Even though this dual-process perspective on evaluative learning has inspired much empirical work on the acquisition and change of explicit and implicit evaluations, it suffers from some notable shortcomings. First, in opposition to the theory, it has been repeatedly demonstrated that implicit evaluations can be flexibly updated via purely verbal instructions and that such updating need not involve direct experience with any stimulus (Kurdi & Banaji, 2017). The preponderance of such findings has prompted some to abandon a dual-process perspective on evaluative learning altogether and to replace it with a model of evaluative learning that relies on a sin-
gle propositional process (De Houwer, 2014; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009).

Second, dual-process theories of evaluation, as currently conceived, are difficult to falsify, and the same applies to single-process alternatives. For instance, whether learning is quick or slow is a matter of judgment and, as such, researchers with different theoretical commitments may make widely divergent inferences from the very same data. Moreover, it is unclear what kind of empirical evidence would be sufficient to discern whether evaluative information is represented in the form of conceptual associations (e.g., FLOWER–GOOD), as posited by dual-process theories of evaluation, or equivalence relationships (e.g., “Flowers are good”), as posited by propositional alternatives.

Third, implicit evaluations exhibit a host of characteristics that are not accounted for by theories that claim that they are subserved by a set of enduring associative representations accumulated over time. For instance, implicit evaluations have been shown to be situationally malleable (Blair, 2002). Specifically, implicit evaluations respond to motivational states, such as nicotine deprivation, thirst, and hunger (Ferguson & Bargh, 2008), as well as to higher-order goals, such as the goal to be egalitarian (Moskowitz, 2014). At the same time, contrary to the prediction by a single-process propositional perspective, implicit evaluations are not indiscriminately sensitive to verbal interventions that have been demonstrated to shift explicit evaluations (Lai et al., 2014).

Finally, dual-process and single-process theories of evaluation both make extreme predictions about the relationship that should emerge between explicit and implicit evaluations. According to dual-process theories, explicit and implicit evaluations are subserved by different learning processes and, as such, any convergence between the two is unexpected. By contrast,
according to single-process theories, explicit and implicit evaluations are subserved by a single learning process and, as such, there is no reason to expect them to diverge. However, the overwhelming majority of empirical data fall between the two extremes: Explicit and implicit evaluations are typically correlated with each other but are rarely redundant (Nosek, 2005), with the magnitude of the correlation modulated by the domain. For instance, explicit and implicit evaluations of political candidates have been found to be highly correlated, whereas explicit and implicit evaluations of racial groups often show considerably lower levels of correlation (Nosek, 2005).

Even though current dual-process theories of evaluation do not explain all the available evidence on the updating of implicit evaluations, they are not easily falsifiable, and are silent on a host of phenomena related to the malleability of implicit evaluations, it may be premature to abandon the class of dual-process theories altogether (Evans & Stanovich, 2013). In this paper, we use reinforcement learning algorithms, which originate from the study of animal learning (Dickinson & Balleine, 2002) and now play an important role in computer science and artificial intelligence (Sutton & Barto, 1998), to provide a new test of the manner in which explicit and implicit evaluations are acquired and updated. Specifically, we investigate whether explicit and implicit evaluations are equally or differentially sensitive to model-free and model-based learning. If both respond in similar ways, we can conclude that in spite of differences in measurement techniques, the representations underlying explicit and implicit evaluations are likely similar to each other. If, on the other hand, the two differ in their sensitivity to model-free and model-based learning, we can conclude that the data are more suggestive of differences in learning and representation.
In a reinforcement learning framework, an agent interacts with its environment and via such interaction it pursues two distinct, but interrelated, goals: (a) to learn about the actions that produce the largest amount of long-term reward and (b) to adjust behavior in line with this learning. Given the generality of this framework, rewards can range from primary reinforcers such as food or sex to more abstract rewards like points in a game or even social reinforcers like smiles and group inclusion. To solve the reinforcement learning problem described above (i.e., to maximize long-term reward), an agent must create internal representations of the subjective value associated with taking different actions. Most important for the present purposes, such representations can be created in two fundamentally different ways (Sutton & Barto, 1998): using model-free or model-based algorithms. The distinction between model-free and model-based processes has already been used with great success to elucidate phenomena across diverse areas of psychology, including moral cognition (Crockett, 2013; Cushman, 2013), impression formation (Hackel, Doll, & Amodio, 2015), and addiction (Everitt & Robbins, 2005). Here we use it to study the acquisition and shift of implicit evaluations for the first time.

Even though model-free and model-based algorithms solve the same problem of maximizing long-term reward, they differ from each other both in the way they learn and the kind of information that they are able to represent. In the present studies (see Figure 6), participants made choices between two fictitious social targets (Laapians vs. Niffians). Depending on their choice, they were then exposed to an intervening stimulus (a horizontal vs. a vertical bar), followed by a win or a loss. In this setting, the goal of both model-free and model-based algorithms is to learn whether, in the long run, choosing a Laapian target or choosing a Niffian target is the more advantageous action. However, they accomplish this goal in fundamentally different ways.
Figure 6. Overview of the learning phase procedure for Studies 1–3. The number of trials (for each part of the learning phase where applicable) is noted after the name of the condition. A hand symbol indicates a choice made by the participant. The assignment of first-stage stimuli (Laapians vs. Niffians) to second-stage stimuli (horizontal vs. vertical bars) as well as the assignment of second-stage stimuli to outcomes (+5 vs. -5 points) was counterbalanced across participants. Transitions between first-stage stimuli, second-stage stimuli, and outcomes were deterministic, with the exception of the control condition in Study 1 where second-stage stimuli were randomly followed by wins or losses, and Study 3 where initial learning in both conditions was probabilistic (one second-stage stimulus followed by wins on 75 percent of trials and by losses on 25 percent of trials and the other second-stage stimulus followed by losses on 75 percent of trials and by wins on 25 percent of trials). In the control (Study 1) and baseline learning (Studies 1–3) conditions, all dependent measures (transition memory, explicit evaluation, and implicit evaluation) were administered following the learning phase. In all remaining conditions, transition memory and explicit evaluation items were administered following each part of the learning phase, whereas implicit evaluations were measured only after the second part of the learning phase.
Model-free algorithms operate over an unordered list of actions, each of which is associated with a positive or negative scalar value. For instance, in the present studies, the model-free system may represent two actions (Choose Laapian vs. Choose Niffian) and, in the absence of prior learning, associate an initial value of zero with each. Over the course of the task, learning unfolds incrementally and based on direct experience: Each time the agent performs an action (e.g., choosing a Laapian target), it updates the value associated with that action based on its outcomes. For instance, if choosing a Laapian target results in a positive outcome (e.g., winning points), the agent increases the value associated with that action, and if it results in a negative outcome (e.g., losing points), the agent decreases the value associated with it. Incremental updating is performed until the prediction error is reduced to zero, i.e., there is no more discrepancy between the reward expected and actually received. Such an algorithm is computationally cheap: It creates action-value pairs, which constitute a highly compressed representation of the past history of rewards. However, the simplicity of this algorithm comes at a cost of reduced flexibility. Specifically, action-value pairs can be updated only upon performing an action. Moreover, given that specific outcomes of actions (e.g., Laapians leading to horizontal bars leading to wins) are not represented, the model-free system has no way to modulate its behavior based on higher-level goals (e.g., “I want to get to the horizontal bar”).

Unlike model-free algorithms that operate exclusively over action-value pairs (e.g., Laapian: +5; Niffian: -5), model-based algorithms operate over a considerably richer cognitive map of the environment that represents the specific outcomes of actions. For instance, in the context of the present experiments, a model-based agent would create a causal model linking first-stage stimuli to second-stage stimuli [e.g., “whenever I choose Laapians, I get to a horizontal bar” or P(horizontal | Laapian) = 1] and second-stage stimuli to rewards [e.g., “whenever I see a
horizontal bar, I win 5 points” or \( P(+5 \mid \text{horizontal}) = 1 \). This representation involves considerably more detail than the highly compressed representation created by model-free learning. Crucially, by virtue of representing specific outcomes of actions [e.g., \( P(\text{horizontal} \mid \text{Laapian}) = 1 \)], model-based learning can bypass the trial-by-trial updating that characterizes model-free learning. Instead, model-based representations enable mental simulation of different courses of action by considering the goal to be achieved (e.g., getting to the horizontal bar) and the probabilities with which different actions (e.g., choosing Laapians) can bring about the desired goal. As such, unlike model-free algorithms, model-based algorithms are highly flexible. However, their flexibility comes at a cost: Effortful planning over different courses of action may be prohibitively complex and time-consuming, especially if the number of potential actions and outcomes to plan over is large.

Both nonhuman animals (Adams & Dickinson, 1981; Dickinson, 1985) and humans (Daw et al., 2005; 2011) have been shown to exhibit both model-free and model-based learning. Under some conditions, model-free and model-based algorithms are difficult to tease apart because they converge on the same behavioral output. For instance, learning in the baseline learning condition of the present experiments (Figure 6) can be accomplished either via model-free or model-based processes: That is, participants may simply learn to associate higher value with the stimulus that led to a positive outcome in the past (e.g., Laapian: +5 and Niffian: -5). Alternatively, participants may explicitly represent the structure of the task, i.e., create a mental model of transition probabilities [e.g., \( P(\text{horizontal} \mid \text{Laapian}) = 1, P(+5 \mid \text{horizontal}) = 1, P(\text{vertical} \mid \text{Niffian}) = 1, P(-5 \mid \text{vertical}) = 1 \)].

However, the results of model-free and model-based learning can diverge when the environment changes in such a way as to modify the motivational relevance of a known stimulus.
Specifically, a paradigm commonly referred to as reward revaluation has been used to discern whether nonhuman animals (Adams & Dickinson, 1981; Dickinson, 1985) and humans (Daw et al., 2005; 2011) rely on model-free or model-based learning. In these studies, participants undergo initial learning that establishes that an action (e.g., pressing a lever or choosing an abstract image) is rewarding (e.g., results in the participant receiving food pellets or monetary rewards). In a second stage, the rewarding quality of the reinforcer is eliminated in the absence of the participant taking any relevant action: For instance, the previously satiating food pellets are paired with illness or the previously winning image is paired with monetary loss.

The signature difference between model-free and model-based learning is then revealed when participants are once again allowed to take the action that produced the previously rewarding outcome. Model-free algorithms are backward-looking and inflexible and, as such, can update values associated with an action only after that action has been performed and a reward has been experienced. Therefore, participants pursuing a purely model-free strategy will continue to consistently perform the action (e.g., pressing the lever or selecting the image) even following revaluation due to its history of producing reward. By contrast, model-based algorithms are forward-looking and flexible and, as such, have the ability to incorporate information about the new state of affairs. As such, participants pursuing a purely model-based strategy now expect that ingesting the food pellets will induce sickness or choosing the abstract image will result in a loss, and will thus consistently refrain from performing the previously reinforced action. In such paradigms, human participants usually pursue a mixture of both strategies (Daw et al., 2005; 2011); however, importantly, any decrease in the tendency to perform the initially reinforced action the can be interpreted as reflecting the contributions of model-based learning.

The present project
The present project has three interrelated goals. First, a large body of research has provided evidence that value representations in humans, as revealed by explicit measures of self-report, can be updated on the basis of the rewards received as a result of interacting with a given stimulus (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman et al., 2014; Otto et al., 2013). However, in spite of a similarly large body of research investigating the effects of mere exposure (Van Dessel et al., 2018), Pavlovian learning (Hofmann et al., 2010), approach–avoidance training (Van Dessel, De Houwer, & Gast, 2015a), and verbal instructions (Kurdi & Banaji, 2017) on implicit evaluations, the effects of reinforcement learning, i.e., rewarding or punishing participants for taking actions involving motivationally relevant stimuli, on implicit evaluations has never been investigated. As such, the first goal of the present project is to establish whether implicit evaluations of novel stimuli can be effectively shifted via this form of learning.

Second, and most important, the present project was designed to probe whether explicit and implicit evaluations of novel targets are equally sensitive to model-free and model-based learning. As mentioned above, explicit evaluations revealed by self-report have been demonstrated to be responsive to both model-free and model-based learning (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman et al., 2014; Otto et al., 2013). However, whether their implicit counterparts are characterized by the same or different patterns of updating is an open question, with different theoretical perspectives and lines of empirical work making opposing predictions about the pattern of data that should emerge.

A prediction of convergence between explicit and implicit evaluations can be made based on propositional theories of implicit evaluation (De Houwer, 2014; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009), widely replicated patterns of empirical data, and the nature of the Im-
plicit Association Test (IAT; Greenwald et al., 1998), which was used as the dependent measure in all three studies reported below. Specifically, propositional theories of implicit evaluation posit that explicit and implicit evaluations do not differ in underlying learning processes or mental representations. As such, if the present revaluation paradigm successfully shifts explicit evaluations of a stimulus, such learning should also be reflected by implicit measures of evaluation.

Second, in studies involving novel targets, such as the present one, patterns of convergence between explicit and implicit evaluations are common because participants do not have access to any information about the targets other than the information provided by the experimenter (Kurdi & Banaji, 2017). Moreover, in this setting, pressures to act in a socially desirable manner, known to result in dissociations between explicit and implicit evaluations (Nosek, 2005), are unlikely to operate. Finally, the IAT, unlike most implicit measures, requires participants to hold two to four categories in working memory while completing the task. This feature of the procedure may activate model-based representations.

On the other hand, dual-process theories of evaluative learning (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) generally predict dissociations between explicit and implicit evaluations. Specifically, reward revaluation, which can be accomplished via model-based but not via model-free processes, should be expected to shift explicit but not implicit evaluations: The use of information that is not represented in precompiled form to evaluate a stimulus may require effortful processing characteristic of the explicit system. Moreover, recent empirical work has revealed that placing participants under cognitive load while performing a reinforcement learning task shifts them toward reliance on the computationally cheap model-free system and away from reliance on the computationally expensive model-based system (Otto et al., 2013). Most implicit measures of evaluation, such as the IAT, place participants
under similar cognitive constraints given that they involve responding under time pressure. As such, this difference in the availability of cognitive resources across the explicit vs. implicit evaluation tasks may also contribute to a pattern of explicit–implicit dissociation.

Finally, to the extent that the present project provides evidence for a dissociation between explicit and implicit evaluations, it is important to demonstrate that such dissociation is due to a computational difference between model-free and model-based reinforcement learning. However, in a standard revaluation paradigm, use of model-free vs. model-based strategies is confounded with primacy vs. recency. Specifically, model-free learning would be revealed via reliance on initially learned information and model-based learning would be revealed via successful updating. Importantly, it has been shown that implicit evaluations, including implicit evaluations of novel targets, may be difficult to change once they are in place (Gregg et al., 2006). As such, any convincing claim about explicit–implicit dissociation being due to the model-free vs. model-based distinction has to involve a condition controlling for the temporal confound inherent in the reward revaluation paradigm.

**Study 1**

**Design**

The experiment consisted of a learning phase and one or two test phases (depending on condition). In the learning phase, participants interacted with two novel groups (Laapians vs. Niffians) and received rewards (positive points) or punishments (negative points) as a result of their choice behavior. In the test phases, participants provided forced-choice judgments probing (a) transition memory and (b) the value of the Laapian and Niffian targets (explicit evaluation), followed by (c) an Implicit Association Test (IAT; Greenwald et al., 1998) probing implicit evaluation of the same targets.
Crucially, for the learning phase of the experiment, participants (final $N = 1,740$) were assigned to one of five between-subjects conditions (Figure 6). In the control and baseline learning conditions, the learning phase consisted of a single part, whereas in the reward revaluation, transition revaluation, and relearning conditions, the learning phase consisted of two parts.

Across all five learning conditions, the first part of the learning phase required participants to complete 20 learning trials on which they made a choice between a Laapian and a Niffian target (first-stage stimuli; see Materials and Methods). Depending on their choice, participants were exposed to a horizontal or a vertical bar (second-stage stimulus), followed by a positive outcome (+5 points) or a negative outcome (-5 points). Participants were instructed to maximize the points received. The relationship between first-stage and second-stage stimuli was deterministic in all five conditions (e.g., Laapians were always followed by horizontal bars and Niffians by vertical bars). In the control condition, second-stage stimuli were randomly followed by wins or losses, thus providing a measure of relative preference at baseline. In all four remaining conditions, the transition between second-stage stimuli and rewards was deterministic (e.g., horizontal bars were always followed by wins and vertical bars by losses).

In the reward revaluation, transition revaluation, and relearning conditions, the first part of the learning phase was followed by a second part, also consisting of 20 trials. In the reward revaluation condition, the transition between second-stage stimuli and rewards was reversed compared to the first part of the learning phase (without participants making any choices or experiencing any first-stage stimuli). In the transition revaluation condition, the transition between first-stage and second-stage stimuli was reversed compared to the learning phase (without participants making any choices or experiencing any rewards). The relearning condition was similar to the reward revaluation condition in that the transition between second-stage stimuli and rewards
was reversed; however, unlike in the reward revaluation condition, participants experienced the
full transition structure from first-stage stimuli to second-stage stimuli to rewards and rather than
passively observing stimuli, they made choices between Laapian and Niffian targets.

In the control and baseline learning conditions, the learning phase was followed by (a) a
set of explicit transition memory items probing memory for the transition between first-stage and
second-stage stimuli, (b) a set of explicit evaluation items probing self-reported subjective value
assigned to each target (Laapians vs. Niffians), and (c) an IAT probing relative implicit evalua-
tion of Laapians vs. Niffians. In the reward revaluation, transition revaluation, and relearning
conditions, the explicit transition memory and explicit evaluation items were administered twice,
once after the first part of the learning phase and once after the second part of the learning phase.
However, to prevent participant fatigue, the IAT was administered only once, following the se-
cond part of the learning phase.

The logic of the statistical analyses reported below is the following. First, a comparison
involving the control and baseline learning conditions can be used to establish whether the re-
wards and punishments used in the present task were effective in shifting participants’ explicit
and implicit evaluations of first-stage stimuli. Explicit evaluations have been demonstrated to
shift as a result of similar manipulations numerous times (Daw et al., 2005; 2011; Fu & Ande-
son, 2008; Gershman et al., 2014; Otto et al., 2013); however, the present project constitutes the
first test of whether a binary choice between two targets, followed by rewards and punishments,
can successfully shift implicit evaluations as measured by the IAT.

Second, a crucial comparison involving the baseline learning and reward revaluation
conditions can be used to probe whether explicit and implicit evaluations are sensitive to model-
based learning. As noted above, successful updating of subjective value in the reward revaluation
condition is commonly interpreted to rely only on model-based processes given that the second part of the learning phase did not involve any experience with first-stage stimuli. Similar to the first comparison, the effectiveness of reward revaluation in shifting explicit evaluations has already been demonstrated (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman et al., 2014; Otto et al., 2013); by contrast, whether reward revaluation can shift implicit evaluations has not been investigated before.

Third, a comparison involving the baseline learning and transition revaluation conditions can be used to probe whether explicit and implicit evaluations are sensitive to a different kind of change in the environment. The predictions for this comparison are less straightforward than for the reward revaluation condition given that updating in this condition may occur via either model-free or model-based processes, or a combination of both: Model-based updating may be performed if participants use their explicit model of the task to cognitively link the second-stage stimuli to rewards (as experienced in the first part of the learning phase). However, because second-stage stimuli were paired with wins and losses in the first part of the learning phase, they might act as valenced stimuli themselves, thus enabling model-free learning (akin to second-order conditioning).

Fourth, a comparison involving the baseline learning and relearning conditions can be used to help disambiguate the results of the reward revaluation condition by revealing whether implicit evaluations are differentially sensitive to (a) model-free vs. model based learning or (b) initial learning vs. subsequent updating (i.e., a primacy effect; Gregg et al., 2006). Specifically, if implicit evaluations were to be insensitive to model-based learning, such insensitivity would be reflected by statistically equivalent responding in the baseline learning and reward revaluation conditions (see above). However, this pattern of responding may also be the result of implicit
evaluations being generally more responsive to initial learning than to updating based on novel information. If this is the case, and implicit evaluations are generally impervious to updating, no difference would be expected between the baseline learning and relearning conditions given that the relearning condition, just as the reward revaluation condition, involves initial learning followed by updating. However, if the defining difference is between model-free and model-based processes, the relearning condition, unlike the reward revaluation condition, should show change given that in the former, unlike in the latter, learning can be accomplished via model-free processes.

Results

The pattern of results obtained with explicit evaluation as the dependent measure (Figure 7) was in line with expectations formulated on the basis of similar studies conducted in the past (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman et al., 2014; Otto et al., 2013) and, as such, underscores the soundness of the design and manipulations.
Figure 7. Study 1 ($N = 1,740$): Mean explicit and implicit evaluations by learning condition. For explicit evaluations (left pane) the y axis shows percentage of responses in line with initial learning; for implicit evaluations (right pane) the y axis shows Implicit Association Test (IAT) D scores (Greenwald et al., 2003) computed such that higher values indicate responses in line with initial learning. For explicit and implicit evaluations, effects of revaluation or relearning are revealed by values closer to 0% or negative D scores, respectively. In the control condition, responses indicating preference in favor of Laapians over Niffians were arbitrarily coded as positive. For visualization purposes, IAT scores have been mean centered using the mean of the control condition. Error bars show 95% confidence intervals.

Specifically, baseline learning was found to be effective in shifting explicit evaluations compared to the control condition, $t(548.86) = 9.88, p < .001, BF_{10} = 3.40 \times 10^{18}$, Cohen’s $d = 0.82$, thus establishing the general effectiveness of the learning task used in the present study (see also Supplementary Studies 1 and 2 in Supplement 2, [https://osf.io/f8pg3/](https://osf.io/f8pg3/)). Also in line with expectations, reward revaluation was effective in shifting explicit evaluations compared to the baseline learning condition, $t(474.09) = 14.49, p < .001, BF_{10} = 5.89 \times 10^{38}$, Cohen’s $d = 1.22$, thus replicating the widely observed finding that explicit evaluations respond to model-based learning (see also Supplementary Study 1 in Supplement 2, [https://osf.io/f8pg3/](https://osf.io/f8pg3)). A similar result was observed for the comparison involving baseline learning and transition revaluation, $t(502.54) = 10.44, p < .001, BF_{10} = 5.06 \times 10^{20}$, Cohen’s $d = 0.86$, which should not be surprising given that such updating could have occurred either via model-free or model-based processes. Finally, although not of major theoretical relevance for the present purposes, the relearning condition was also found to effectively shift explicit evaluations compared to the baseline learning

---

41 In this condition, as well as the transition revaluation and relearning conditions, participants completed two sets of explicit evaluation items, one following the first part of the learning phase (initial learning) and one following the second part of the learning phase (revaluation or relearning). To ensure compatibility of analyses across conditions and across explicit and implicit measures, here we report comparisons between the baseline learning condition and the second set of explicit evaluation items completed following revaluation or relearning. In Supplement 2 ([https://osf.io/f8pg3/](https://osf.io/f8pg3)) we report additional within-participant analyses comparing the first set to the second set of explicit evaluation items in the reward revaluation, transition revaluation, and relearning conditions. These within-participant analyses reinforce the conclusions reported here.
condition, \( t(793.91) = 24.55, p < .001, \) BF\(_{10} = 1.62 \times 10^{85}, \) Cohen’s \( d = 1.59 \) (see also Supplementary Study 3 in Supplement 2, [https://osf.io/f8pg3/](https://osf.io/f8pg3)).

Given that learning in the baseline learning and relearning conditions could have been accomplished either in a model-free or a model-based way, we investigated \((a)\) whether explicit evaluations differed from neutrality at chance-level responding to the transition memory item (a signature of model-free processes) and \((b)\) whether accuracy of transition memory predicted explicit evaluations (a signature of model-based processes) in each condition. As revealed by a significant intercept, explicit evaluations differed from neutrality at chance-level responding to the transition memory item both in the baseline learning condition, \( b = 0.90, t(307) = 11.76, p < .001, \) and in the relearning condition, \( b = 0.47, t(573) = 6.07, p < .001, \) thus providing evidence for the operation of model-free processes. At the same time, accurate transition memory positively predicted explicit evaluations in both conditions, \( b = 0.43, t(307) = 8.77, p < .001, \) and \( b = 0.38, t(573) = 8.83, p < .001, \) revealing the contribution of model-based processes to the acquisition of explicit evaluations.

A comparison involving the control and baseline learning conditions revealed that implicit evaluations, like explicit evaluations, were sensitive to reinforcement learning: Scores on the Implicit Association Test (IAT) exhibited significant change away from control as a result of the rewards received in the baseline learning condition, \( t(565.06) = 4.35, p < .001, \) BF\(_{10} = 9.11 \times 10^{2}, \) Cohen’s \( d = 0.36 \) (see also Supplementary Studies 1 and 2 in Supplement 2, [https://osf.io/f8pg3/](https://osf.io/f8pg3)). The crucial comparison in this experiment involved the baseline learning vs. reward revaluation conditions given that this comparison establishes whether implicit evaluations responded to model-based reinforcement learning. This comparison provided evidence in favor of the null hypothesis, \( t(569.05) = 1.06, p = .287, \) BF\(_{01} = 6.22, \) Cohen’s \( d = 0.09, \) suggest-
ing that implicit evaluations are impervious to model-based updating (see also Supplementary Study 1 in Supplement 2, https://osf.io/f8pg3/). In line with the expectation that updating in the transition revaluation condition may emerge from model-free or model-based processes, we found weak evidence that the transition revaluation condition may have differed from the baseline learning condition, $t(591.29) = 2.47, p = .013, BF_{10} = 1.85$, Cohen’s $d = 0.20$.

Finally, given that we found no updating compared to baseline in the reward revaluation condition, a comparison involving the baseline learning and relearning conditions can be used to establish whether such lack of updating occurred due to a general primacy effect or due to the more specific effect of implicit evaluations being impervious to model-based, but not model-free, updating. The baseline learning and relearning conditions were found to significantly differ from each other, $t(649.58) = 6.04, p < .001, BF_{10} = 2.40 \times 10^6$, Cohen’s $d = 0.42$, suggesting that already established implicit evaluations can be effectively updated provided that the updating can be performed via model-free mechanisms. As such, this result eliminates a general primacy effect as an explanation for the present findings (see also Supplementary Study 3 in Supplement 2, https://osf.io/f8pg3/).

**Studies 2–3**

**Design**

Study 1 has provided initial evidence that, unlike their explicit counterparts, implicit evaluations are impervious to model-based learning. Studies 2 and 3 were designed to provide a direct replication of this result as well as to produce evidence about its generality (see Figure 6).

In line with our primary focus on the sensitivity of implicit evaluations to model-based learning, Studies 2 and 3 consisted only of baseline learning and reward revaluation conditions. In Study 2 (final $N = 245$), in addition to the baseline learning condition, two versions of the re-
ward revaluation condition were implemented. In the first version (revaluation 20 condition), participants were exposed to 20 revaluation trials. As such, this condition provides a direct replication of the reward revaluation condition from Study 1. In the second version (revaluation 40 condition), participants were exposed to 40, rather than 20, revaluation trials. As such, in this condition, participants experienced twice as many trials in the second part of the learning phase (revaluation) than in the first part (initial learning). If a comparison of implicit evaluations across the baseline learning and revaluation 40 conditions reveals no difference, this would suggest that the lack of updating observed in Study 1 was likely due to the insensitivity of implicit evaluations to model-based learning rather than a lack of sufficient training in the second part of the learning phase.

In Study 3 (final $N = 369$), reinforcement in the baseline learning condition and in the first part of the revaluation condition (initial learning) was probabilistic rather than deterministic, with one of the targets (e.g., Laapians) followed by wins 75 percent of the time and the other target (e.g., Niffians) followed by losses 75 percent of the time. To provide a conservative test of the null hypothesis of model-based learning being ineffective in shifting implicit evaluations, revaluation was deterministic. This study was designed to probe whether the insensitivity of implicit evaluations to model-based learning, as established by Study 1, may be modulated by the ambiguity of the initially received evaluative information. As such, the reinforcement contingencies in this study were more ecologically realistic than contingencies in Studies 1 and 2 where one of the targets was deterministically followed by wins and the other target was deterministically followed by losses.

**Results**
In Study 2, explicit evaluations shifted significantly as a result of revaluation in both the revaluation 20 condition, $t(124.17) = 8.31, p < .001, BF_{10} = 2.04 \times 10^{12}$, Cohen’s $d = 1.33$, and in the revaluation 40 condition, $t(115.47) = 7.42, p < .001, BF_{10} = 1.44 \times 10^{10}$, Cohen’s $d = 1.22$, thus replicating the results obtained in Study 1. In Study 3, explicit evaluations were also found to shift significantly as a result of reward revaluation, although the evidence in favor of change was considerably weaker than in Studies 1 and 2, $t(351.96) = 2.16, p = .031, BF_{10} = 1.10$, Cohen’s $d = 0.23$.

Crucially, replicating the results of Study 1, implicit evaluations were found to be impervious to reward revaluation in both Study 2 and Study 3. Specifically, Study 2 provided evidence in favor of the null hypothesis when the number of trials was the same across the first and second parts of the learning phase (baseline learning vs. revaluation 20 conditions), $t(154.74) = -0.87, p = .386, BF_{01} = 4.19$, Cohen’s $d = -0.14$. A similar result was obtained in the revaluation 40 condition where the number of revaluation trials was double the number of the initial learning trials, $t(152.71) = -0.51, p = .612, BF_{01} = 5.28$, Cohen’s $d = -0.08$. Implicit evaluations also remained impervious to reward revaluation in Study 3, demonstrating that their insensitivity to model-based learning does not depend on the deterministic vs. probabilistic nature of initial reinforcement, $t(366.99) = 0.32, p = .747, BF_{01} = 8.26$, Cohen’s $d = -0.03$.

**Results combined across experiments**

Bayesian meta-analyses were conducted to obtain an aggregate measure of differences across the baseline learning and reward revaluation conditions in Studies 1–3. Explicit evaluations were found to be sensitive to model-based learning, $BF_{10} = 2.58 \times 10^{43}$, Cohen’s $d = 0.87$, 95% HDI: [0.77; 0.99], replicating previous work (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman et al., 2014; Otto et al., 2013). By contrast, implicit evaluations were found to
be impervious to model-based learning, BF$_{01}$ = 13.18, Cohen’s $d = 0.03$, 95% HDI: [-0.08; 0.15].

Additional meta-analyses conducted only with participants who had perfect transition memory revealed the same pattern of results (Supplement 2, [https://osf.io/f8pg3/](https://osf.io/f8pg3/)), suggesting that lack of updating in implicit evaluations did not result from an erroneous representation of the structure of the environment.

To further compare the relative importance of transition memory in shaping implicit versus explicit evaluations, we conducted a small-sample corrected robust meta-regression (Tipton, 2015) with correlation between transition memory and evaluation as the dependent measure, a fixed effect for type of evaluation (implicit vs. explicit), and a random effect for study and condition to account for dependency in the data. For implicit measures, no relationship was found between transition memory and evaluation, $b = 0.013$ [-0.028; 0.054], $t(7.91) = 0.75$, $p = .476$. By contrast, for explicit measures, transition memory positively and significantly predicted evaluations, $b = 0.237$ [0.072; 0.390], $t(7.91) = 3.29$, $p = .011$. As such, this meta-analysis provides additional correlational evidence for the idea that only explicit, but not implicit, evaluations are responsive to model-based learning.

**Discussion**

We conducted three experiments relying on the distinction between model-free and model-based reinforcement learning (Daw et al., 2005; Sutton & Barto, 1998) and involving novel stimuli to arrive at a better understanding of the updating of implicit (indirectly measured) evaluations. Model-free algorithms are backward-looking, incremental, and computationally cheap: They adjust the value of an action upon experiencing its motivationally relevant outcomes. By contrast, model-based algorithms are forward-looking, flexible, and computationally expensive:
They perform planning over a causal model of the environment to choose the best course of action in light of current goals.

The model-free vs. model-based distinction seemed ideal as a theoretical basis for this investigation because, unlike existing dual-process and single-process theories of evaluation, it is computationally well-specified: The signatures of model-free vs. model-based processes can be revealed in a so-called reevaluation paradigm (Adams & Dickinson, 1981; Daw et al., 2005; 2011; Dickinson, 1985). In this paradigm, subjective evaluation of a well-known and previously rewarding stimulus is measured after the stimulus loses its rewarding quality. Change in choice behavior as a result of this new information reveals the operation of model-based learning, whereas persistence of the old choice behavior is characteristic of a model-free strategy. Explicit evaluations have been known to reflect a combination of model-free and model-based processes (Daw et al., 2005; 2011; Fu & Anderson, 2008; Gershman et al., 2014; Otto et al., 2013) and here we replicate this result.

The novel contribution of the present project is twofold. First, we show that, in addition to other forms of evaluative learning such as mere exposure (Van Dessel et al., 2018), Pavlovian learning (Hofmann et al., 2010), approach–avoidance training (Van Dessel, De Houwer, & Gast, 2015a), and verbal instructions (Kurdi & Banaji, 2017), implicit evaluations of stimuli, similar to their explicit counterparts, are amenable to updating as a result of reinforcement learning, i.e., experience with the positive and negative outcomes of actions involving those stimuli. Second, we demonstrate both a commonality and a difference in the computations underpinning the updating of explicit vs. implicit evaluations via reinforcement learning: Just like explicit evaluations, implicit evaluations were found to be responsive to model-free processes both at baseline
and following initial model-free learning with different reinforcement contingencies. However, unlike their explicit counterparts, implicit evaluations were insensitive to model-based learning.

Such dissociation between explicit and implicit evaluations is surprising for a number of reasons. First, our own previous work has shown that implicit evaluations can shift in the face of propositional processes traditionally thought of as uniquely influencing explicit evaluations, suggesting underlying commonality in learning (Kurdi & Banaji, 2017). However, here we provide clear evidence for a theoretically meaningful explicit–implicit dissociation. Second, to provide a conservative test of dissociation, explicit items were always administered to participants before the Implicit Association Test (IAT) and, as such, responding on the IAT could have been influenced not only by the learning manipulations but also by responding on the explicit measure of evaluation. Yet, no spillover effects were observed, suggesting separate underlying representations. Third, the IAT, used as a measure of implicit evaluation in the present studies, involves explicit categorization of stimuli, which may have been predicted to activate model-based value representations, but the data showed no such evidence. Fourth, all three experiments involved novel social groups as targets. This feature should have minimized social desirability concerns, which are known to contribute to explicit–implicit dissociations in tests of real social targets (Nosek, 2005).

However, although the present studies created a pattern of dissociation between explicit and implicit evaluations, it should be noted that a reinforcement learning perspective, unlike a traditional dual-process perspective (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004), does not make an unqualified prediction of explicit–implicit dissociation.

Even though the stimuli were introduced to participants as “names […] com[ing] from two groups,” it should be pointed out that the stimuli were not imbued with any further social meaning, nor did we replicate the present experiments using stimuli explicitly described as nonsocial. As such, we leave a direct comparison of social vs. nonsocial and novel social vs. real social stimuli in the present paradigm to future work.
tions, for multiple reasons. First, in many situations, including the baseline learning condition of the present studies, model-free and model-based algorithms converge on the same value representation. Second, as demonstrated by the present studies, explicit and implicit evaluations can both be updated by model-free processes. This shared learning process should generally lead to some degree of association between explicit and implicit evaluations. Third, recent research has shown that model-free and model-based learning need not be antagonistic: On the contrary, a model of the environment can be used to modulate model-free value representations via simulated experience (Gershman et al., 2014; Gershman, Zhou, & Kommers, 2017). Future work may test this idea in the context of implicit evaluations by imposing a delay between the revaluation and test phases of the experiment.

In addition, the distinction between model-free and model-based learning processes provides a theoretical framework to explain why certain interventions, even those that do not involve valenced feedback upon performing an action, can successfully shift implicit evaluations, whereas others seem to be ineffective. Out of 17 interventions implemented in a recent large-scale collaboration, with individual investigators submitting their chosen intervention, eight shifted implicit evaluations of African Americans toward neutrality, whereas nine produced no change (Lai et al., 2014). Among the eight interventions that were effective, five were clearly better characterized as model-free: They included direct experience with African American exemplars paired with positive stimuli or outcomes [e.g., evaluative conditioning (Hofmann et al., 2010)]. The remaining three manipulations that were effective required a mental model given that they were based on verbal instructions rather than direct experience; however, this mental model was of the simplest possible form: \( P(\text{positive} \mid \text{African American}) = 1 \) and \( P(\text{negative} \mid \text{White American}) = 1 \). This group of interventions included a vivid story in which the protagonist
was assaulted by a White American and saved by a Black American as well as two manipulations involving implementation intentions (Gollwitzer, 1993). From a reinforcement learning perspective, it could be argued that the causal model involved in the latter group of interventions is sufficiently simple to be able to train model-free values almost immediately (see also Kurdi & Banaji, 2017).

By contrast, among the nine ineffective interventions, eight involved a complex causal model of the environment, including a model of another persons’ mind, a model of a positive encounter with an outgroup member, and a model of racial injustice. Crucially, unlike the successful interventions described above, this set of interventions did not provide participants with pre-compiled value representations (e.g., African American: good; White American: bad) that could be activated quickly and effortlessly while responding under time pressure on an implicit task. Given this time pressure, participants may not have had sufficient opportunity to discern what modulation of existing value representations a complex causal model would imply. Finally, the only model-free intervention that remained ineffective involved pairings of both Black and White Americans with (a) positive and negative facial expressions and (b) positive and negative feedback. Given the nature of reinforcement provided in this intervention, no change in the model-free values associated with each target should be expected.

This novel perspective on the results reported by Lai et al. (2014) is consistent with the idea that evaluative representations acquired in ways other than via reinforcement learning may generally be able to effectively drive responding on implicit measures such as the IAT only if they are sufficiently compressed to enable automatic activation under time pressure. Future work will be able to offer more systematic tests of this idea. For instance, the model-free vs. model-based distinction underpinning the present project may be used to probe whether implicit evalua-
tions are amenable to revaluation in a Pavlovian setting (Dayan & Berridge, 2014). Moreover, when it comes to purely language-based learning (De Houwer, 2014), the present results suggest that the effectiveness of verbal statements in updating implicit evaluations may be moderated by the complexity of the propositional reasoning required to assign the appropriate truth value to those verbal statements or, in the terminology of reinforcement learning, by the complexity of the implied causal model.

Moreover, the present results as well as a general reinforcement learning framework provide a new perspective on what is usually described as the sensitivity of implicit evaluations to higher-order goals (Ferguson & Bargh, 2008; Moskowitz, 2014). In studies of this kind, activation of a goal (such as hunger, achievement, or egalitarianism) leads to a modulation of implicit evaluations such that objects that can contribute to achieving the goal are temporarily evaluated more positively until the goal is successfully completed. These findings are seemingly at odds with the present results, given that, as mentioned above, only model-based, and not-model free, value representations can be modulated in the face of higher-order goals.

However, the contradiction between both perspectives may be illusory. One group of variables investigated in this set of studies (including nicotine deprivation, thirst, and hunger) are more appropriately described as motivational states rather than goals. Sensitivity of model-free reinforcement learning to motivational states is compatible both with the theoretical formulation of model-free algorithms (Niv, Joel, & Dayan, 2006) and empirical findings (Dickinson & Balleine, 2002): Based on past experience, a model-free learner can represent multiple value estimates associated with the same action (e.g., smoking a cigarette). Smoking a cigarette in a nicotine-deprived state is highly rewarding and smoking a cigarette nicotine-satiated state is much less so. Accordingly, over time, a smoker should learn to associate higher value with cigarettes in
the former compared to the latter context and activate the appropriate value representation depending on their current motivational state.

A second set of variables used in this literature can be described as genuine higher-order goals (such as achievement or egalitarianism). However, the general finding involving such goals is that they modulate responding on implicit measures only to the extent that participants have protracted past experience with them (e.g., professional athletes or chronic egalitarians). In a reinforcement learning framework, such past experience is equivalent to having accumulated corresponding model-free value representations over time. These model-free representations can then be activated automatically upon encountering the relevant motivational state without such activation requiring genuine goal-directed behavior involving effortful planning over a causal model. As such, in line with our observation above, motivational states and goals seem to modulate implicit evaluations only to the extent that they provide a precompiled value representation that can be activated automatically and effortlessly during an implicit task.

The present findings demonstrating the sensitivity of implicit evaluations to model-free learning and their insensitivity to model-based learning may be expanded upon in a number of ways in future work. For instance, as mentioned above, model-free learning is inherently state-dependent, which may provide an explanation for the highly contextualized nature of implicit evaluation (Blair, 2002) as well as its resistance to long-term change (Lai et al., 2016). By mapping out the space of relevant states and providing model-free training across a large number of them, change in implicit evaluations may become more robust, enduring, and generalizable. Moreover, as mentioned above, the present studies were designed to produce a dissociation between model-free and model-based learning; however, recent work on offline updating of model-free value representations via model-based algorithms (Gershman et al., 2014; 2017) suggests
that model-based interventions may also be successfully used to shift implicit evaluations. Beyond these specific proposals for future work, it is our hope that the theoretical framework outlined here will generally inspire much insightful theorizing and empirical research on when, how, and why implicit evaluations change.

Materials and Methods

Institutional approval and informed consent

All studies reported here were granted ethical approval by the Committee on the Use of Human Subjects at Harvard University. Participants provided informed consent at the beginning of each study.

Participants

Participants in all studies were American adult volunteers recruited from the Project Implicit educational website (http://implicit.harvard.edu). Exclusion criteria are reported in Supplement 2 (https://osf.io/f8pg3/).

Learning phase

In the initial part of the learning phase of Studies 1 and 2, participants were exposed to 20 forced choice trials. In Study 3, the number of forced choice trials was increased to 30. On each trial, a Laapian stimulus (randomly selected from Caalap, Feelslap, Gabeelap, Ineelap, and Maasolap) and a Niffian stimulus (randomly selected from Ibbenif, Jabbunif, Lebbunif, Mettanif, and Oballnif) were presented side-by-side on the screen. Participants selected the left-hand stimulus by pressing the E key or the right-hand stimulus by pressing the I key. The side on which Laapian and Niffian stimuli were presented was randomly selected for each trial. Following participants’ choice, a second-stage stimulus (horizontal or vertical bar) was displayed. Once participants pressed the space bar, the second-stage stimulus was removed and a reward (+5 or -5) ap-
peared. The next trial started upon pressing the space bar. The transition from first-stage stimuli (Laapians vs. Niffians) to second-stage stimuli (horizontal vs. vertical bars) to rewards (+5 vs. -5 points) was counterbalanced across participants.

The learning phase of the control (Study 1) and baseline learning (Studies 1–3) conditions consisted of only the initial learning described above. In the reward revaluation (Studies 1–3), transition revaluation (Study 1), and relearning (Study 1) conditions, a second part followed. In Study 1, the second part of the learning phase consisted of 20 trials, in Study 2 it consisted of 20 or 40 trials (depending on condition), and in Study 3 it consisted of 30 trials. In the reward revaluation conditions, participants were exposed to a randomly selected second-stage stimulus (horizontal or vertical bar) on each trial. Once they pressed the C key, a reward (+5 or -5 points) was revealed. The next trial started upon the participant pressing the space bar. The transition revaluation condition was similar, with the exception that participants were exposed to first-stage stimuli (Laapians or Niffians) and, upon pressing the C key, a second-stage stimulus (horizontal or vertical bar) appeared. The second part of the learning phase in the relearning condition of Study 1 was identical to the first part, with the exception that the transition from second-stage stimuli (horizontal vs. vertical bars) to rewards (+5 vs. -5 points) was reversed.

**Explicit evaluation**

Explicit evaluation items were identical to the forced-choice trials used in the first part of the learning phase; however, on these trials participants received no feedback. Participants in the control and baseline learning conditions completed a single set of four explicit evaluation items, whereas participants in the reward revaluation, transition revaluation, and relearning conditions completed two sets of four explicit evaluation items: one set following the first part of the learning phase (initial learning) and a second set following the second part of the learning phase (re-
valuation or relearning). Responses on each set of explicit evaluation items were summed (1 = normatively accurate response, 0 = normatively inaccurate response) to create an index of explicit evaluation.

**Transition memory**

Transition memory items were identical to the explicit evaluation items, with the exception that participants were asked to select the first-stage stimulus (Laapian vs. Niffian) that leads to a certain second-stage stimulus (horizontal vs. vertical bar) rather than to a positive outcome. The administration of transition memory items followed the same schedule as the administration of explicit evaluation items. Responses on each set of transition memory items were summed (1 = normatively accurate response, 0 = normatively inaccurate response) to create an index of transition memory.

**Implicit evaluation**

Implicit evaluations were measured using a standard five-block Implicit Association Test (IAT; Greenwald et al., 1998). The categories were “Laapians” and “Niffians” and the attributes were “good” and “bad.” Category items were identical to the items used during the learning phase. Good attribute items included love, peace, joy, happy, peace, glory, and lucky; bad attribute items included hate, war, devil, bomb, bitter, agony, and grief. Implicit evaluations were calculated using the improved scoring algorithm (Greenwald et al., 2003) such that higher D scores indicate evaluations in line with initial learning. Further details of the IAT procedure are reported in Supplement 2 (https://osf.io/f8pg3/).

**Statistical analyses**

All statistical analyses were conducted in the R statistical computing environment. The R code for all analyses, as well as data files (including trial-level IAT data) are freely available
from the Open Science Framework (https://osf.io/f8pg3/). The Bayesian $t$ tests and Bayesian meta-analyses were performed using the BayesFactor package (Morey, Rouder, & Jamil, 2015). The small-sample corrected robust variance meta-regression was conducted using the robumeta package (Fisher & Tipton, 2015).
General Conclusion

This dissertation consists of three papers, relying on data from 11,962 participants across 13 studies, that use a learning and memory approach to investigate the nature of the processes that give rise to implicit, i.e., automatically activated, evaluations of social and nonsocial stimuli.

Paper 1 (Kurdi & Banaji, 2017, *J Exp Psychol Gen*) provides evidence that, when measured immediately following learning, verbal information about upcoming stimulus pairings (“group X will be paired with positive images”) is at least as effective, if not more effective, than actual experience with stimulus pairings in shifting implicit evaluations. Moreover, a combined intervention of verbal information followed by actual stimulus pairings has been shown not to outperform an intervention relying on verbal information alone.

In Paper 2 (Kurdi & Banaji, 2019, *J Pers Soc Psychol*), we conducted further tests of the processes by which exposure to stimulus pairings influences implicit evaluations. We found that learning asymptoted quickly, after as few as four stimulus pairings, and that learning from stimulus pairings was modulated by purely verbal information on the nature of stimulus pairings. Finally, the results of this paper suggest that exposure to stimulus pairings may create more durable implicit evaluative learning than purely language-based interventions; however, this advantage is eliminated if stimulus pairings are preceded by verbal information.

Finally, in Paper 3 (Kurdi, Gershman, & Banaji, 2019, *Proc Natl Acad Sci USA*), we investigated, for the first time, the effects of motivationally relevant consequences of interacting with stimuli on the implicit evaluations of those stimuli. We found that reinforcement learning significantly modulated not only explicit attitudes, but also their implicit counterparts, suggesting that this kind of learning may represent an important, and as yet largely unexplored, avenue toward implicit attitude change. Importantly, this paper also provides evidence for a computation-
ally tractable single dissociation between explicit and implicit evaluations: Whereas explicit
evaluations were found to respond both to model-free and model-based learning processes, im-

clict evaluations were sensitive to the former but impervious to the latter.

Taken together, these findings (a) can help arbitrate between currently available theories
of implicit evaluation, (b) can form the basis of future theory development within a reinforce-
ment learning framework that goes beyond both existing associative theories and existing proposi-
tional accounts, and (c) have practical implications for the more applied endeavor of achieving
long-term change in implicit evaluations of real-world stimuli, such as social groups (Lai et al.,
2016) or motivationally relevant stimuli such as cigarettes or alcohol (Lindgren, Neighbors, Gass-
er, Ramirez, & Cvencek, 2017).

With regard to theories of implicit evaluation, the present findings are difficult, if not im-
possible, to reconcile with associative accounts, as currently formulated (Rydell & McConnell,
2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004). Specifically, these theories posit
that implicit evaluations (a) should shift exclusively, or at least primarily, in response to direct
experience with valenced stimuli and not in response to purely verbal manipulations, (b) in a slow,
incremental, and purely stimulus-driven manner, and (c) should be subserved by associ-
ative representations stored in long-term memory (e.g., OTTERS–GOOD). By contrast, the present
findings seem more compatible with more recent propositional accounts (De Houwer, 2014;
2018; De Houwer & Hughes, 2016) according to which implicit evaluations (a) may shift either
in response to direct experience with valenced stimuli or in response to purely verbal manipula-
tions, (b) in a quick, sudden, and inferential manner, and (c) are subserved by propositional rep-
resentations stored in long-term memory (e.g., “Otters are good”).
Specifically, the present studies provide evidence that direct experience with valenced stimuli has no privileged role in shifting implicit evaluations: In Paper 1, repeated evaluative pairings were often outperformed by evaluative statements in their effectiveness in creating implicit attitude change and, in Paper 2, verbal information on diagnosticity was productively combined with information derived from stimulus pairings in giving rise to implicit attitudes. Such results directly contradict current associative theories of implicit evaluation that assign a special role to stimulus pairings in driving implicit attitude change. By contrast, they are easily accommodated by propositional theories according to which implicit attitudes can shift as a result of information derived from multiple sources, including stimulus pairings and purely verbal manipulations.

Moreover, the present studies (Papers 1–3) provide joint evidence that implicit attitudes can shift quickly and implicit attitude change does not require the kind of protracted experience with valenced stimuli that, according to associative theories, should be necessary for learning to emerge. Similar to the data on the role of direct experience versus language-based learning, this aspect of the data seems to favor propositional over associative approaches to implicit evaluation. However, the present work provides mixed evidence on the role of explicit recollection of the learning phase in predicting implicit evaluative learning: In the context of evaluative conditioning, Paper 2 has demonstrated, in line with previous work (Van Dessel, De Houwer, & Gast, 2015a; Van Dessel, De Houwer, Roets, & Gast, 2016b; Van Dessel, Mertens, Smith, & De Houwer, 2017b), that contingency awareness is a major moderator of implicit attitude change. On the other hand, Paper 3 found that, in the context of reinforcement learning, transition memory is a major moderator of explicit, but not implicit, attitude change. Future work should more systematically address this potential difference, especially in view of evidence provided by
Wimmer and Poldrack (2017) according to which the effects of conscious recollection may differ depending on whether learning takes place in a single session or distributed across multiple sessions over time. Related to this point, the present Paper 2 also points to the importance of distinguishing between immediate learning effects and the durability of learning effects over time. Whereas both current associative and propositional approaches are silent on this critical distinction, it can be easily accommodated in the context of a reinforcement learning perspective (see below).

Finally, although the current investigation does not directly speak to the content and format of the mental representations created as a result of evaluative conditioning (Papers 1–2), the redundancy observed between repeated evaluative pairings and evaluative statements in Paper 1 strongly speaks against the core idea of associative theories according to which experience-based learning and language-based learning create separate mental representations. Rather, in line with propositional theories, it seems that learning from stimulus pairings and from purely language-based instructions can give rise to overlapping evaluative representations that are then both activated on implicit measures of attitude.

In ongoing work with several collaborators, we are exploring further predictions derived from associative (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) versus propositional accounts of implicit evaluation (De Houwer, 2014; 2018; De Houwer & Hughes, 2016), with the bulk of initial evidence appearing to favor the latter over the former. First, in a set of developmental studies (Charlesworth, Kurdi, & Banaji, in prep), we have demonstrated that implicit attitude acquisition from evaluative statements shows adultlike levels in children aged 7 to 11; by contrast, learning from repeated evaluative pairings seems impaired in this population. As such, this finding suggests that, contrary to the position taken by associative
theories, implicit attitude acquisition via evaluative conditioning requires relatively complex relational reasoning that may emerge only relatively late in development.

Second, in a set of studies directly pitting associative and propositional information against each other (Mann, Kurdi, & Banaji, in prep), we have provided further evidence that implicit attitude acquisition in no way favors the former over the latter. Specifically, in these experiments, a novel target was paired with human screams in an evaluative conditioning paradigm, followed by diagnostic verbal information about the target’s character. Consistently across three studies, diagnostic verbal information, but not a simple instruction to mentally reverse the previously experienced stimulus pairings (Gregg et al., 2006), eliminated the effects of the prior conditioning experience. As such, in combination with Papers 1 and 2, this project provides strong evidence against any privileged role of direct experience with valenced stimuli in driving implicit evaluative learning. At the same time, given that the effectiveness of certain kinds of verbal information in shifting implicit evaluations has already been established, these studies highlight the need to establish what kinds of verbal information have impact and what kinds of verbal information do not (see in more detail below).

Third, as mentioned above, existing studies have failed to provide much direct evidence on the format and content of the mental representations underlying implicit evaluation. In a novel set of studies (Kurdi & Dunham, in prep), we probe whether implicit evaluations can be updated in response to information that can be represented only in propositional, and not in associative, form. Specifically, in these experiments, participants will receive information about the traits of fictitious targets (Laapians versus Niffians; Gregg et al., 2006) in the form of conditional sentences that is subsequently disambiguated via the presentation of a visual stimulus (e.g., “If we show you a green triangle next, you can conclude that Imolaap is sincere; if we show you a red
circle next, you can conclude that Imolaap is a liar.”). Crucially, in this design, (a) Laapians and Niffians are paired with positive and negative traits the same number of times and (b) the truth value of the yoked sentences can be appropriately updated only if the information was initially represented in propositional form. Initial evidence suggest that implicit evaluations can be updated in such a paradigm; in future work, we plan to investigate whether implicit evaluations reflect errors in propositional reasoning (e.g., denying the antecedent) in a similar setting.

A fourth line of ongoing studies seeking to provide evidence on the associative–propositional debate is being conducted within the general framework of causal learning (Kurdi, Morris, & Cushman, in prep). Specifically, in these experiments, participants are being exposed to a machine that is being activated by a stimulus (causally responsible stimulus) and then outputs a different stimulus (associated stimulus) and a reward. Notably, under associative theories, implicit evaluations should be responsive only to the covariance between stimuli and rewards and, as such, causally responsible and associated stimuli should not differ in implicit evaluation. By contrast, under propositional theories, implicit evaluations should be responsive not only to the fact that two stimuli are related to each other but also to the nature of their relationship. As such, the causally responsible stimulus should be evaluated more positively than the associated stimulus. Initial results are in line with the second hypothesis suggested by propositional theories. In ongoing work, we are investigating (a) to what extent the effects are dependent on verbal scaffolding provided prior to exposure to the learning experience and (b) whether implicit evaluations reflect well-known signatures of causal learning, including abnormal selection (A. Morris et al., 2018).

At the same time, it seems that although the associative–propositional debate has led to considerable advances in terms of our understanding of the processes that give rise to implicit
attitude acquisition and change, a novel approach with a higher level of specificity regarding (a) the posited mental computations and (b) the implementation of those computations, and, as such, a higher level of falsifiability would be desirable for further progress to be possible. Specifically, even though we interpreted Papers 1–2 to provide evidence in favor of propositional over associative theories, it should be pointed out that propositional theories are silent on a host of aspects of the present data that seem critical for a deeper understanding of implicit evaluative learning processes. Specifically, propositional theories (a) do not explain why verbal instructions lead to stronger immediate implicit attitude change than experience with stimulus pairings (Paper 1); (b) do not specify under what conditions implicit attitudes should be acquired quickly versus slowly (Paper 2, Study 1); (c) do not explain why implicit attitude acquisition from stimulus pairings should be modulated by verbal information if such information is presented before but not after exposure to stimulus pairings (Paper 2, Study 2); (d) do not explain why implicit attitude acquisition from stimulus pairings should be more durable than implicit attitude acquisition from verbal information (Paper 2; Studies 3–4); and (e) are generally unable to explain dissociations of explicit and implicit evaluations, and specifically the fact that only explicit, and not implicit evaluations, are responsive to model-based learning (Paper 3).

As explained in more detail in Paper 2, some of the shortcomings of current propositional theories could be addressed by relying on Bayesian models of cognition [especially points (a), (b), and (c)] (Tenenbaum et al., 2011) and general research on episodic memory [point (d)]. However, the propositional approach has at least three shortcomings that appear to be difficult to address in a Bayesian framework: First, the propositional theory cannot explain dissociations of explicit and implicit evaluations without unprincipled and ad-hoc assumptions about which propositions can or cannot be automatically activated on implicit measures of attitude. Second, in
a related vein, propositional theories posit that implicit evaluations should be responsive to both
direct experience with valenced stimuli and purely language-based learning. At the same time,
the propositional theory is silent on the fact that implicit evaluations are responsive to some
kinds of direct experience and not others (e.g., to rewards received as a result of binary choice
behavior but not to revaluation; Paper 3) and to some kind of verbal information (e.g., Paper 1)
but not others (Lai et al., 2014). Importantly, the kinds of direct experience and verbal infor-
mation that are effective in shifting explicit versus implicit attitudes show dissociations: For in-
stance, whereas explicit attitudes are responsive to model-based processes, implicit attitudes
seem impervious to them. Moreover, whereas explicit evaluations of African Americans are
modulated by appeals to be more egalitarian, implicit evaluations are not (Lai et al., 2014).
Third, the propositional theory does not appear to have the ability to address the conceptually
and practically important distinction between temporary malleability (Blair, 2002; Lai et al.,
2014) and long-term change (Forscher et al., 2017; Lai et al., 2016). Specifically, the proposi-
tional approach is silent on the fact that the effects of many interventions that have immediate
impact on implicit evaluations seem ephemeral when tested following a delay.

On the other hand, the reinforcement learning framework introduced in Paper 3 seems
particularly well-suited to address these shortcomings of current theories of implicit evaluation.
First, given that both explicit and implicit evaluations were found to respond to model-free learn-
ing but only explicit evaluations were found to respond to model-based learning, this finding
seems to provide an immediate explanation for why explicit and implicit evaluations are usually
related to each other but rarely redundant (Hofmann et al., 2005; Nosek, 2005). Specifically, be-
because explicit and implicit evaluations share model-free representations, they should be related to
each other; however, this relationship should become weaker as a result of explicit evaluations
incorporating increasing amounts of model-based information. This implication of the current results can be more directly tested in future work.

Second, with regard to the informational inputs to implicit evaluative learning, the present findings suggest that the crucial distinction may not concern the format in which information is presented to participants (e.g., as stimulus pairings versus verbally) but rather the extent to which participants are able to represent this information in sufficiently compressed form to access it under time pressure when completing an implicit measure of attitude such as the Implicit Association Test (Greenwald et al., 1998). A reexamination of the results of Lai et al. (2014) suggests that only interventions (a) allowing for a pure model-free strategy or (b) involving the simplest possible kind of model \( P(\text{good} | A) = P(\text{bad} | B) = 1 \) may have impact. This idea should also be more systematically investigated in future work by directly manipulating the complexity of the mental model required for the successful updating of evaluative representations. For instance, this perspective suggests that implicit evaluations should be updated in response to a verbal description of a two-armed bandit task but not in response to a verbal description of a revaluation procedure.

Third, with regard to the distinction between temporary shifts (Blair, 2002; Lai et al., 2014) and durable change (Forscher et al., 2017; Lai et al., 2016) in implicit evaluations, the reinforcement learning perspective suggests that these issues are not empirically and conceptually separate; rather, they are deeply intertwined. Specifically, in a reinforcement learning framework evaluations of stimuli are always tied to particular states of the environment, i.e., they are context-specific. As such, in this framework, generic or context-free evaluations do not exist; any new learning that takes place is attached to a particular context, be it a physical space (Wittenbrink et al., 2001), a motivational state (Fishbach & Ferguson, 2007; Moskowitz, 2014), or the
presence of an experimenter (Lowery et al., 2001). Accordingly, it should not be surprising that interventions taking place in a particular experimental context (Forscher et al., 2017; Lai et al., 2016) do not remain effective in modulating implicit evaluations, especially following a delay. In this context, the most important open question seems to be the adequate characterization of the state space (Gershman, 2018; Gershman et al., 2015) that drives the generalization of implicit evaluations across contexts. Once the state space has been adequately characterized, it may be possible to design interventions that span the most consequential regions of state space in order to drive the widest possible generalization to new situations and, as such, the highest possible degree of temporal stability.

To summarize, the reinforcement learning approach to implicit evaluation introduced in Paper 3 is already able to account for a wide range of findings regarding implicit attitudes, including their relationship with their explicit counterparts (Hofmann et al., 2005; Nosek, 2005), their responsiveness to certain interventions but not others (Lai et al., 2014), and their fundamentally context-dependent nature (Blair, 2002), including their sensitivity to motivational states (Fishbach & Ferguson, 2007; Moskowitz, 2014). Future research within the same framework will be able to offer more systematic evidence on (a) changes in explicit–implicit correlations depending on the magnitude of model-based contributions to explicit attitudes, (b) the idea that to drive responding on implicit measures of attitude, value representations must be highly compressed, and (c) the state space that determines the generalization of implicit evaluations across different contexts. In addition, future work will also be able to provide more clarity on (a) whether model-based representations might be used to train implicit evaluations offline (Gershman et al., 2017), (b) the differences between experienced and instructed reinforcement learning in shifting implicit attitudes, and (c) how reinforcement learning might interact with other
sources of learning, such as causal learning (Cheng, 1997) or mental state inferences (Baron-Cohen, Leslie, & Frith, 1985), in driving implicit evaluations.
Postscript: Outline of the Compression–Computation Theory of Implicit Evaluation

Relying on the empirical results reported in this dissertation, especially Paper 3, the goal of the present postscript is to sketch out the fundamentals of a new theory of implicit attitude acquisition and change. This theory, which I will refer to as the compression–computation theory of implicit evaluation, uses important insights from both associative (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) and propositional theories of implicit evaluation (De Houwer, 2014; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009) but departs from both in crucial ways. On the one hand, like associative theories, the compression–computation theory posits a dissociation between implicit and explicit evaluations in terms of underlying representations. On the other hand, like propositional theories, the compression–computation theory proposes that implicit evaluations can be updated on the basis of many forms of learning, including direct experience and verbal information.

The novel insight of the compression–computation theory, inspired by a reinforcement learning perspective, is that whereas explicit evaluations can emerge either from compressed (precompiled) representations of value or from online computations of value, implicit evaluations can rely only on the former but not on the latter. As such, unlike existing associative theories, the compression–computation theory posits partial, rather than full, dissociation between implicit and explicit evaluations in terms of underlying representations. Moreover, the compression–computation theory recognizes the possibility that propositional information may be compressed into propositional format. Each of these points is discussed in more detail below.

Under the compression–computation theory, implicit (automatically activated) and explicit (self-reported) attitudes are thought to differ from each other in underlying representations.
Specifically, I propose that explicit attitudes can reflect both (a) representations that are highly compressed and precompiled (such as Q values derived from model-free reinforcement learning) and (b) representations that require a considerable amount of online computation (such as dynamic programming over a causal model of the environment). By contrast, implicit evaluations must rely on highly compressed representations: Given that they are activated promptly and without much effort, they cannot emerge from representations that require extensive online computations to arrive at an estimate of value.

In this regard, the proposed theory is similar to existing dual-process theories (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) in that it posits a dissociation between explicit and implicit attitudes in terms of underlying representations. However, notably, the dissociation proposed here is not a double dissociation but rather a single dissociation: Whereas explicit evaluations are thought to reflect a mix of compressed value representations and value representations that are dynamically computed online, implicit evaluations reflect only the former but not the latter. Moreover, the dissociation proposed here is not between associative and propositional representations but rather between value representations that are pre-compiled versus computed online. As such, the present theory departs from propositional theories (De Houwer, 2014; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009) given that in these theories, the same (propositional) representations underlie both implicit and explicit evaluations. Such departure seems necessary because the propositional theory has difficulty accounting for dissociations of explicit and implicit attitudes, which are both numerous and well-documented in the literature (Hofmann et al., 2005; Nosek, 2005). Proponents of the propositional theory usually argue that such dissociations can be explained by the fact that certain propositions but not others can be activated automatically. From this perspective, the present theory
provides a more specific delineation of (possibly propositional) representations that can or cannot be subject to automatic activation.

Although the distinction between value representations that are (a) compressed and pre-compiled and (b) more rich in detail and requiring online computation is based on the more specific distinction between model-free and model-based reinforcement learning, I do not wish to argue that reinforcement learning is the only, or even the most important, path toward implicit attitude change. Such a view would be difficult to defend given that implicit attitudes have been demonstrated to also respond to mere exposure (Van Dessel et al., 2018), Pavlovian learning (Hofmann et al., 2010), approach–avoidance training (Van Dessel, De Houwer, & Gast, 2015a), and verbal instructions (present Paper 1). Rather, the basic argument of the compression–computation theory is that any kind of learning that has the ability to create compressed representations of value should be able to shift implicit evaluations; by contrast, learning resulting in more complex representations that require online computations to arrive at an estimate of value will have impact on explicit, but not implicit, evaluations.

As such, the compression–computation theory is again similar to existing dual-process theories (Rydell & McConnell, 2006; E. R. Smith & DeCoste, 2000; Strack & Deutsch, 2004) in that it posits that implicit evaluations will not respond to all learning interventions to which explicit evaluations are responsive. However, importantly, unlike existing dual-process theories, the compression–computation theory does not argue that the defining difference is between language-based and experience-based learning interventions, with the former primarily affecting explicit attitudes and the latter primarily affecting implicit attitudes. This view is extremely difficult to reconcile with empirical evidence. Implicit attitudes have been shown to respond both to (a) certain kinds of experience-based learning (e.g., experienced evaluative conditioning) and (b)
certain kinds of language-based learning (e.g., instructed evaluative conditioning; present Paper 1). At the same time, implicit attitudes have also been demonstrated not to respond to (a) certain kinds of experience-based learning, such as reward revaluation (present Paper 3) and (b) to certain kinds of language-based learning, such as the instruction to be more egalitarian (Lai et al., 2014). Rather, the compression–computation theory calls for a reexamination of these learning interventions on the basis of whether they create compressed representations of value irrespective of their procedural details (e.g., whether information is being presented verbally or nonverbally to participants). A first attempt at such reexamination of the learning manipulations featured in Lai et al. (2014) is included in the present Paper 3.

To summarize, the compression–computation theory posits that not all kinds of mental representations can underlie implicit evaluations: Specifically, in order to be able to impact implicit (automatic) evaluations, representations of value must be sufficiently compressed to be able to activated quickly and without effort. As such, unlike existing dual-process theories (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) and propositional theories (De Houwer, 2014; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009), the basic distinction drawn by the present theory is not between associative and propositional value representations but rather between value representations that are precompiled versus computed online. From the perspective of the present theory, the question of whether compressed representations of value are stored as associations (Laapians: +5, Niffians: -5) or propositions (“Laapians are good and Niffians are bad”) is both difficult to decide empirically and potentially not as consequential as previously assumed. One potential advantage of scalar value representations, as posited for instance in the Q learning algorithm (Sutton & Barto, 1998), is that they provide a “common currency” for preference and choice (D. J. Levy & Glimcher, 2012). Without such a
common scale for preferences, the benefit of compressed value representations would be dubious: Incommensurable value representations do not seem well-suited to quickly and adaptively guide choice and behavior. (For instance, if X caused a reward of +3 and Y was associated with a reward of +4, should X or Y be preferred?)

To reiterate, these scalar representations of value need not arise from reinforcement learning, or even any kind of experience-based learning. Rather, they can be created via any learning process, including purely verbal instructions, that has the ability to create highly compressed representations of value. In this context, it should be noted that the present theory, unlike traditional dual-process theories (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) is not opposed to the possibility of implicit evaluations responding to propositional information, as long as such propositional information is incorporated into a highly compressed value representation at the time of learning. For instance, in ongoing work (Kurdi, Morris, & Cushman, in prep), we have demonstrated that of two stimuli equally associated with a positive outcome, the one causally responsible for the outcome is implicitly preferred to the one merely co-occurring with it. This result clearly shows that implicit evaluations can respond to propositional information (i.e., information specifying the kind of relationship between two entities beyond mere co-occurrence); however, it leaves open the format of the representation created. Specifically, in line with the common currency idea, observations implying propositional relationships need not be represented in propositional format. For instance, the observation “X causes the machine to output pleasant objects” may translate to a higher scalar value than the observation “the machine outputs Y along with pleasant objects”, i.e., $Q_X > Q_Y$, and these relative scalar values may then be automatically activated on implicit measures of attitude such as the IAT.
This idea may be directly tested in a modified revaluation paradigm in which participants are first exposed to the same contingencies as before (i.e., one stimulus causing a machine outputting a reward and the other stimulus co-occurring with a machine outputting a reward) and are then provided with (either verbal or experiential) information about the machine being broken. If the reasoning above is accurate, unlike explicit evaluations, implicit evaluations should not be responsive to this novel information given that the highly compressed representation of value associated with the two stimuli does not contain symbols for the machine or the transitional probabilities between events and outcomes. By contrast, a second kind of manipulation providing direct experience with the diminished causal power of the previously causally efficacious stimulus should have impact given that it can directly modify the scalar value associated with the stimulus without requiring detailed representations of the entire causal process.

The compression–computation theory departs from existing dual-process accounts (Rydell & McConnell, 2006; E. R. Smith & DeCoster, 2000; Strack & Deutsch, 2004) not only in terms of the posited inputs to implicit attitude acquisition and the resulting mental representations but also in terms of the nature of the posited learning process. Specifically, existing dual-process theories claim that implicit evaluations should always be updated in a slow and piecemeal fashion. In Paper 2 of the present dissertation, we provide clear evidence against this idea, demonstrating that implicit evaluations can be updated as a result of as few as four stimulus pairings, with further pairings providing no additional benefit. By contrast, propositional theories (De Houwer, 2014; De Houwer & Hughes, 2016; C. J. Mitchell et al., 2009) allow for either quick or slow acquisition of implicit attitudes; however, they fail to specify under what conditions learning should be quick or slow. The present theory takes the position that both quick or slow learning processes should be possible and, at the same time, highlights the importance of
considering computational models of the learning processes underlying implicit attitude change in predicting the kind of learning process that should unfold.

As explained above, the compression–computation theory posits that implicit evaluations should reflect the output of any kind of learning process provided that that learning process creates compressed representations of value. Computational models of several such processes, including model-free reinforcement learning (Sutton & Barto, 1998) and causal learning (Gopnik et al., 2004), exist, and the compression–computation theory has the ability to incorporate such models. For instance, in the context of the reinforcement learning manipulation reported in Paper 3, a standard Q learning algorithm predicts that learning should asymptote faster in a setting where reinforcement is deterministic (Studies 1–2) rather than probabilistic (Study 3). In Paper 2, we sketched the outlines of a Bayesian model of evaluative conditioning that has the ability to take into account how verbal instructions structuring the hypothesis space entertained by participants and actual experience with stimulus pairings may interact with each other in producing implicit attitude change. Specifically, the more highly structured the hypothesis space, the less experiential learning should be necessary for learning to asymptote.

More generally, computational models of learning make it clear that whether learning is fast or slow should be a function of the particulars of the learning situation and not whether the output of learning is measured via self-report (explicitly) or via response interference tasks (implicitly). Beyond making quantitative predictions about the unfolding of learning processes, computational models may be beneficial for the study of implicit attitude acquisition in additional ways, for instance by providing individual-level parameters that may be used in theory testing. As such, in future work, it should be possible to correlate (changes in) individual-level Q value estimates with (changes in) individual-level implicit evaluations.
Beyond considering existing evidence, the compression–computation theory may also be submitted to further scrutiny by testing some of its specific predictions in future empirical work. First, a central prediction of the theory is that implicit evaluations should generally be immune to revaluation manipulations, whether in the context of reinforcement learning or other learning paradigms such as causal learning or Pavlovian learning. Above I describe a potential paradigm for testing this idea in the context of causal learning. In the context of evaluative conditioning, the same idea may be tested in a three-stage design: In stage 1, two initially neutral categories will be conditioned to become either positive or negative (A+ and B-). In stage 2, these two categories will be associated with a pair of novel, also initially neutral categories (AC and BD). Finally, in stage 3, the first set of categories will undergo counterconditioning (A- and B+). The prediction of the theory is that implicit evaluations of C and D should reflect second-order conditioning (stage 2) but not revaluation (stage 3) given that the mental representations underlying implicit evaluations encode only the fact that C and D became valenced but do not encode how they became valenced (via their pairings with A and B; but see Gershman, 2015). By contrast, explicit evaluations rely on a detailed model of the conditioning process and, as such, should be responsive to the revaluation manipulation.

Second, in Paper 3, the revaluation manipulation was created to result in divergent model-free and model-based value representations. However, as demonstrated by recent investigations (Gershman et al., 2014; 2017), a model of the environment can be used to train model-free value representations offline. As mentioned multiple times above, the compression–computation theory of implicit evaluation posits that any kind of learning resulting in compressed representations of value should be able to shift implicit attitudes. As such, unlike in Paper 3, implicit evaluations should be able to arise from model-based learning provided that sufficient time elapses.
between learning and testing in order for offline retraining to occur. This idea may be tested in a paradigm similar to that used in Paper 3, but with a time delay implemented between the learning and test phases of the experiment. During this delay, participants may be asked to engage in one of the following three activities: (a) a somewhat but not excessively taxing secondary task, (b) listening to music, or (c) performing instructed mental simulation of the novel action–outcome contingencies following revaluation.

Third, from the perspective of the compression–computation theory of implicit evaluation, the distinction between experience-based and language-based interventions is largely irrelevant: As long as a manipulation is able to create compressed representations of value, it should have impact on implicit evaluations. In line with this idea, in Paper 1 of the present dissertation we demonstrated that instructed evaluative conditioning is at least as effective as, if not more so than, experienced evaluative conditioning in shifting implicit attitudes. However, as discussed above, certain kinds of language-based interventions, such as the instruction to be more egalitarian, do not have impact (Lai et al., 2014). In a follow-up study, an initial theoretically motivated attempt may be made at teasing apart what kinds of language-based intervention are effective or ineffective in shifting implicit attitudes. Specifically, under the compression–computation theory, if experienced action–outcome contingencies were replaced by instructed action–outcome contingencies, the results should remain unchanged. That is, implicit attitudes should respond to both experienced and instructed reinforcement learning. Similarly, just as experienced revaluation, instructed revaluation should also have impact exclusively on explicit and not on implicit evaluations.

Finally, as is well-established by decades of research in social psychology, mental state inferences are often made automatically and on the basis of minimal information (Heider &
Simmel, 1944). If the compression–computation theory is accurate, mental state inferences made at the time of learning should be incorporated into implicit evaluations of human targets given that such inferences can contribute to a compressed representation of value. Specifically, mentalistic inferences about intention may override or at least be integrated with representations of reward in acquiring implicit evaluations of agents that are perceived as having a mind. This idea may be tested in a framework where a novel agent has inaccurate information about a participant’s preferences (e.g., they erroneously assume that the participant prefers Coke Zero to Diet Coke). From a model-free reinforcement learning perspective, this agent should be evaluated negatively given that they are associated with an outcome that is negative for the perceiver. However, if the present theory is accurate, and any information that can be incorporated into compressed representations of value at the time of learning can influence implicit evaluations, implicit evaluations of the well-meaning agent should be (at least to some degree) positive.

In addition to generating novel predictions, such as the ones outlined above, the compression–computation theory also has implications for longstanding issues in the study of implicit evaluations. Specifically, as discussed in more detail in Paper 3, the theory speaks to issues such as (a) the relationship between implicit and explicit attitudes, (b) the goal-dependent nature of implicit attitudes, and (c) the context-specific nature of implicit attitudes.

First, given that the representations underlying implicit and explicit evaluations are posited to be partially, but not fully, overlapping, implicit and explicit evaluations are predicted to be related to each other but not redundant. Empirical evidence in favor of this idea is ubiquitous (Hofmann et al., 2005; Nosek, 2005).

Second, in line with the compression–computation theory, implicit evaluations have been shown to be sensitive to goals in two contexts (Ferguson & Bargh, 2004; Moskowitz, 2014): (a)
when what has usually been referred to as a goal is better characterized as a motivational state (such as thirst, hunger, or nicotine deprivation) and (b) among individuals who have protracted experience pursuing a certain goal, presumably resulting in a shift away from initially effortful online computations toward highly compressed representations of value. This idea is also discussed in more detail in Paper 3.

Third, implicit evaluations have been shown to be highly context-dependent (Blair, 2002). This characteristic of implicit evaluations, again, follows directly from the definition of model-free learning algorithms under which representations of value are attached to different states of the environment. A challenge for future work is to explore how our highly complex (social) environments are represented in such a way as to be able to guide automatic responding. It seems obvious that an exhaustive and unstructured list of states, as often used in toy Q learning problems, would be highly inefficient and intractable. However, either feature-based representations (Gershman, 2018) or latent cause representations (Gershman et al., 2015) may provide viable alternatives.

Beyond these issues described in some depth in Paper 3, the compression–computation theory also speaks to two fundamental issues regarding the nature of implicit evaluations. First, implicit evaluations are often described as reflecting the heuristic or irrational part of the human mind; in fact, both in scientific and popular communication, (some) implicit attitudes are often referred to as implicit biases. The present perspective, in combination with a host of empirical findings, including the papers forming the basis of the present dissertation, seems to challenge or at least constrain the validity of this claim. Specifically, it appears that seemingly irrational implicit biases emerge, at least in part, not because human minds operate in a biased way but rather because the input to which humans are exposed is biased. For instance, in a recent paper we have
demonstrated that in a vast repository of online text, some group labels (such as “professional” or “educated”) appear systematically closer to positive than to negative terms, whereas other group labels (such as “homeless” or “unemployed”) appear systematically closer to negative than to positive terms (Kurdi, Mann, Charlesworth, & Banaji, 2019). Given inputs of this kind to evaluative learning, it should not surprise if implicit evaluations deviate from neutrality even if the cognitive machinery used to derive implicit evaluations from such inputs operates in a fully rational way.

Second, the human mind is often described as consisting of multiple systems with their distinct modes of operation: For instance, system 1 is posited to be fast, parallel, automatic, effortless, associative, slow-learning, and emotional, whereas system 2 is posited to be slow, serial, controlled, effortful, rule-governed, flexible, and neutral (Kahneman, 2003). A serious issue with characterizations of this kind is that such bundles of features are neither logically coherent nor empirically correlated (Melnikoff & Bargh, 2018). For instance, a process can be associative and fast or effortful and emotional. In this context, it is crucial to distinguish central features of a system or process from incidental or secondary features. With regard to the distinction between explicit and implicit evaluations, the compression–computation theory suggests that the crucial difference has to do with the degree of compression of the underlying mental representations. Obviously, this difference may have consequences for other features of implicit vs. explicit attitudes; however, these are better seen as secondary.

The compression–computation theory may also provide some insight into why implicit evaluations sometimes seem sticky, especially relative to their explicit counterparts. For instance, in a paper by Gregg, Seibt, and Banaji (2006), participants were able to acquire novel implicit evaluations of fictitious groups based on an elaborate manipulation describing one of the groups
as savage and aggressive and the other group as docile and peaceful. However, when participants were informed that the group labels had been erroneously swapped by the computer, implicit evaluations of the two groups remained unchanged. Under the compression–computation theory, such stickiness is understandable: Considering that the initial manipulation gave rise to highly compressed representations of value (possibly of the form Laapian–positive, Niffian–negative), subsequent information that referred to the details of the manipulation for which these representations did not contain any symbols could not have had impact.

By contrast, the compression–computation theory has considerable difficulty explaining one set of findings where implicit attitude change relies on reinterpretation of the specific details of previously acquired episodic memories (Mann & Ferguson, 2015; 2017). In this paradigm, participants learn that a novel target called Francis West ransacked the home of his neighbors, destroyed their property, and took precious things from the bedrooms. In the reinterpretation condition, participants are then presented with the additional information that Francis broke into the neighbors’ home because he saw that it was on fire, and the “precious things” that he removed from the bedrooms were children. Initially negative implicit evaluations of the target were reversed in this condition but were merely mitigated in a condition in which information unrelated to the breaking and entering story was presented. Under the compression–computation theory, the reinterpretation manipulation should not have had impact (or at least no more impact than unrelated positive information) given that implicit evaluations are posited to emerge from highly compressed value representations and, as such, are not expected to be able to access narrative details of the initial story.

On a related note, the compression–computation theory does not readily explain why the relearning condition of Study 1 in Paper 3 did not produce complete reversal of implicit evalua-
tions, as may be expected on the traditional formulation of model-free learning algorithms that privilege recent over more temporally distant information (Sutton & Barto, 1998). Similarly, the theory does not offer an immediate explanation for why implicit evaluations created via instructions are less durable than implicit evaluations created via exposure to stimulus pairings (Studies 3–4, Paper 2) given that both of these manipulations are posited to give rise to the same, highly compressed, value representations. Finally, the finding that explicit but not implicit evaluations are able to shift in the face of revaluation manipulations (Paper 3) raises the intriguing idea that the same learning may simultaneously give rise to two representations: (a) a detailed causal model of the environment and (b) highly compressed scalar values associated with stimuli. Such redundancy seems wasteful and therefore requires a compelling justification whose details should be worked out in the future.
References


Baggett, P., & Ehrenfeucht, A. (1983). Encoding and retaining information in the visuals and ver-
Ballew, C. C., II, & Todorov, A. T. (2007). Predicting political elections from rapid and unreflec-
tive face judgments. *Proceedings of the National Academy of Sciences, 104*(46), 17948–
17953. http://doi.org/10.1073/pnas.0705435104
social beliefs. In D. L. Schacter & E. Scarry (Eds.), *Memory, Brain, and Belief* (pp. 139–
0410-6
and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3–51). New
York, NY.
Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stere-
type effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology*
(pp. 361–382). Guilford Press.


http://doi.org/10.1080/15248372.2012.742902


Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message


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). Elsevier Inc. http://doi.org/10.1016/bs.aesp.2017.03.001


De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and per-


Devos, T., Gavin, K., & Quintana, F. J. (2010). Say “adios” to the American dream? The interplay between ethnic and national identity among Latino and Caucasian Americans. *Cultural Di-


http://doi.org/10.1037/h0057856


http://doi.org/10.1080/00952990.2016.1229324

http://doi.org/10.1037//0022-3514.81.5.842

http://doi.org/10.1111/cogs.12236


210


Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, A. Nowak, & S. J. Read (Eds.), *Computational models in social psychology*.


