DO WE TRUST OUR GUT?

THE CAUSAL RELATIONSHIP BETWEEN IMPLICIT GROUP ATTITUDES AND BEHAVIOR

By

KATHERINE J. WOLSIEFER

B.A., Bellarmine University, 2009

M. A., University of Colorado Boulder, 2014

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Psychology and Neuroscience

2017

This thesis entitled:
Do We Trust Our Gut? The Causal Relationship Between Implicit Group Attitudes and Behavior
written by Katherine Wolsiefer
has been approved for the Department of Psychology and Neuroscience

_____

Dr. Irene Blair (Chair)

_____

Dr. Charles Judd

_____

Dr. Christopher Loersch

_____

Dr. Lewis Harvey

_____

Dr. Lawrence Williams

Date: April 25, 2017

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline

IRB protocol # 14-0042

Wolsiefer, Katherine J. (Ph.D., Psychology and Neuroscience)

Do We Trust Our Gut? The Causal Relationship Between Implicit Group Attitudes and Behavior

Thesis directed by Professor Irene V. Blair

Considerable evidence suggests that implicit attitudes co-vary with behavior (Greenwald, Poehlman, Uhlmann & Banaji, 2009). Within the domain of stereotyping and prejudice, in particular, implicit group attitudes have been shown to correlate with behavior towards individual group members. Notably, little experimental evidence demonstrates that implicit group attitudes *cause* behavior towards individual group members. In five experiments, I created (Experiments 1, 3, 4, & 5) or manipulated (Experiment 2) implicit attitudes, and measured these attitudes as well as behavior towards individual group members. Although an evaluative conditioning procedure reliably affected implicit attitudes, it did not have any impact on behavior by itself (Experiments 1 & 2). The addition of a narrative vignette to the manipulation increased condition differences in implicit attitudes (Experiment 3) and impacted behavior (Experiments 4 & 5). However, multiple mediation analysis revealed conflicting evidence regarding the roles of implicit and explicit attitudes in affecting behavior. In Experiment 4, implicit but not explicit attitudes mediated condition difference in behavior; in Experiment 5, explicit but not implicit attitudes mediated condition differences in behavior. This suggests that any causal relationship between implicit group attitudes and individual level behavior may be smaller and more tenuous than previously assumed.

Keywords: implicit attitudes, causal inference

DEDICATION

I would like to dedicate this dissertation to my parents. They have always encouraged me to pursue my goals, even when they didn't understand why I would want to pursue them. I would not have completed this work without the lessons they have taught me and without their unwavering love and support.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the continued support from the faculty members in our department. In particular, I would like to thank my advisor, Dr. Irene Blair, for her careful mentorship and guidance throughout graduate school. Throughout this process, Irene has been willing to provide feedback at every stage and has provided invaluable suggestions, insights, and moral support. I would also like to thank Dr. Chick Judd for providing years of support and training in statistics and for being willing to answer questions no matter how large or small.

Finally, I would like to thank the research assistants who worked with me on this project: Roxanne Ross, Sierra Barnes, Zach Lee, and Preston Godkin. These students are all highly driven and intelligent people without whom material creation and data collection would have been impossible, thank you!

CONTENTS

TABLES

Table

FIGURES

Figure

CHAPTER I: General Introduction

I think implicit bias is a problem for everyone...

- Hilary Clinton

Ms. Clinton's statement during the first 2016 presidential debate illustrates popular interest in implicit attitudes and their role in behavior (Blake, 2009). In the past several years, news outlets have been quick to point to implicit bias, relatively uncontrollable associative biases, as the cause of behaviors ranging from police shootings of unarmed Black men (Cummins, 2016) to hiring decisions (Wall Street Journal, 2017). These popular accounts coincide with the application of dual process theories of attitudes to explain broad evidence of racial discrimination in employment, housing, credit markets, and incarceration (Bertrand & Mullainathan, 2004; Pager and Shepherd, 2012; Spohn & Holleran, 2000; Turner, Ross, Galster, & Yinger, 2002; F. D. Wilson, Tienda, & Wu, 1995), despite dramatic decreases in reported prejudice over the past several decades (Marsden, 2012; Schuman, Steeh, Bobo, & Krysan, 1997)[1]. An oft-cited explanation for disparate outcomes in the absence of explicit prejudice is that bias now operates in subtler or more automatic ways (Dovidio, Gaertner, Kawakami & Hodson, 2002; Rudman, 2004; Smith & Levinson, 2011; Ziegert & Hanges, 2005). That suggests that implicit attitudes should be particularly helpful at explaining behaviors that explicit attitudes do not, either because individuals do not want to self-report their attitudes towards groups (e.g. people may be reluctant to admit that they view one racial group more positively than another) or because the behavior is driven by more automatic processes that are not captured by explicit attitude measures.

[1] For example, from 1960 to 1995 the percentage of White Americans who claimed they would vote for a Black president rose from 50% to 95% (Gallup; Shuman, et al., 1997, p. 106-107).

In support of this argument, researchers have amassed substantial evidence that implicit bias against outgroups (particularly those that are socially stigmatized) is more the norm than the exception (Nosek, Banaji & Greenwald, 2002). For example, a large online sample completed implicit attitude measures for many social category dimensions and revealed implicit biases that favored White over Black individuals, straight over gay individuals and thin over heavy people (Nosek et al, 2007). Such evidence suggests that all forms of bias have not been eliminated and the high rates of implicit bias are consistent with findings of societal-level discriminatory outcomes across domains. Furthermore, a number of studies have shown that on an individual-level, those with higher levels of implicit bias are the same individuals who show higher levels of biased behavior (e.g. Agerström & Rooth, 2011; Bessenoff & Sherman, 2000; Dovidio, Kawakami & Gaertner, 2002; Ziegert & Hanges, 2005). These data are often cited in support of the conclusion that implicit attitudes are at least partially responsible for discriminatory outcomes (Dasgupta & Greenwald, 2001; Devine, Forscher, Austin & Cox, 2012; Dovidio & Gaertner, 2010; Fiske & Molm, 2010), and thus efforts to change implicit attitudes are necessary if one wishes to address discrimination in today's society (Dasgupta & Greenwald, 2001; Devine et al., 2012).

The goal of this dissertation is to examine the strength and validity of the claim that implicit group attitudes (implicit biases) cause discriminatory behavior towards group members. I focus on group attitudes and behavior for two reasons. First, a considerable amount of the literature that focuses on implicit attitude-behavior relations is in the domain of stereotypes and prejudice. Implicit attitudes are thought to be especially useful at predicting behavior in this domain because self-presentational concerns may prevent individuals from self-reporting intergroup biases (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). Thus, implicit

attitudes may be particularly likely to act as a unique cause of behavior (over and above explicit attitudes) within this domain. The group context is also unique because the implicit group attitude to individual level behavior relation is quite commonly studied and the lack of correspondence between the level of the (group) attitude being assessed and the (individual) behavior may make detecting such a relationship more challenging (Azjen & Fishbein, 1977). Second, there is no evidence, to my knowledge, that supports the assumption that implicit group attitudes cause behavior towards group members. I begin by examining the existing evidence on the implicit attitude to behavior relationship (with a focus on the domain of stereotypes and prejudice). Then I present 5 experiments and a re-analysis that explore whether and when implicit group attitudes are likely to cause behavior.

Dual-Process Models

Models of implicit group attitudes are typically situated in the context of dual process theories, which suggest that two types of processes influence behavior: one fast, associative and efficient; the other slow, deliberative, and effortful (e.g. Olson & Fazio, 2009; Wilson, Lindsey, & Schooler, 2000). These dual process theories suggest that attitudes can operate through either the more associative or more deliberative route, and can result in either greater (in the case of the associative route) or less (in the case of the deliberative route) reliance on associative content. Whereas implicit attitudes are thought to represent traces of past experience or strength of associations and map onto faster, associative processes; explicit attitudes are thought to represent personal beliefs and (Devine, 1989) and likely influenced by self-presentational concerns raised by more deliberative processes (Fazio & Towles-Schwen, 1999; Gawronski & Bodenhausen, 2006). Additionally, although implicit attitudes are thought to be automatically activated and operate on behavior relatively unconsciously (Greenwald & Banaji, 1995, Fazio & Towles-

Schwen, 1999; Dovidio, et al., 2002), explicit attitudes are thought to be effortful and to affect behavior through more deliberative processes (Fazio & Towles-Schwen, 1999; Dovidio, et al., 2002). Consistent with the theory that implicit and explicit attitudes originate from or operate through different processes, correlations between implicit and explicit attitudes towards the same targets tend to be small but reliable and depend on a number of factors[2] (Hofmann et al., 2005; Nosek et al., 2007).

Dual process models concerning the attitude-behavior relationship differ in the ways they suggest that implicit and explicit attitudes may influence behavior. Some accounts of the attitude-behavior relationship suggest that implicit and explicit attitudes may be activated and operate simultaneously on behavior, producing an additive effect (Greenwald & Banaji, 1995; Wilson, et al., 2000) whereas others suggest that implicit and explicit attitudes may act in competition with one more likely to influence behavior under certain circumstances and the other more likely to influence behavior under other circumstances (e.g. Fazio & Towles-Schwen, 1999). Regardless of these specific differences, two important themes exist across dual process models. First, most of these models suggest that implicit attitudes uniquely contribute to behavior over and above explicit attitudes. That is, implicit attitudes capture variability in evaluation that is not detected by self-report measures. Second, most dual process models suggest that situational factors can increase reliance on automatic vs. controlled processes. In situations for which more automatic processing is likely, for example, when individuals are under cognitive load or time pressure, implicit attitudes are thought to more strongly influence

[2] One of these factors is the social sensitivity of the domain in which implicit and explicit attitudes are assessed. Implicit-explicit correlations are considerably lower in domains for which individuals are uncomfortable self-reporting attitudes such as, stereotyping and prejudice, and much higher in domains low in social sensitivity (e.g. politics; Greenwald, et al. 2009).

behavior. In situations for which more controlled processing is likely, such as situations in which an individual has sufficient time and motivation to think through their actions, explicit attitudes are thought to more strongly influence behavior. The situational factors that impact process reliance are thought to do so in two ways. Some situational factors decrease individuals' opportunity to control their behavioral responses, resulting in greater reliance on automatic processes. Other situational factors decrease individuals' motivation to control their behavioral responses, which is also thought to increase reliance on automatic processes (Bless & Schwarz, 1999; Fazio & Towles-Schwen, 1999; Smith & DeCoster, 2000).

Associations Between Implicit Group Bias and Behavior Towards Individuals

In this section I present correlational evidence that suggests that implicit attitudes and behavior co-vary, that implicit attitudes co-vary with behavior over and above explicit attitudes, and that implicit attitudes are more likely to co-vary with behavior when controlled processing is low. Evidence from over 350 studies of the correlational relationship between implicit attitudes and behavior (across many attitude domains) suggests that implicit attitudes co-vary with behavioral outcomes. Three meta-analyses[3] demonstrate small, but statistically significant, implicit attitude-outcome relationships ranging from $r = 0.14$ to $r = 0.28$ (Cameron, Brown-Iannuzzi, & Payne, 2012; Greenwald, et al., 2009; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013)[4]. These meta-analyses indicate that, on average across studies, implicit attitudes

[3] These meta-analyses do not distinguish between studies that correspond with regard to the specificity of the attitude and behavioral targets so it is not possible to know what the average effect size is for the implicit group attitude to individual-level behavior relationship.

[4] These effect sizes are across domains and various theoretical and methodological moderators. Greenwald et al. (2009) do report correspondence (i.e. the degree to which the target of the implicit attitude measure and the target of the behavior correspond) as a significant theoretical moderator in their analysis. They find that implicit attitude-behavior relations are stronger when there is a higher level of correspondence between the implicit attitude measure and the outcome. Although this is not exactly the same as the specificity of the targets of the attitude

are related to behavior and that they are especially predictive of behavior that may be considered socially sensitive (Cameron, et al., 2012; Greenwald et al., 2009; Oswald, et al., 2013). Two of these meta-analyses (the third did not report a test of significance for this effect) also demonstrated that implicit attitudes, on average, offered incremental predictive validity over and above the explanatory power of explicit attitudes (Cameron, et al., 2012; Greenwald, et al., 2009). More specifically, Implicit group attitudes are related to many different types of outcomes, including monetary allocations to individuals (Stanley, Sokol-Hessner, Banaji, & Phelps, 2011; Stepanikova, Triplett, & Simpson, 2011), decisions to vote for a Black candidate (Greenwald, Smith, Sriram, Bar-Anan, & Nosek, 2009; Payne, Krosnick, Pasek, Lelkes, Akhtar, & Tompson., 2010), self-reported verbal discriminatory behavior (Rudman & Ashmore, 2007), medical doctors' hypothetical treatment decisions of a Black (compared to White) patient (Green, Carney, Pallin, Ngo, Raymond, Iezzoni, & Banaji, 2007), school performance of minority students[5] (Jacoby-Senghor, Sinclair & Shelton, 2016; Peterson, Rubie-Davies, Osborne & Sibley, 2016; van den Bergh et al., 2010), and callbacks of obese individuals for a job (Agerström & Rooth, 2011). Notably, all of the above-mentioned studies find evidence of significant partial correlations of implicit attitudes with behavior, over and above any effects of explicit attitudes. Such evidence supports the meta-analytic conclusion that implicit attitudes offer additional explanatory power and are not simply accounting for the same variance in behavior as explicit attitude measures.

---

and behavior, it does suggest that measures of attitudes at a group level may not strongly predict behavior at the individual level.

[5] Note that the studies that measured teacher implicit bias and student achievement measured both implicit attitudes and behavior at the group level.

Some evidence suggests that implicit attitudes are more likely to relate to behavior when the opportunity or motivation to control behavior is low[6]. Several studies demonstrate that the implicit attitude-behavior relationship is moderated by opportunity to control behavior by establishing that implicit attitudes are related to non-verbal behavior, which is thought to be less controllable than verbal behavior (Amodio & Devine, 2006; Bessenoff & Sherman, 2000; Dovidio, et al., 2002; Gonsalkorale, von Hippel, Sherman, & Klauer, 2009; Hofmann, Gshwendner, Castelli, & Schmitt, 2008; Neumann, Hülsenbeck, & Seibt., 2004; Olson & Fazio, 2007). For example, Dovidio and colleagues (2002) measured implicit and explicit racial attitudes and had participants interact with both Black and White confederates. They found that implicit (but not explicit) attitudes were significantly related to racial bias in non-verbal behavior, and that explicit (but not implicit) attitudes were significantly related to verbal behavior.

Some work also provides evidence that motivation-related factors moderate the implicit attitude-behavior relationship. Individuals low in Need for Cognition (Florack, Scarabis & Bless, 2001), low in motivation to control prejudice (Gabriel, Banse, & Hug, 2007; Olson & Fazio, 2004), and who perceive less within group variability (Lambert, Payne, Ramsey, & Shaffer, 2005) all tend to exhibit a stronger relationship between implicit attitudes and behavior. These moderating effects impacted the relationship between implicit attitudes and trait ratings of individual targets (Florack et al., 2001; Lambert et al., 2005; Olson & Fazio, 2004) and monetary

[6] Notably, although 15-20 studies demonstrated significant moderation of the implicit attitude-behavior relationship by motivation- or opportunity-related moderators, only one meta-analysis finds any evidence that such conceptual moderators are effective on average across studies (Cameron et al., 2012). Other studies either do not test opportunity and motivation related moderators or they test controllability of the outcome (which could be considered an opportunity-related moderator) and do not find evidence that implicit attitude-behavior correlations depend on this factor (Cameron et al., 2012; Greenwald et al, 2009).

allocations to stigmatized groups[7] (Gabriel, et al., 2007). Evidence of motivation-related moderating effects also includes situational moderators. Evidence suggests that the implicit attitude-behavior relationship is stronger when exhibiting bias is more socially normative (Pryor, Reeder, Wesselmann, Williams, & Wirth, 2013; Ziegert & Hanges, 2005).

It should be noted that not all evidence is in favor of a reliable relationship between implicit attitudes and behavior. Carlsson and Agerström (2016) conducted a meta-analysis which examined the relationship between implicit attitudes (as measured by the IAT) and discriminatory behavior using a stringent definition of discrimination and a subset of studies from the Oswald et al. (2013) meta-analysis (k = 13). Specifically, to be included by Carlsson and Agerström (2016), the behavioral measure had to be one of relative discrimination (i.e. relative treatment of a minority group member compared to a majority group member) and the discrimination measure had to yield evidence of a main effect of discrimination. Using this definition, the authors found no evidence of a relationship that was reliably different from zero (r = 0.03). Notably, when the researchers relaxed their criterion requiring an overall main effect of discrimination, the average effect size mirrored that of Oswald et al. (2013) and was statistically significant.

In sum, a wide variety of behaviors appears to be correlated with implicit attitudes. These behaviors range from trait ratings to non-verbal behavior to academic performance of minority students. Both a large number of individual studies and two meta-analyses also support the conclusion that implicit attitudes provide incremental predictive capability over-and-above explicit attitudes (at least in particularly socially sensitive domains) and several moderator

[7] The authors report both an overall relationship between implicit attitudes and behavior and report that this effect is stronger for individuals low in Motivation to Control Prejudiced Responding.

studies support the idea that implicit attitudes and behavior are more likely to co-vary when motivation and opportunity to control behavior are low. Importantly, not all evidence is in favor of covariance. The small average effect sizes reported in the existing meta-analyses in addition to a more recent meta-analysis that finds no correlation between implicit attitudes and behavior when discriminatory behavior is more carefully defined leave some room for doubt as to whether implicit group attitudes are robustly related to behavior.

Even if the correlational evidence perfectly supported the idea that implicit bias is related to biased behavior, covariance is not sufficient for establishing a causal relationship. To strongly establish causality, an experiment must demonstrate that individuals randomly assigned to different levels of implicit attitudes demonstrate differential behavior and that these differences in behavior can be explained by differences in implicit attitudes. A small number of experiments that meet some of these criteria have examined the relationship between implicit attitudes and behavior.

Experimental Evidence

The studies cited in the preceding section all measure implicit attitudes and measure behavior. Although some studies measure implicit attitudes and prejudiced behavior with the correct temporal ordering, they lack random assignment. Despite substantial evidence that prejudicial implicit attitudes are related to behavior, there is no evidence that prejudiced implicit attitudes cause behavior.

A small number of studies do use experimental designs that can be used to examine the role of implicit bias in discriminatory behavior: each study used random assignment and manipulated implicit attitudes, each one measured implicit attitudes, and each one measured a behavioral outcome expected to have been influenced by implicit attitudes (Dasgupta & Rivera,

2008; Gapinski, Schwartz, & Brownell, 2006; Kawakami, Dovidio & van Kamp, 2005; Mann & Kawakami, 2012; Yoshida, Peach, Zanna, & Spencer, 2012; Rudman & Lee, 2002; Saleem & Anderson, 2013; Zogmaister, Arcuri, Castelli, & Smith, 2008). Unfortunately, each study has limitations that prevent it from offering strong causal evidence for the relation between implicit bias and behavior[8] (see Table 1 for a summary of these experiments and their limitations in terms of establishing causation).

Table 1

Summary of Implicit Attitude-Behavior Experiments

Citation	Description	Limitation
Dasgupta & Rivera, 2008	Exposure to positive gay and lesbian exemplars leads to reduced implicit anti-gay and lesbian bias and increased support of lesbian and gay civil rights legislation.	Test of whether implicit attitudes mediated the effect of the intervention on voting behavior was non-significant.
Gapinski et al., 2006	Exposure to positive (versus negative) media portrayals of obese individuals did not lead to reductions in anti-fat implicit bias. Exposure to negative media portrayals of obese individuals resulted in greater preference for living with an overweight roommate.	Manipulation did not impact implicit attitudes.
Kawakami et al., 2005	Counterstereotypic training led to less gender discrimination (i.e. less choosing of males over females for a job).	No measurement of implicit attitudes.

[8] It is important to note that this statement is not meant as a serious criticism of the research presented in this section. None of the studies presented here state establishing a causal relationship as the primary goal of their research.

Mann & Kawakami, 2012	Feedback that (White) participants were becoming more egalitarian led to greater seating distance from a Black interaction partner (Studies 1 & 3) and greater levels of implicit racial bias (Studies 1 & 2).	No evidence that implicit attitudes were related to seating distance and no test of mediation.
Yoshida et al., 2012	Participants who learned that an audience responded favorably to an anti-Muslim joke demonstrated greater anti-Muslim implicit normative bias and greater willingness to cut money from a Muslim student group budget.	The manipulation impacted implicit normative evaluations but not implicit attitudes. Normative evaluations, but not implicit attitudes mediated the effect of the manipulation on the discrimination outcome measure.
Rudman & Lee, 2002	Participants primed with violent rap music demonstrate greater anti-Black implicit bias and greater bias in trait ratings of a Black target (Study 1). Participants primed with violent rap music also show greater racial bias in ratings of ambiguous behavior of a Black (versus White) target.	Mediation is not tested (Study 1). Implicit attitudes are measured before manipulation so mediation cannot be tested (Study 2)
Saleem & Anderson, 2013	Participants who played a video game that portrayed Arab terrorists demonstrated greater anti-Muslim implicit attitudes and drew more stereotypic pictures of Muslim people.	No test of mediation. Outcome is stereotypic drawing of a Muslim person.
Zogmaister et al., 2008	Priming equality reduces implicit anti-Muslim bias and reduces seating distance of participants from an ostensible Muslim interaction partner.	Implicit anti-Muslim bias and seating distance were unrelated. Mediation was not tested.

Note. Contains citations for experiments that attempt to manipulate implicit attitudes and measure behavior. The second column provides a brief summary of relevant results and the third column identifies limitations in detecting a causal relationship between implicit group attitudes and behavior.

In perhaps the best example of a test of the causal relationship between implicit group attitudes and behavior, Saleem and Anderson (2013) randomly assigned participants to play either a violent video game portraying Arab terrorists, a violent video game portraying Russian terrorists or a non-violent control video game. Participants who played a violent anti-terrorist video game displayed stronger implicit and explicit anti-Muslim attitudes and drew pictures of Muslim people with more stereotypic traits, weapons and displaying negative affect (compared to drawings of White people). The authors also found that implicit Muslim attitudes were related to drawing characteristics. However, the authors did not report the test of the full mediational path controlling for explicit attitudes. Thus, the study does not provide evidence that implicit attitudes uniquely explained the effect of the video game manipulation on drawings. The behavioral task in this measure, while interesting, is also not the strongest measure of behavior towards individual group members as participants were simply asked to draw a picture of a White and an Arab-Muslim person. This is quite different from the measures of behavior usually predicted by performance on implicit attitude measures (e.g. seating distance, non-verbal behavior, etc.).

Two additional papers provided similarly strong tests of the causal relationship between implicit attitudes and behavior. Across three studies, Mann & Kawakami (2012) demonstrated that feedback suggesting the participant was becoming more positive towards Blacks led to increases in implicit racial bias and increases in seating distance from a Black (compared to White) interaction partner. However, in two studies, the authors did not find a significant link

between implicit bias and behavior (this link was not measured in the third study) and they did not examine the role of implicit attitudes in explaining condition differences in behavior. In a study of bias towards gay and lesbian people, Dasgupta and Rivera (2008) found that exposure to positive exemplars from these groups led to more positive implicit associations and led individuals to be more likely to report voting intentions in favor of pro-gay policies. Although the authors tested mediation, they did not find any evidence of attenuation of condition differences in behavior when controlling for implicit attitudes.

Other studies attempted to experimentally manipulate implicit attitudes and measure effects on behavior, but fall short of establishing a causal relationship for one or more reasons. Some studies did not measure implicit attitudes and cannot demonstrate that implicit attitudes were a unique cause of behavior (Kawakami et al., 2005; Yoshida et al., 2012; Saleem & Anderson, 2013). Other studies either did not test whether implicit attitudes mediated condition differences in behavior (Gapinski et al., 2006; Kawakami, et al., 2005; Rudman & Lee, 2002; Saleem & Anderson, 2013; Zogmaister, et al., 2008) or tested for this mediating effect and did not find it (Dasgupta & Rivera, 2008; Yoshida et al., 2010) and one study reports that the manipulation did not have a significant effect on implicit attitudes (Gapinski et al., 2006).

In sum, there are a small number of studies that used the experimental designs necessary for making causal claims about the relationship between implicit attitudes and behavior. However, these studies either do not measure behavior towards targets of the attitudes, do not measure all the relationships needed to establish significant mediation, or do not find evidence of the links needed to establish mediation. Even though such a causal relationship is yet to be demonstrated, a common implication from the literature on implicit prejudice is that implicit attitudes cause prejudiced behavior.

The Present Research

In the present research, I present 5 experiments and a reanalysis of existing data that examine whether implicit group attitudes cause behavior towards individual group members. In these experiments, I manipulate implicit group attitudes, measure implicit attitudes and measure behavior using multiple manipulations of implicit attitudes and behavioral outcomes. Since implicit and explicit attitudes are often related, explicit attitudes were also measured. This allows me to examine whether my manipulation of implicit attitudes is independent of explicit attitudes and allows me to ensure that any evidence in favor of causality of implicit attitudes is not driven by explicit attitudes. Finally, I present a reanalysis of existing data to further explore exactly what these manipulations of implicit attitudes are changing and to better understand how implicit attitudes may cause behavior.

CHAPTER II: Experiment 1

An experiment would provide the strongest evidence that implicit attitudes cause behavior towards individual group members. In Experiment 1, participants were randomly assigned to develop different implicit preferences towards two groups. Implicit attitudes were measured to confirm that the manipulation was successful and behavior towards members of the two groups was assessed. Since implicit and explicit attitudes are typically weakly to moderately correlated (Greenwald et al., 2009), explicit attitudes were also measured. This allowed me to examine whether any relationship between implicit attitudes and behavior was an artifact of the impact of explicit attitudes on behavior.

I also opted to use groups of fish (rather than human groups) as stimuli in this experiment for several reasons. First, participants seemed likely to experience less social desirability pressure when making judgments and decisions about fish rather than humans. Second, fish were relatively neutral stimuli for which individuals likely did not have strong existing attitudes. A lack of strong pre-existing attitudes may make the manipulation of implicit preferences easier. Third, the fish stimuli were created (using photo morphing software) such that they exhibited features that communicated group membership (color, shape, etc.) but such that each exemplar varied regarding internal features.

Method

Participants and Design. One-hundred and sixty-three undergraduates (73 female, 90 male, 0 other gendered, M_{age} = 19.24, age range: 18-25) at the University of Colorado Boulder participated in this study in exchange for partial course credit. Six participants were excluded for exhibiting extremely high error rates or other behavior indicating that they were not fully attending to tasks on at least 2 occasions during the study. Thus, 157 (72 female, 85 male; M_{age} =

19.26, age range: 18-25) participants remained in the final sample. Given the design, this final sample size gave me the ability to detect a relatively small effect (cohen's d = 0.32) with 80% power.

Participants were randomly assigned to one of two conditions: the orange-good condition in which participants were assigned to develop positive implicit attitudes towards the orange group of fish; and the purple-good condition, in which positive implicit attitudes towards purple fish were created. Implicit and explicit attitudes were both measured variables.

Materials. ***Stimuli.*** Stimuli were created by morphing the prototype orange fish with other species of fish along a continuum. A purple fish prototype was also morphed with several other species of fish (although the other species did not overlap with those used to create the orange stimuli). Fish prototypes from each group were morphed along a continuum from 0% prototype to 100% prototype, in 10% increments. This yielded orange and purple fish stimuli that retained group attributes but also varied in terms of individual characteristics. From this pool of 110 stimuli, 30 purple and 30 orange fish were selected based on pre-test ratings (using a separate sample of participants from Amazon's Mechanical Turk website) of liking of the individual images (see Appendix A for stimulus images and pretest ratings). As a set, liking ratings of the orange and purple fish used for this study were not significantly different, $M_{\text{orange}} = 54.66$, $M_{\text{purple}} = 54.82$, $t(28) = 0.30$, $p = 0.76$. Fish were then divided by task so that no overlapping stimuli appeared in the tasks manipulating implicit attitudes, measuring implicit attitudes and measuring behavior. This was to ensure that any significant effects on behavior would be due to conditioning of group-based attitudes and not due to conditioning attitudes towards an individual stimulus. Ten fish from each group were selected for the sorting and

evaluative conditioning tasks[9], 8 fish from each group were selected for the implicit attitude measure and 12 fish from each group were selected for the behavioral tasks. Within each task, orange and purple fish did not significantly differ in their pre-test liking ratings, all *p*s > 0.71.

Manipulation of Implicit Attitudes. Two tasks were used to manipulate implicit attitudes: a sorting task and an evaluative conditioning task. The first task followed the procedures of Greenwald, Pickrell, and Farnham (2002) to create new implicit attitudes. Specifically, the participants were asked to sort 16 fish stimuli, 8 fish from the purple group ("Group P") and 8 fish from the orange group ("Group O"), with a cue to prompt the group to which each image belonged. Next, participants viewed 8 images of either purple (in the purple-good condition) or orange (in the orange-good condition) fish for 60 seconds. Finally, participants were asked to sort the same 16 fish (8 orange and 8 purple) without prompting.

Next, an evaluative conditioning procedure was used to continue to build more positive implicit attitudes towards either the orange group of fish (orange-good condition) or the purple group of fish (purple-good condition), following the procedures of Olson & Fazio (2002). Participants were informed that their task was to identify target images among other distracting images by pressing the spacebar when a target image appeared on the screen. The target image was always a fish that clearly did not belong to either the purple or the orange group. Participants viewed 88 trials per block in which one or two words, one or two images or a combination of one word and one image were displayed on the screen for 1200ms (with a 1000ms interstimulus interval). Critically, five of the trials simultaneously paired fish from the orange group (in the

[9] Since both tasks were part of the manipulation of implicit attitudes, the same stimuli were used in both the sorting and evaluative conditioning tasks.

orange-good condition) with evaluatively positive words[10] and another five trials simultaneously paired fish from the purple group (in the orange-good condition) with negative words. The pairings of fish groups and valenced words were reversed in the purple-good condition so that participants viewed purple fish paired with positive words and orange fish paired with negative words (see Appendix B for a schematic and stimuli used in this task). Participants completed 5 blocks of this task. Each block contained a different neutral target image. Ten orange and 10 purple fish were selected to appear in the evaluative conditioning task. Five fish from each group were randomly selected to appear in each block of this task.

Implicit Attitude Measure. The Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) was used to assess implicit attitudes towards the two groups of fish because it is the most commonly used measure of implicit attitudes (Nosek, Hawkins, and Frazier, 2011) and because responses on the IAT may be less sensitive to the individual stimuli presented during the task (De Houwer, 2001; Mitchell, Nosek, & Banaji, 2003; Olson & Fazio, 2003; Wolsiefer, Westfall, and Judd, 2016). That is, the IAT may provide a better group-based measure of attitudes compared to other implicit attitude measure. In the IAT, participants were asked to use the "E" and "I" keys to sort photos of orange and purple fish while simultaneously sorting positive and negative words. Participants were instructed to sort fish based on whether they belonged to Group O or Group P and to sort words based on whether they were "good" or "bad" (see Appendix C for a list of these words). After completing practice blocks in which they sorted just positive and negative words and just fish from the orange and purple groups, participants

[10] Positive and negative words were selected based on valence ratings data from Warriner, Kuperman, and Brysbaert (2013). The words selected were chosen to be near the extreme positive or negative ends of the response scale and roughly equated so that the positive words were rated as positively as the negative words were rated negatively.

completed four critical blocks. In two critical blocks (one 20-trial and one 40-trial block) participants pressed the same key in response to both orange fish and positive words and another key in response to both purple fish and negative words. In the remaining two critical blocks (also 20-trials and 40-trials in length) participants completed the task with the opposite response mappings (i.e. purple fish and positive words shared a response key and orange fish and negative words shared a response key). The order in which participants completed these two types of blocks was counterbalanced. For each participant, an IAT d-score was calculated following the recommendations of Greenwald, Nosek, and Banaji (2003). A higher, positive d-score indicated an implicit preference for orange fish over purple; whereas a more negative d-score indicated an implicit preference for purple fish over orange fish.

Fish Rescue Game. A fish rescue game was used as one behavioral measure and was designed to minimize the opportunity participants had to rely on more deliberative processes (Fazio & Towles-Schwen, 1999). In this task, participants were told there had been a "devastating spill of toxic waste into a local lake and the delicate ecosystem is in danger. Your job is to save as many fish as possible." Participants were also told that some fish had already been contaminated by the toxic waste and that catching an unhealthy fish would harm any healthy fish that the participant caught. Participants were instructed to click on healthy fish to catch and save them while avoiding any unhealthy fish. Participants were further incentivized by being awarded 10 points for each healthy fish saved and losing 10 points for each unhealthy fish captured.

In each block, participants viewed a screen with 12 healthy and 12 unhealthy fish (healthy and unhealthy fish were equally represented across purple/orange groups, see Appendix D). The fish moved around the screen quickly in a pseudorandom pattern. Participants were

given 5 seconds to catch as many healthy fish as possible. Between blocks, participants were told their score and encouraged to amass more points. Participants completed 10 blocks of this task. Behavioral preference scores were calculated by subtracting the total number of purple fish saved from the total number of orange fish saved. Thus, a more positive difference score indicated that the participant was more likely to save orange compared to purple fish. A negative difference score indicated that the participant was more likely to save purple compared to orange fish.

Forced choice task. Previous work has demonstrated that evaluative conditioning of preference for an individual stimulus is related to preference for that stimulus in a subsequent forced choice task (Kendrick & Olson, 2012); therefore, I included the forced choice task as a second behavioral measure. In the forced choice task, participants were asked to help select stimuli for a future study by selecting, from pairs of images, the image they preferred. Participants were instructed to "go with their first instinct" and to not spend a lot of time deliberating on their choice. Participants then saw 30 pairs of images, one at a time, and were instructed to click on the image they preferred. Twenty-four of these trials were filler trials in which two fish that were not members of the target groups were paired with each other or in which neutral images were paired with each other. On six trials, participants saw orange fish paired with neutral fish (2 trials), purple fish paired with neutral fish (2 trials) or purple and orange fish paired with each other (2 trials). For each trial, the image that the participant selected was recorded. The six orange and purple images selected for this task were randomly selected from the group of behavioral orange and purple stimuli. A behavioral preference score was calculated as the number of orange fish selected minus the number of purple fish selected across

all six trials[11]. A positive score on this measure also indicated a relative preference for orange over purple fish and a negative score indicated the reverse.

Explicit Attitude Measure. Thermometer ratings were used to assess explicit attitudes towards the two fish groups. For each group, participants were asked to "Click the button that corresponds with your feelings towards the orange [purple] group." Participants then clicked to select a button ranging from 0 (cool and unfavorable feelings) to 100 (warm and favorable feelings) in 10-point increments[12]. For each participant, their ratings of the purple fish group were subtracted from their ratings of the orange fish group such that a positive number indicated an explicit preference for orange fish; whereas, a negative number indicated a preference for purple fish.

Procedure. After obtaining informed consent, participants were seated at a computer in an individual cubicle. All tasks were completed on the lab computer and were self-paced. First, participants completed the sorting and evaluative conditioning tasks. Next, implicit attitudes were measured using the IAT and then participants completed the fish rescue followed by the forced choice task. Next, participants completed the explicit attitude measure, a series of questions assessing the difficulty and engagement of each task, and awareness of the research hypothesis. Finally, participants completed demographic items. After completing the study, all participants were thanked, debriefed and awarded credit for participating in the study.

---

[11] I also examined relative preference for orange compared to purple fish for just trials in which participants had to choose between an orange and purple fish. This did not change the results.

[12] To ensure that thermometer ratings of the two groups of fish were truly group ratings (and not ratings of a subset of the members of the two fish groups), no images of either group were shown during the thermometer rating items.

Results[13]

Three primary questions examined whether there was evidence that implicit or explicit group attitudes caused behavior towards individual group members: 1. Were there condition differences in implicit or explicit attitudes? 2. Were there condition differences in behavior? And 3. Were implicit or explicit attitudes related to behavior? Finally, a mediation model could examine whether condition differences in behavior could be independently explained by either implicit or explicit attitudes.

Condition Differences in Attitudes. To assess condition differences in implicit attitudes, participant IAT scores were regressed on a contrast coded condition variable (orange-good condition = 0.5, purple-good condition = -0.5). Figure 1 depicts mean implicit and explicit attitudes by condition. There was evidence that implicit attitudes differed by condition, $b = .18$, $t(155)=2.56$, $p = .01$, $R^2 = 0.04$, with participants in the orange-good condition showing more favorable implicit attitudes towards orange fish than participants in the purple-good condition. This difference occurred even though participants in both conditions demonstrated implicit attitudes that were relatively more positive toward orange fish than purple fish (orange-good condition: $M_{IAT} = 0.43$, $t(155) = 8.81$, $p < .001$, $R^2 = 0.33$; purple-good condition: $M_{IAT} = 0.25$, $t(155) = 4.88$, $p < .001$, $R^2 = 0.13$). Because implicit and explicit attitudes were also significantly related to each other, $b = 0.003$, $t(132) = 2.74$, $p = 0.006$, $R^2 = 0.05$, it was important to examine whether condition differences in implicit attitudes were explained by differences in explicit attitudes. There was no evidence that this was the case. Condition differences in implicit attitudes

[13] I also analyzed these data treating stimuli as a random factor in the IAT. See Appendix E for a summary of that analysis.

persisted even after controlling for explicit attitudes and IAT block order, $b = 0.15$, $t(130) = 2.05$, $p = .04$[14], $R^2_{partial} = 0.03$.

Finally, it should be noted that a marginal effect of condition on explicit attitudes was obtained, $b = 9.98$, $t(132) = 1.88$, $p = 0.06$, $R^2 = 0.03$. This effect indicated that individuals in the orange-good condition also self-reported a significant explicit preference for orange over purple fish, $M_{orange\text{-}good} = 13.04$, $t(132) = 3.83$, $p < 0.001$, $R^2 = 0.10$, whereas participants in the purple-good condition did not demonstrate a significant explicit preference for either group, $M = 3.97$, $t(132) = 1.03$, $p = 0.31$, $R^2 = 0.03$. These condition effects on explicit attitudes were attenuated after controlling for implicit attitudes and IAT block order, $b = 7.00$, $t(130) = 1.32$, $p = 0.19$, $R^2_{partial} = 0.01$.

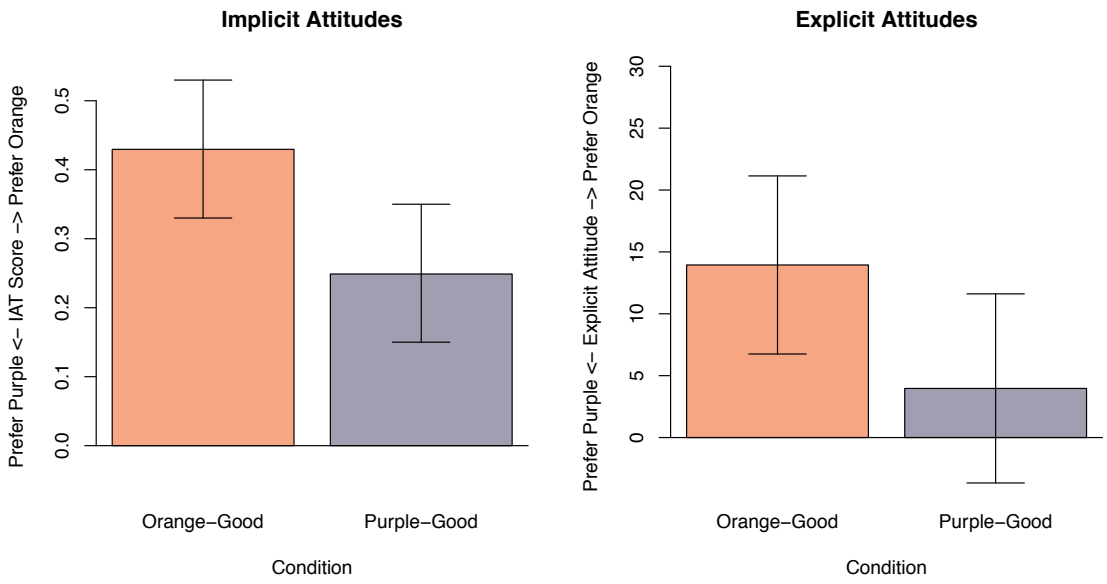


Figure 1. Condition Differences in Attitudes, Experiment 1. Average IAT and explicit attitude scores by condition. Higher values on the y-axis indicate a greater preference for orange over purple fish. In both panels, zero represents no preference for one fish group over the other. The error bars represent 95% confidence intervals.

[14] Degrees of freedom are lower here because 23 participants did not provide explicit attitude ratings towards the two fish groups.

Condition Differences in Behavior. *Fish Rescue Game*. Across all 10 blocks, participants had the opportunity to save 240 fish and 120 of these were unmarked healthy fish—120 of these were healthy and evenly split between orange and purple groups. On average, participants chose to save 67.72 (SD = 14.48), 89.78% of them healthy. Table 2 presents the average number of orange and purple fish saved by health status (healthy vs. unhealthy) and condition (orange-good vs. purple-good).

To analyze data from this task, I calculated the number of orange and the number of purple fish "rescued" (regardless of healthy/unhealthy status) and calculated a difference score for each participant. A positive number indicated that a participant saved a greater number of orange fish and a negative score indicated a participant saved a greater number of purple fish. Next, this difference score was regressed on condition. By this metric, there was no evidence of condition differences in behavior, $b = 0.99$, $t(154) = 0.78$, $p = 0.44$, $R^2 = 0.00$. Examining these effects separately for healthy and unhealthy fish yielded the same results.

Table 2

Saves by Fish Color, Health Status, and Participant Condition- Fish Rescue Game

		Orange Fish		Purple Fish	
Condition	N	Healthy	Unhealthy	Healthy	Unhealthy
Orange-good	82	31.06 (6.07)	3.65 (2.34)	30.57 (6.69)	3.68 (2.36)
Purple-good	74	28.79 (8.21)	4.64 (3.83)	30.50 (8.26)	3.98 (3.20)

Note. Above are the mean (SD) number of healthy and unhealthy fish saved during the fish rescue game reported by fish color and participant condition (Experiment 1).

Forced Choice Task. For the forced choice task, the difference in orange – purple fish selections was regressed on the contrast-coded condition variable. This simple model revealed no evidence of condition differences in fish selection, $b = 0.16$, $t(155) = 0.50$, $p = 0.62$, $R^2 = 0.00$.

Relationship between Attitudes and Behavior. ***Fish Rescue Game.*** Figure 2 represents the relationship between attitudes and behavior for the fish rescue game. To examine whether implicit attitudes were related to behavior, I regressed the difference in orange fish saved minus purple fish saved on IAT score and found no evidence of a relationship between implicit attitudes and biased behavior, $b = 0.22$, $t(154) = 0.15$, $p = 0.88$, $R^2 = 0.01$ (this remained the case even after controlling for explicit attitudes and IAT block order, $b = 0.35$, $t(129) = 0.20$, $p = 0.84$, $R^2_{\text{partial}} = 0.00$).

Explicit attitudes also neglected to yield a statistically significant relationship with behavior in the fish rescue game, $b = -0.02$, $t(131) = 1.06$, $p = 0.29$, $R^2 = 0.01$. This relationship remained non-significant after controlling for IAT score and IAT block order, $b = -0.02$, $t(129) = -1.07$, $p = 0.29$, $R^2_{\text{partial}} = 0.01$.

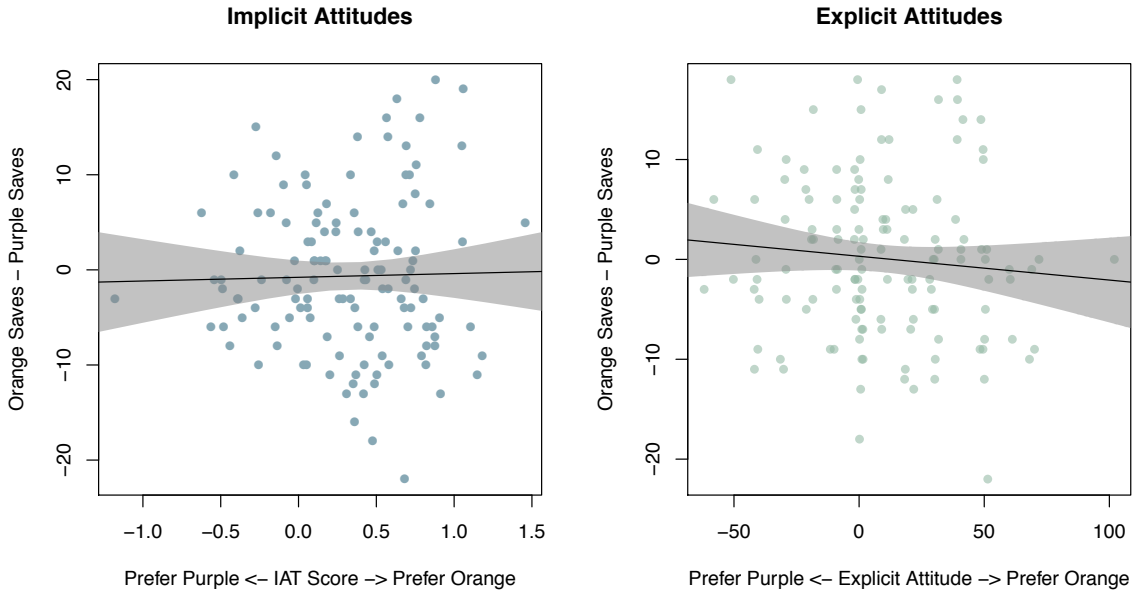


Figure 2. *Attitude Behavior Relationship, Fish Rescue Game.* The two panels represent the relationship between IAT scores and behavior (left panel) and the explicit attitudes and behavior (right panel) for the fish rescue game in Experiment 1. The solid black line represents the line of best fit and the gray bands depict the 95% confidence interval. Values for explicit attitudes are jittered to display all points.

Forced Choice Task. The forced choice task also yielded no evidence of a relationship between implicit attitudes and behavioral preference (see Figure 3). A simple model regressing the difference in number of orange fish selected minus number of purple fish selected, on IAT score, did not demonstrate a significant relationship, $b = 0.16$, $t(155) = 0.50$, $p = 0.62$, $R^2 = 0.00$. This remained true after controlling for explicit attitudes and IAT block order, $b = -0.11$, $t(130) = -0.30$, $p = 0.76$, $R^2_{\text{partial}} = 0.00$.

In contrast to implicit attitudes, explicit attitudes did demonstrate a significant relationship with behavior on the forced choice task. Self-reported preferences for orange fish over purple fish were related to preferences for orange over purple fish in the forced choice task, $b = 0.02$, $t(132) = 3.55$, $p < 0.001$, $R^2 = 0.09$. This significant relationship remained after

controlling for implicit attitudes and IAT block order, $b = 0.02$, $t(130) = 3.48$, $p < 0.001$, $R^2_{\text{partial}} = 0.09$.

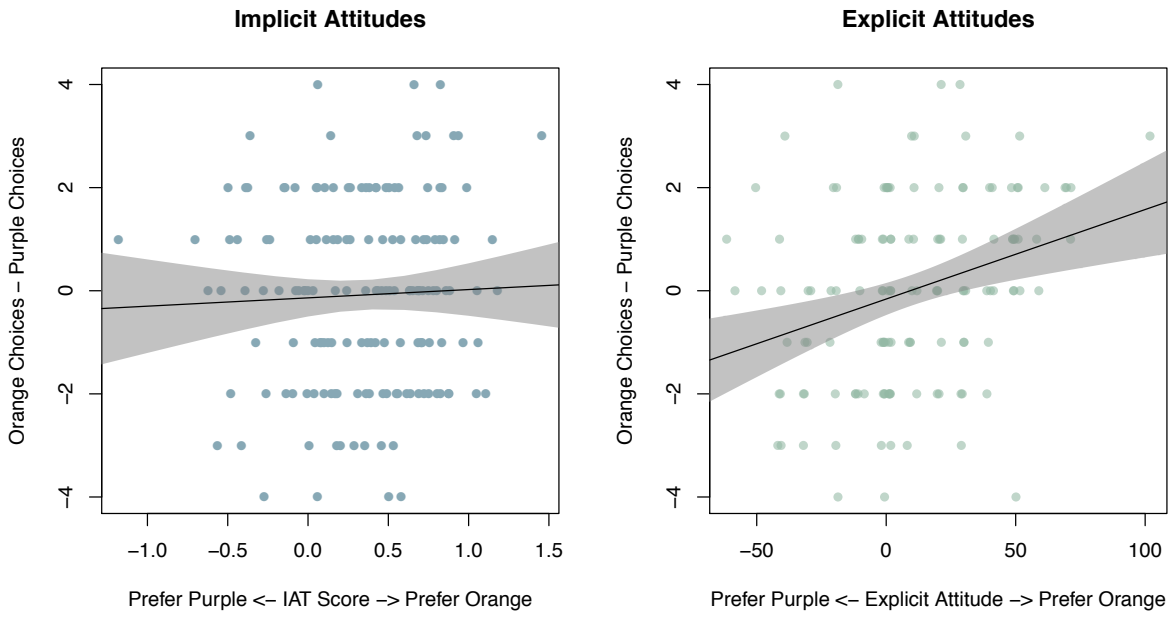


Figure 3. *Attitude Behavior Relationship, Forced Choice Task.* The two panels represent the relationship between IAT scores and behavior (left panel) and the explicit attitudes and behavior (right panel) for the fish rescue game in Experiment 1. The solid black line represents the line of best fit and the gray bands depict the 95% confidence interval. Values for explicit attitudes are jittered to display all points.

Mediational Models. Models testing multiple mediation were used to examine causal evidence in a slightly different way. Even though there is no evidence of statistically significant condition differences in behavior, it could be the case that the explained variance in the outcome by condition (even if not statistically significant) is due to either implicit or explicit attitudes (or both). Two mediational models, one for each behavioral task, were tested to examine this possibility.

Fish Rescue Game. See Figure 4 for the path model for the fish rescue game. Like the simple linear regression analysis, there was evidence of condition differences in implicit attitudes, $b = 0.19$, $z = 2.31$, $p = 0.02$, and marginal condition differences in explicit attitudes, $b =$

0.16, $z = 1.90$, $p = 0.06$. Also in parallel with the linear regression analysis, there was no evidence of a significant total effect of condition on the difference in orange vs. purple fish saved, $b = 0.04$, $z = 0.41$, $p = 0.68$. Neither indirect effect was statistically significant, explicit: $b = -0.02$, $z = -1.03$, $p = 0.31$; implicit; b = 0.01, z = 0.30, p = 0.76, indicating that any variance in behavior in the fish rescue game accounted for by condition was explained by neither implicit nor explicit attitudes. Such null effects are not supportive of either implicit or explicit attitudes as a cause of behavior in the fish rescue game.

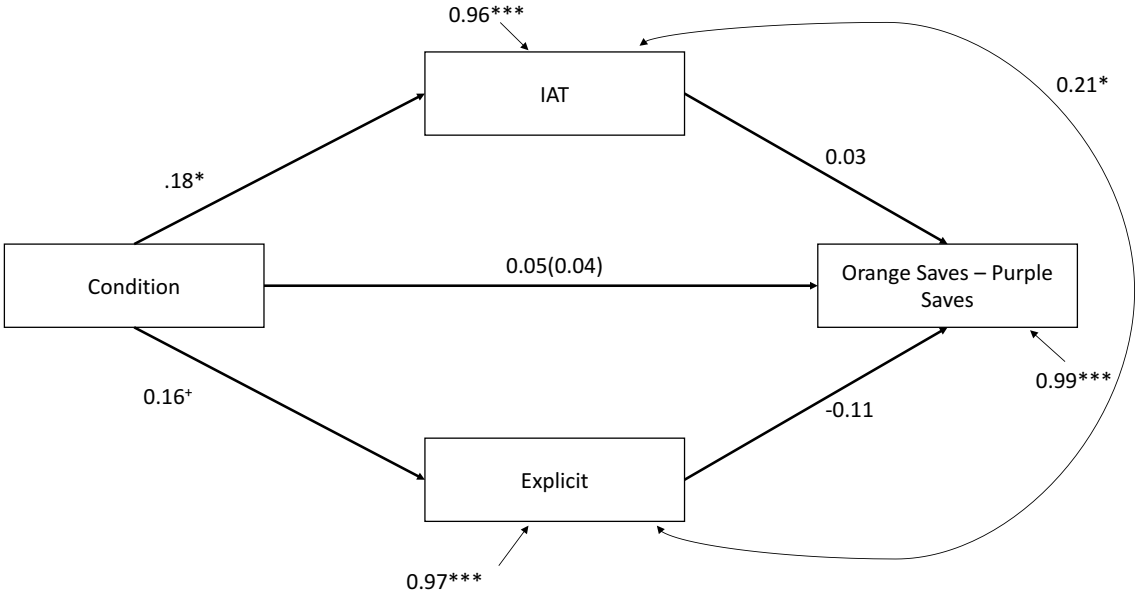


Figure 4. *Multiple Mediation Model, Fish Rescue Game*. Multiple mediation path model examining the impact of condition on differences in the number of orange versus purple fish saved during the fish rescue game. Total effect of condition on differences in orange vs. purple saves is represented in parentheses. Numeric values represent standardized path estimates. Neither indirect effect was statistically significant, explicit: $b = -0.02$, $z = -1.03$, $p = 0.31$; implicit; b = 0.01, z = 0.30, p = 0.76+p < 0.1; *p < .05; **p< 0.01; ***p < 0.001.

Forced Choice Task. See Figure 5 for the detailed path model for this outcome measure. Again, this model yielded significant condition differences in implicit attitudes, $b = 0.18$, $z = 2.23$, $p = 0.03$ and marginally significant condition differences in explicit attitudes, $b = 0.16$, $z = 1.93$, $p = 0.05$. Although the total effect of condition on the forced choice difference score was non-significant, $b = -0.06$, $z = -0.67$ $p = 0.50$, the indirect effect for explicit attitudes was in the expected direction, $b = 0.05$, $z = 1.72$, $p = 0.09$. In contrast, the indirect effect of implicit attitudes was non-significant, $b = -0.003$, $z = -0.22$, $p = 0.83$.

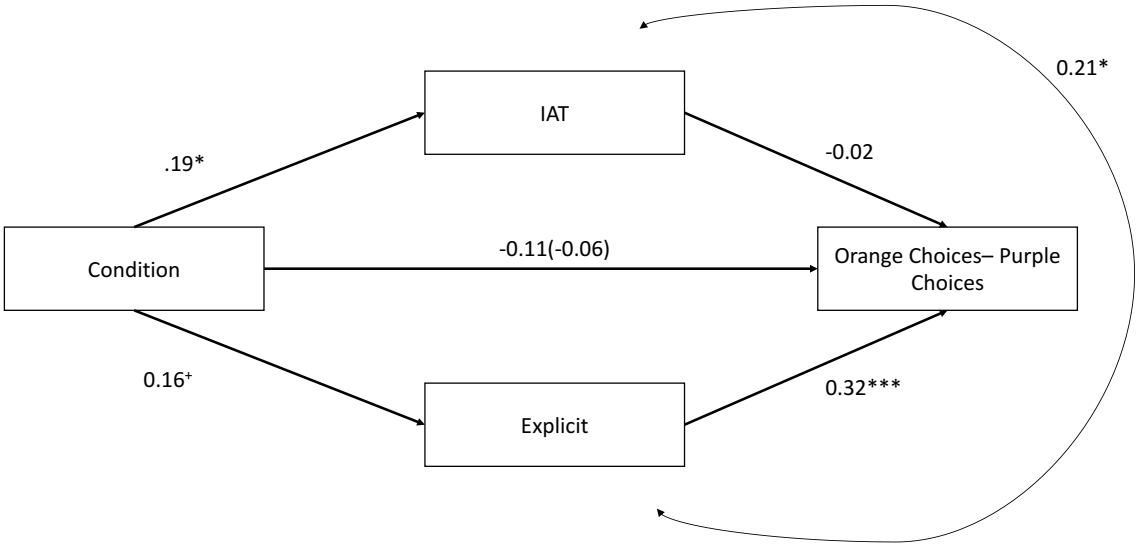


Figure 5. *Multiple Mediation Model, Forced Choice Task*. Multiple mediation path model examining the impact of condition on differences in the number of orange versus purple fish selected during the forced choice task. Total effect of condition on differences in orange vs. purple selections is represented in parentheses. Numeric values represent standardized path estimates. The indirect effect for explicit attitudes was marginally significant in the expected direction, $b = 0.05$, $z = 1.72$, $p = 0.09$. The indirect effect of implicit attitudes was non-significant, $b = -0.003$, $z = -0.22$, $p = 0.83$. $^{+}p < 0.1$; *p < .05; **p< 0.01; ***p < 0.001.

Discussion

Although implicit attitudes were successfully manipulated over and above explicit attitudes, there was no evidence that these implicit attitudes impacted behavior. The only hint that attitudes explained behavior at all was through explicit attitudes, which is consistent with the MODE model (Fazio & Towles-Schwen, 1999). Theoretical models such as the MODE model (Fazio & Towles-Schwen, 1999) suggest that more deliberative processes, which might increase reliance on explicit attitudes, would be more likely to impact outcomes when individuals have the motivation and opportunity to engage in more deliberative behavior. Since participants were given an unlimited amount of time to make selections during the forced choice task and there were no other constraints placed on working memory, most participants should have had the opportunity to engage in more deliberative thinking. Although no part of the research design was specifically created to motivate participants to respond deliberatively, the very nature of participation in a lab study may have motivated participants to think carefully about their decisions.

Several possibilities may help to explain the null results regarding implicit attitudes. First, it may be the case that implicit group attitudes do not cause behavior towards individual group members. Second, it may be that the fish stimuli used in this experiment were too neutral. Although the neutral nature of my stimuli was initially considered a strength of my design, the fish stimuli may have been uninteresting to participants and resulted in random responding. Additionally, it is possible that the fish rescue game was simply too chaotic to pick up on preferences for one group over the other. This possibility was supported by the fact that, on average, only about half of all possible healthy fish were “saved” during the fish rescue game. In contrast, the forced choice task might have invoked (deliberate) processes that minimized

implicit bias. In Experiment 2, I attempted to address potential issues related to stimuli and the behavioral outcome measures.

CHAPTER III: Experiment 2

Experiment 2 was designed to address some potential explanations for the null results from Experiment 1. First, I attempted to address the issue of participant engagement by using more engaging stimulus groups—dogs and cats. Second, I changed the nature of the behavioral outcome to decrease the likelihood of random responding. Finally, since Experiment 1 (and another pilot test) yielded little evidence of a relationship between implicit attitudes and performance on the forced choice task, I eliminated this task from Experiment 2.

Method

Participants and Design. Two-hundred fourteen University of Colorado Boulder undergraduates (103 female, 109 male, 2 other gendered, M_{age} = 19.07, age range: 18-31) completed this study in exchange for partial course credit or for payment. Seven participants were excluded for exhibiting at least two instances of high error rates, failed attention checks or other patterns of responding that indicated they were not following instructions. The left 207 participants in the final sample (100 female, 106 male, 1 other gendered, M_{age} = 19.00, age range: 18-31).

Participants were randomly assigned to one of two conditions: the dogs-good condition in which implicit attitudes were manipulated to be positive towards dogs; and the cats-good condition, in which participants' implicit attitudes were manipulated to be positive towards cats.

Materials. ***Stimuli.*** Twenty-five dog and twenty-five cat images of roughly equal quality were obtained from humane society websites (see Appendix F for stimulus images). Images were edited to display only cat and dog faces and to appear as the same size. These images were divided up so that stimuli used to manipulate implicit attitudes, measure implicit attitudes and measure behavior did not overlap. Sixteen (eight dogs and eight cats) of these images were used

for practice sorting exemplars from the two groups and to display members of one group to increase implicit preference for one group over the other. Twenty dog and cat images (including the 16 from the initial sorting task) were used in the evaluative conditioning task.[15] Ten images (five dogs and five cats) were used for the IAT and 20 images (10 dogs and 10 cats) were used for the behavioral task.

Manipulation of Implicit Attitudes. Participants were randomly assigned (within the computer program) to either a dogs-good condition or a cats-good condition. Implicit attitudes were manipulated using the same combination of sorting, exposure to one animal group and evaluative conditioning tasks from Experiment 1. The only difference was the substitution of non-dog and non-cat animals (primarily mammals) as the targets in the evaluative conditioning task.

Implicit Attitude Measures. The measure of implicit attitudes (the IAT) was identical to the IAT used for Experiment 1 with the exception that dog and cat images replaced images of fish. IAT d-scores were again calculated for each participant following the recommendations of Greenwald, Nosek, and Banaji (2003).

Pet Rescue Game. The fish rescue game was modified to decrease random responding. Rather than display all stimuli at one time moving on the screen, images were displayed sequentially. Participants were instructed to "Imagine there is an abandoned town filled with stray cats and dogs. A rabies epidemic is spreading and we would like you to help save as many healthy animals as possible." As with Experiment 1, participants were instructed that their job was to save healthy animals and leave behind unhealthy animals. Participants then viewed

[15] Since both the sorting procedure and the evaluative conditioning task are part of the manipulation of implicit attitudes, I used overlapping sets of stimuli.

images of dogs and cats, one-at-a-time (in a randomized order) and pressed the "S" key to save healthy animals or the "L" key to leave behind unhealthy animals. Healthy animals were unmarked, whereas unhealthy animals appeared with a skull and bones symbol in a random location on the image. Participants were given a 1000ms response window but to encourage more automatic processing, participants were given a warning if they responded slower than 500ms. Between trials, participants viewed a fixation cross for the 500ms interstimulus interval (Appendix G includes a schematic for this task). Participants completed two blocks of this task with 96 trials per block. Statistics from signal detection theory were used to analyze data in this study. The calculation of these outcome measures is described, in detail, in the results section.

Explicit Attitude Measures. Recent work suggests that multiple regression analysis with even moderately reliable covariates can yield results with inflated type I error rates (Westfall and Yarkoni, 2016). To increase the reliability of my explicit attitude measure, I added several items to the thermometer ratings scales. For both dogs and cats, participants completed a thermometer rating (i.e. "Click the button that corresponds to your feelings towards dogs [cats])," with responses ranging from 0 (cool and unfavorable feelings) to 100 (warm and favorable feelings) in increments of 10. Participants were also asked to respond to four semantic differential items assessing explicit attitudes towards dogs (e.g. Unpleasant-Pleasant; Disliked-Liked; Bad-Good; Boring-Fun). Responses on the semantic differential items ranged from 0 (most negative rating) to 4 (most positive rating). For each of the explicit attitude items, the participants rating of cats was subtracted from their rating of dogs such that five difference scores were created. For each difference score a positive number indicated an explicit preference for dogs over cats and a negative score indicated the reverse. These items were standardized and averaged to form an

overall explicit attitudes scale (α = 0.92). An attention check was also embedded in these explicit attitude items.

Procedure. The procedure was the same as Experiment 1, with two exceptions. First, the new pet rescue game replaced the fish rescue game from Experiment 1. Second, participants did not complete a forced choice task in this study.

Results

As before, I tested the relevant paths needed to determine a causal relationship using both linear models and path models.

Condition Differences in Attitudes[16]. To examine condition differences in implicit attitudes, IAT d-scores were regressed on contrast-coded condition (dogs-good condition = 0.5, cats-good condition = -0.5). As in Experiment 1, there were significant condition differences in implicit attitudes, b = 0.13, $t(205)$ = 2.21, p = .03, R^2 = 0.02 (see Figure 6). Participants in the dogs-good condition demonstrated a positive implicit preference for dogs over cats, b = 0.12, $t(205)$ = 2.92, p = .003, R^2 =04 , whereas participants in the cats-good condition demonstrated a small, but non-significant, implicit preference for cats over dogs, b = -0.01, $t(205)$ = -0.22, p = 0.82, R^2 = 0.00. Again, implicit and explicit attitudes were significantly related to each other, b = 0.03, $t(205)$ = 4.17, $p < 0.001$, R^2 = 0.08. However, condition differences in implicit attitudes remained significant after controlling for explicit attitudes and IAT block order, b = 0.15, $t(203)$

[16] There is some debate about whether evaluative conditioning effects can occur outside of the awareness of the contingencies between the condition stimuli (in this case dogs and cats) and the unconditioned stimuli (in the present work-- evaluative words; Baeyens, Eelen, & van den Bergh, De Houwer, Thomas, & Baeyens, 2001; 1990; Pleyers, Corneille, Luminet, & Yzerbyt, 2007). Thus, it seemed possible that contingency awareness could moderate effects of conditioning on implicit attitudes, explicit attitudes or on the effects of condition or attitudes on behavior. Contingency awareness was measured in Experiments 2-5. Moderating effects are summarized in Appendix H.

= 2.74, $p = 0.007$, $R^2_{partial} = 0.04$. The manipulation did not appear to impact explicit attitudes, $b = 0.01$, $t(205) = 0.04$, $p = 0.97$, $R^2 = 0.00$. Thus, the manipulation successfully manipulated implicit, but not explicit attitudes.

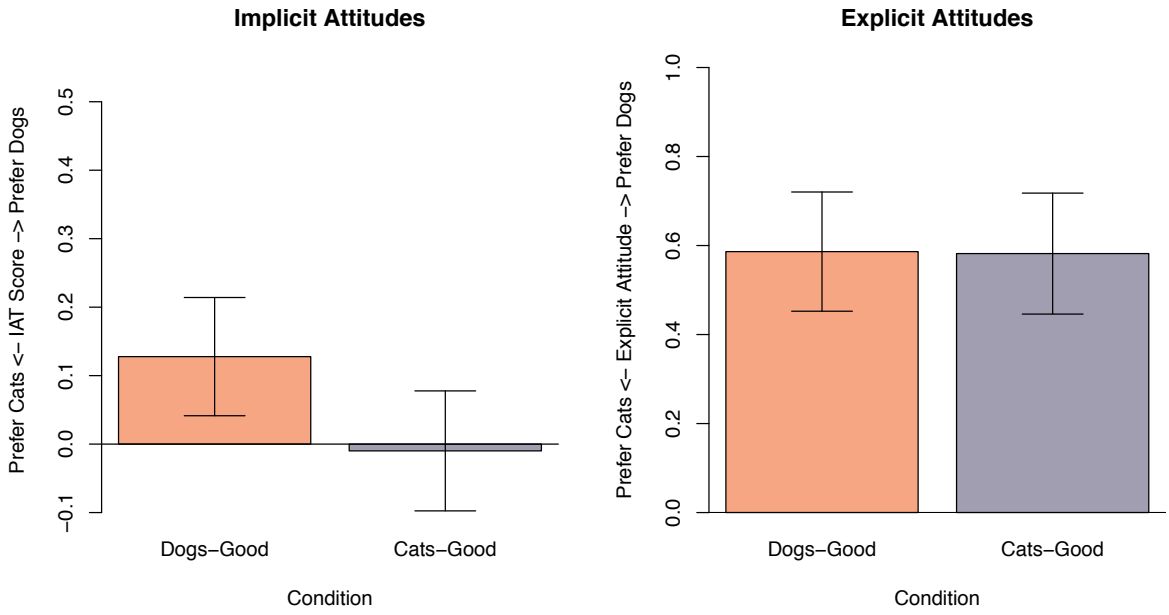


Figure 6. *Condition Differences in Attitudes, Experiment 2*. Average IAT and explicit attitude scores by condition. Higher values on the y-axis indicate a greater preference for dogs over cats. In both panels, zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

Condition Differences in Behavior. Condition differences in behavior were examined using signal detection analysis. Signal detection analysis (SDT) involves calculating two different metrics for participant performance in the pet rescue game: discriminability (*d'*) and response bias (*c*; Stanislaw & Todorov, 1999). The use of signal detection theory allows me to consider both whether participants decided to save or leave animals and whether they did so correctly with two independent metrics. Discriminability (*d'*) is a metric that describes the extent to which the participant was correctly able to differentiate sick from healthy animals. That is, the extent to which participants correctly saved healthy animals and left behind unhealthy ones. This

metric was calculated by subtracting the z-transformed proportion of false alarms from the z-transformed proportion of hits, where a hit was correctly saving a healthy animal. This means that positive values of d' indicate greater discriminability. Response bias (*c*) indicates the extent to which a participant tended to make one type of response (save) or the other (leave behind) regardless of what the correct decision was. This metric was calculated as the average of the z-transformed proportion of hits and the z-transformed proportion of false alarms where a hit is defined as correctly saving a healthy animal and a false alarm is defined as accidentally saving an unhealthy animal[17]. In this analysis, a more positive value indicates a more liberal bias in favor of saving animals[18]. Table 3 contains the average values of *d'* and *c* by condition and trial type.

In this task, behavioral bias in favor of dogs vs. cats might be indicated by either *d'* or *c*. Since participants were warned that poor performance on the pet rescue game would lead to poor outcomes for all animals, it was possible that participants would be particularly careful to discriminate between healthy and sick animals more for their preferred group. That is, a participant with a strong dog preference may demonstrate a higher *d'* score for dog trials compared to cat trials. Bias in behavior may also be represented by differences in response bias for dog vs. cat trials. For example, a participant who favors dogs more than cats may reasonably be expected to set a lower threshold for saving dogs than for saving cats. That is, a person who

[17] Typically, this calculation involves multiplying this result by -1 (MacMillan & Creelman, 2004), but, to ensure that higher values on the outcome measures meant preferential treatment for orange over purple fish (mirroring the attitude measures), I did not take this step.

[18] The analyses reported in the main text are based on hand calculations of *d'* and *c*, and assume the same criterion for all trials. Additional models that use maximum likelihood estimation and allow for the estimation of different criterion levels for dog and cat trials were estimated but were nearly identical (r = 0.99 to 1.0 for both *d'* and *c*). More information about these models can be found in Appendix I.

prefers dogs may be more likely to save dogs regardless of their status as healthy/unhealthy. As such, I analyzed both *d'* and *c* to examine whether there were condition differences in behavior. For both metrics, a participant-level difference score was calculated such that a positive value would indicate a higher *d'* (or *c*) for dog vs. cat trials and lower value would indicate a higher *d'* (or *c*) for cats vs. dog trials. As such positive values of *d'* and *c* represent behavioral bias in favor of dogs and negative values represent behavioral bias in favor of cats. These difference scores were then regressed on participant condition (and other predictors).

Table 3

Signal Detection Statistics by Trial Type and Condition, Experiment 2

	Dogs-Good Condition		Cats-Good Condition	
	Cat Trials	Dog Trials	Cat Trial	Dog Trial
Discriminability (*d'*)	2.07 (0.78)	2.18 (0.84)	2.25 (0.84)	2.31 (0.87)
Response Bias (*c*)	0.10 (0.21)	0.14 (0.25)	0.08 (0.21)	0.09 (0.23)

Note. Values represent means (standard deviations) for each signal detection statistic by type of trial in the behavioral task (dog vs. cat) and condition (dogs-good vs. cats-good). For d-prime, higher numbers indicate greater discriminability. For c, more positive numbers indicate a greater bias in favor of saving animals.

Discriminability (d'). Participants were successful at discriminating healthy from unhealthy animals on both dog ($M_{d'}$ = 2.24, $t(207) = 37.95$, $p < 0.001$, $R^2 = 0.87$) and cat trials ($M_{d'}$ = 2.16, $t(207) = 38.27$, $p < 0.001$, $R^2 = 0.88$). Overall, participants were better able to discriminate healthy from unhealthy animals on dog trials compared to cat trials, $b = 0.08$, $t(205) = 2.37$, $p = 0.02$, $R^2 = 0.03$. However, there was no evidence that bias in *d'* was reliant on participant condition, $b = 0.05$, $t(205) = 0.74$ $p = 0.46$, $R^2 = 0.00$. This remained the case after

controlling for explicit attitudes and IAT block order, $b = 0.05$, $t(203) = 0.05$, $p = 0.73$, $R^2_{partial} = 0.00$.

Response Bias (c). On average, Participants demonstrated a liberal save threshold for both dog ($M_c = 0.11$, $t(207) = 6.72$, $p < 0.001$, $R^2 = 0.18$) and cat trials ($M_c = 0.09$, $t(207) = 6.20$, $p < 0.001$, $R^2 = 0.16$). Across condition, response bias did not significantly differ between cat- and dog-trials, $b = 0.02$, $t(205) = 1.22$, $p = 0.23$, $R^2 = 0.01$. Further, there was no evidence that differences in response bias on dog versus cat trials depended on condition, $b = 0.03$, $t(205) = 0.67$, $p = 0.50$, $R^2 = 0.00$. This non-significant effect remained non-significant after controlling for explicit attitudes and IAT block order, $b = 0.02$, $t(203) = 0.60$, $p = 0.50$, $R^2_{partial} = 0.00$.

Relationship Between Attitudes and Behavior. The outcomes used to examine the relationship between implicit attitudes and behavior were the same as those used to examine condition differences in behavior.

Discriminability (d'). Figure 7 represents the relationship between implicit and explicit attitudes and differences in discriminability. There was a significant relationship between IAT scores and difference in d' on dog vs. cat trials, $b = -0.15$, $t(206) = -1.99$, $p = 0.048$, $R^2 = 0.02$. Contrary to hypotheses, individuals with stronger implicit preferences for dogs, demonstrated less ability to discriminate sick from health animals on dog trials compared to cat trials (see Table 1 for mean *d'* and *c* by condition and target animal)**.** This relationship between implicit attitudes and differences in *d'* remained significant after controlling for explicit attitudes and IAT block order, $b = -0.18$, $t(203) = -2.08$, $p = 0.04$, $R^2_{partial} = 0.02$. There was no evidence that explicit attitudes were related to differences in discriminability, $b = -0.002$, $t(205) = -0.30$, $p = 0.76$, $R^2 = 0.00$.

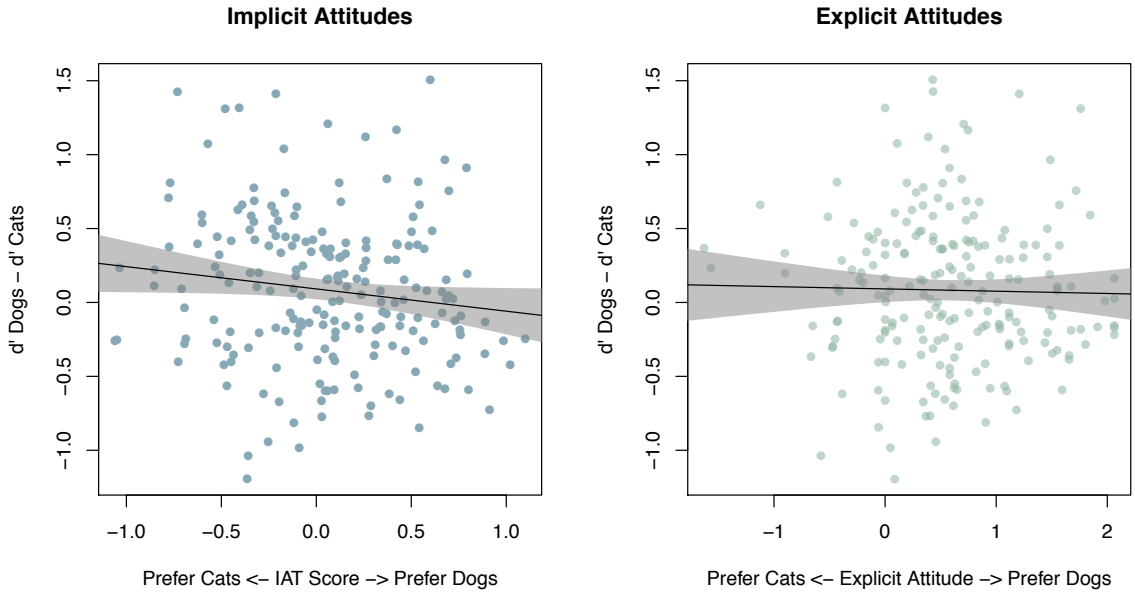


Figure 7. *Experiment 2: Attitude Behavior Relationship, d' Differences*. Relationship between implicit attitudes and differences in d' (left panel) and explicit attitudes and differences in d' (right panel) for dog versus cat trials. Values greater than 0 on the y-axis indicate participants were better able to discriminate healthy from unhealthy dogs compared to cats while negative values indicate the reverse. Gray bands represent the 95% confidence interval around the line of best fit (black line).

Response Bias (c). Figure 8 represents the relationship between implicit and explicit attitudes and differences in response bias. There was a significant relationship between IAT d-scores and differences in response bias for dog vs. cat trials, $b = 0.09$, $t(206) = 2.24$, $p = 0.03$, $R^2 = 0.02$. That is, for individuals with more positive implicit attitudes towards dogs, response bias in favor of saving was greater for dog trials than for cat trials[19]. This relationship between implicit attitudes and differences in response bias remained after controlling for both explicit attitudes and IAT block order, $b = 0.12$, $t(203) = 2.64$, $p = 0.01$, $R^2_{partial} = 0.03$.

[19] Further consideration of this relationship indicated that there was a significant relationship between IAT d-score and response bias (c) for dog trials, $b = 0.11$, $t(206) = 3.14$, $p = 0.002$, $R^2 = 0.05$, but not for cat trials, $b = 0.02$, $t(206) = 0.69$, $p = 0.49$, $R^2 = 0.00$. Any relationship between implicit attitude and behavior appears to be impacting behavior on dog, but not on cat trials.

The relationship between explicit attitudes and response bias difference was marginally significant, $b = 0.008$, $t(205) = 1.96$, $p = 0.05$, $R^2 = 0.02$[20]. Notably, the relationship between explicit attitudes and differences in response bias was attenuated after controlling for implicit attitudes and IAT block order, $b = 0.004$, $t(203) = 0.92$, $p = 0.36$, $R^2_{\text{partial}} = 0.00$.

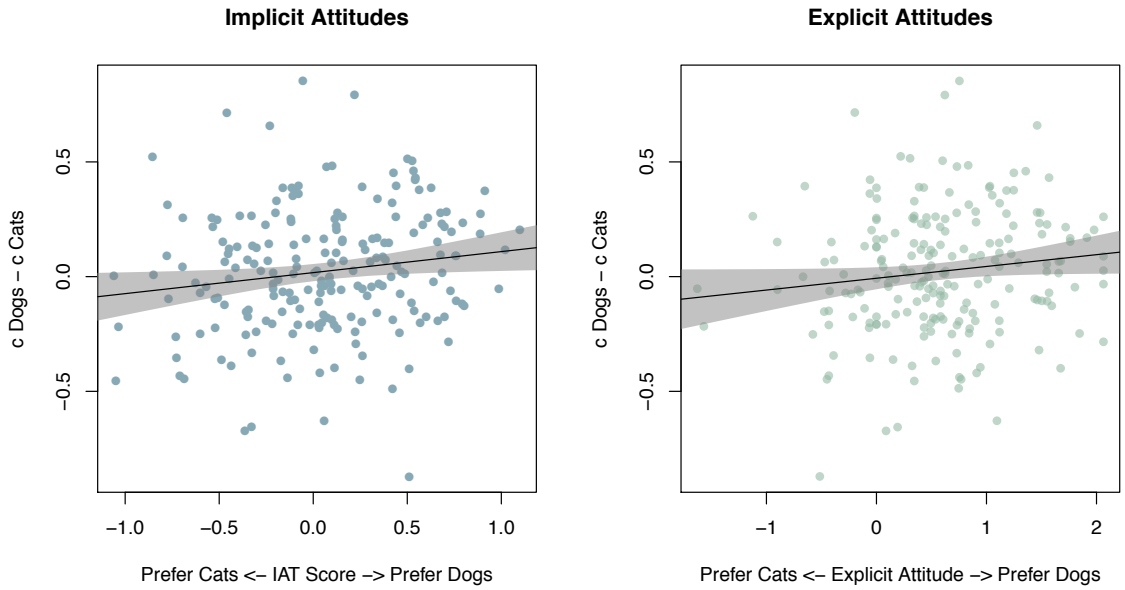


Figure 8. *Experiment 2: Attitude Behavior Relationship, c Differences*. Relationship between implicit attitudes and differences in c (left panel) and explicit attitudes and differences in c (right panel) for dog versus cat trials. Values greater than 0 on the y-axis indicate participants were respond in favoring of saving dogs compared to cats while negative values indicate the reverse. Gray bands represent the 95% confidence interval around the line of best fit (black line).

Mediation Models. Two mediation models, one for each signal detection outcome, were used to simultaneously examine the ability of implicit and explicit attitudes to explain any condition differences in behavior. For each model, condition differences in behavior were

[20] For dog trials, greater explicit dog preferences were associated with greater response bias in favor of saving dogs, $b = 0.01$, $t(205) = 2.56$, $p = 0.01$, $R^2 = 0.03$. However, the relationship between explicit attitudes and response bias on cat trials was non-significant, $b = 0.001$, $t(205) = 0.41$, $p = 0.68$, $R^2 = 0.00$.

estimated and implicit and explicit attitudes were simultaneously entered into the path model as mediators.

Discriminability (d'). Figure 9 contains the path model diagram for the model estimating differences in discriminability. Overall, there were condition differences in IAT score, $b = 0.15$, $z = 2.26$, $p = 0.02$ but not explicit attitudes, $b = 0.003$, $z = 0.04$, $p = 0.97$. The total effect of condition on *d'* difference scores was also non-significant, $b = -.08$, $z = 1.09$, $p = 0.28$. Implicit attitudes, but not explicit attitudes ($b = 0.02$, $z = 0.32$, $p = 0.75$) were related to differences in d' for cat vs. dog trials, $b = -0.15$, $z = -2.25$, $p = 0.02$. However, there was no evidence of mediation by either implicit or explicit attitudes (indirect effect implicit attitudes: $b = -0.02$, $z = -1.58$, $p = 0.11$; indirect effect explicit attitudes: $b = 0.00$, $z = 0.04$, $p = 0.97$).

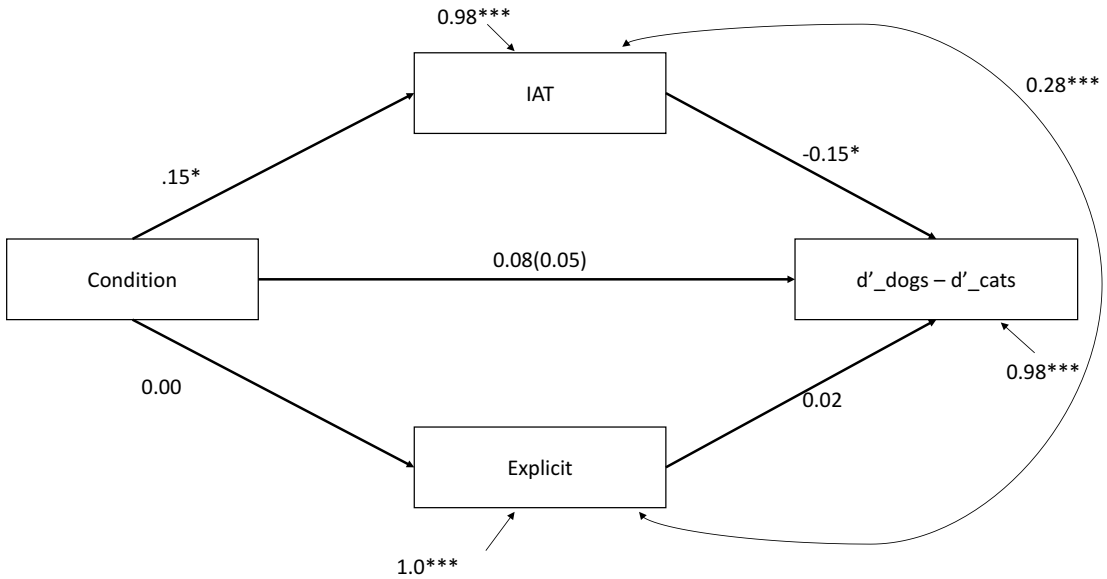


Figure 9. Experiment 2: *Multiple Mediation Model, d' Differences.* Multiple mediation path model examining the impact of condition on differences in d' for dog versus cat trials. Total effect of condition on differences in d' is represented in parentheses. Numeric values represent standardized path estimates. Neither indirect effect was significant (indirect effect implicit attitudes: $b = -0.02$, $z = -1.58$, $p = 0.11$; indirect effect explicit attitudes: $b = 0.00$, $z = 0.04$, $p = 0.97$). $^{+}p < 0.1$; *p < .05; **p< 0.01; ***p < 0.001.

Response Bias (c). Figure 10 displays the path model diagram for the model predicting response bias difference scores. This model yielded the same condition differences in implicit attitudes, $b = 0.15$, $z = 2.26$, $p = 0.02$ and non-significant condition differences in explicit attitudes, $b = 0.003$, $z = 0.04$, $p = 0.97$. Condition differences in response bias for cats vs. dogs were non-significant, $b = 0.05$, $z = 0.68$, $p = 0.50$. Implicit attitudes were marginally related to differences in response bias, $b = 0.12$, $z = 1.78$, $p = 0.08$. Explicit attitudes were not significantly related to differences in response bias, $b = 0.10$, $z = 1.49$, $p = 0.14$. Notably, there was no evidence of mediation by implicit or explicit attitudes. The indirect effect of implicit attitudes was non-significant, $b = 0.02$, $z = 1.39$, $p = 0.17$, as was the indirect effect of explicit attitudes, $b = 0.00$, $z = 0.42$, $p = 0.97$.

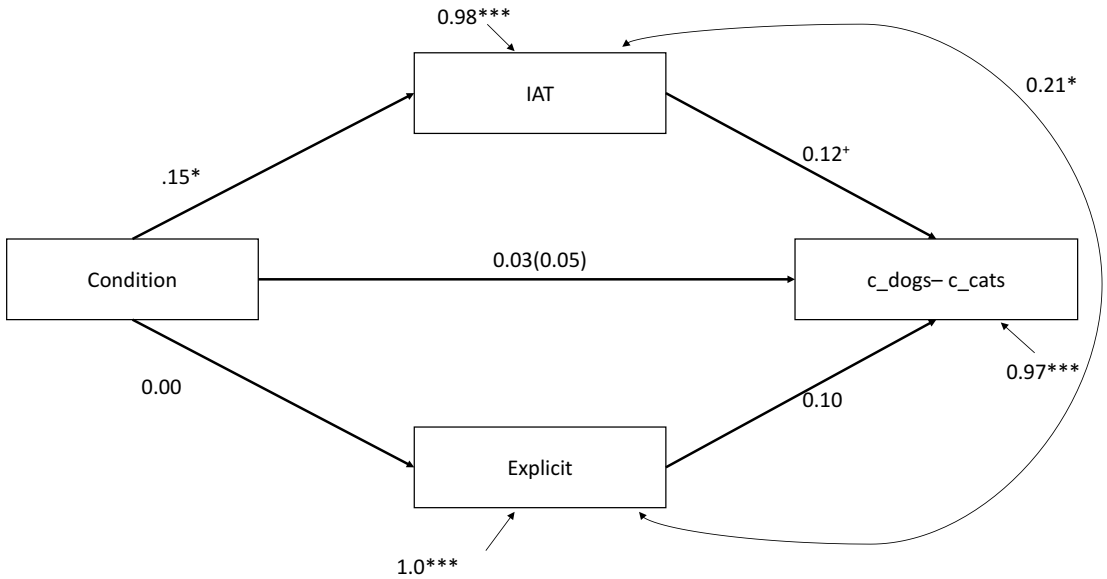


Figure 10. Experiment 2: *Multiple Mediation Model, c Differences.* Multiple mediation path model examining the impact of condition on differences in c for dog versus cat trials. Total effect of condition on differences in c is represented in parentheses. Numeric values represent standardized path estimates. Neither indirect effect was significant (indirect effect implicit attitudes: $b = 0.02$, $z = 1.39$, $p = 0.17$; indirect effect explicit attitudes $b = 0.00$, $z = 0.42$, $p = 0.97$). $^{+}p < 0.1$; *p < .05; **p< 0.01; ***p < 0.001.

Discussion

Again, there was evidence that exposure to one group and evaluative conditioning successfully manipulated implicit attitudes. Individuals who studied the dog group and saw pairings of dogs with positive words in the evaluative conditioning task demonstrated stronger implicit preferences in favor of dogs than individuals who studied cats and saw cats paired with positive words in the evaluative conditioning task. The effect of condition on implicit attitudes existed over and above both explicit attitudes and the order in which IAT blocks were presented. Unlike Experiment 1, there was no evidence that the manipulation impacted explicit attitudes.

In Experiment 2, the use of signal detection analysis allowed me to examine whether behavior was biased in two different ways. Examination of differences in discriminability for dog vs. cat trials allowed me to examine whether individuals performed better on trials involving members of their preferred group. Examination of differences in response bias allowed me to examine whether individuals demonstrated a relative preference for saving either cats or dogs regardless of whether the cue for illness was present. There was no evidence that the manipulation impacted behavior towards cats versus dogs by either metric.

In contrast to Experiment 1, there was some evidence of a correlational relationship between implicit attitudes and behavior. Regarding discriminability (d'), individuals with more positive implicit attitudes towards dogs were less able to discriminate sick from healthy animals when those animals were dogs versus cats. Additionally, participants with greater implicit preferences in favor of dogs demonstrated greater "save" response bias on dog (compared to cat) trials.

The findings from the regression analysis were paralleled by path models estimating multiple mediation. These models allowed me to examine the above questions simultaneously as well as examine whether implicit or explicit attitudes could explain any condition differences in behavior. These models again demonstrated evidence of condition differences in implicit, but not explicit attitudes and found some evidence of a relationship between implicit attitudes and behavior. However, there was no evidence that either implicit or explicit attitudes mediated any condition differences in behavior.

Such findings indicate that, although it is possible to manipulate implicit attitudes, this manipulation does not appear to impact behavior. Notably, there was evidence of correlational relationships between implicit attitudes and behavior for both discriminability and response bias. Although the relationship between implicit attitudes and response bias differences was in the expected direction, the relationship between implicit attitudes and discriminability was the opposite of the expected direction. This may be because participants who held an implicit preference for one group over the other had to work harder to ignore members of the preferred group when determining whether the animal was sick or healthy. However, it is important to replicate this finding before placing too much weight on this effect.

In sum, although the manipulation of implicit attitudes was successful, there was little evidence that implicit (or explicit) attitudes caused behavior in the serial pet rescue game. Although implicit attitudes were correlated with behavior, the failure of the manipulation to impact behavior (directly or indirectly through implicit attitudes) rules out a causal interpretation. The results of Experiment 2 largely mirror those of Experiment 1 with slight differences in the effect of the manipulation on explicit attitudes and with the correlational relationship of implicit and explicit attitudes.

There are several possibilities for the lack of a causal effect on behavior. First, it is possible the magnitude of condition differences in implicit attitudes was simply not large enough to impact behavior. Second, it could be that the manipulation of implicit attitudes changed some aspect of IAT scores other than automatic associations. For example, since the pairings of the fish and valenced words in the evaluative conditioning task were supraliminal this task could have changed conscious, propositional knowledge about the two groups of fish and such propositional knowledge could have been reflected in IAT d-scores (De Hower, 2006). Finally, it is possible that implicit attitudes towards groups do not actually cause behavior towards individual group members or that the relative contribution of implicit attitudes to behavior is so small that it is very difficult to detect.

Experiment 3 examines an approach to examining the first explanation for the null results from Experiments 1 and 2. Specifically, I examined whether the addition of a new task to my manipulation might increase the effect of condition on implicit attitudes. Since this manipulation is decidedly explicit in its presentation, I first examined whether this additional task increased the size of my manipulation of implicit attitudes and, secondly, whether it did so though changes in explicit attitudes.

In this study (and future studies) I return to the use of fish stimuli from Experiment 1 for three reasons. First, the use of dog and cat stimuli did not yield larger condition differences in implicit attitudes or behavior. Second, the use of dog and cat stimuli did not appear to increase participant engagement in the tasks. An examination of the level of self-reported focus in the behavioral tasks from the first two studies indicated that participants reported feeling significantly more focused in Study 1 compared to Study 2, $M_{Study1} = 81.79$, $M_{Study2} = 68.70$,

$t(361) = 5.62$, $p < .001$. Third, the fish stimuli were more carefully selected and have been equated on liking based on pre-test ratings.

CHAPTER IV: Experiment 3

Experiment 3 was designed to examine whether the addition of a narrative to the implicit attitudes manipulation would increase the magnitude of condition differences in implicit attitudes. Previous research has demonstrated that implicit attitudes can be created through both automatic processes and concrete learning. Gregg, Seibt and Banaji (2006) demonstrated that participants who read a vignette portraying one group in a positive light and another group in a negative light developed implicit preferences based on this manipulation (see also Cone & Ferguson, 2015; Mann & Ferguson, 2015). As such, I opted to use a similar narrative to increase the effect conditioning on implicit attitudes. It is also possible that the addition of a manipulation that relies on such deliberative processes may alter explicit attitudes or may change implicit attitudes through changes in explicit attitudes (Gawronski & Bodenhausen, 2006). As such, before running the full study with a new manipulation, I opted to first test the impact of the narrative on implicit and explicit attitudes.

Method

Participants and Design. One-hundred participants (48 female, 52 male, 0 other gendered, M_{age} = 34.79, age range: 20-71) from Amazon's Mechanical Turk website participated in this study in exchange for monetary compensation. Seven participants were excluded for responding in ways that indicated they were not following instructions on more than two occasions[21]. This left a final sample of 93 participants (47 female, 46 male, 0 other gendered, M_{age} = 35.47, age range: 20-71)

[21] Analyses were also conduced excluding participants who demonstrated lack of compliance on 2 or more tasks (mirroring other studies in this prospectus), but the pattern of results was the same no matter exclusion level and this broader exclusion criterion resulted in exclusion of 13 participants. I opted to use the more narrow exclusion criteria to retain more of the sample since I had fewer participants to begin with.

The design was a 2 (condition: orange-good vs. purple-good) X 2 (vignette: no vignette vs. vignette) between-subjects factorial design.

Materials. ***Manipulation of Implicit Attitudes.*** All participants completed the sorting task, exposure to one group of fish and evaluative conditioning task as outlined in Experiment 1. In addition, half of participants were exposed to a narrative about the two fish groups that portrayed one fish group in a positive light and the other fish group in a negative light (See Appendix J for the narrative). For participants who viewed the narrative, they learned that the orange group of fish (in the orange-good condition) were beneficial to the environment and improved the quality of American beaches and coastal waters. Furthermore, (in the orange-good condition) the purple fish were described as an invasive species that was harmful to the environment, polluted coastal waters and beaches and was decimating the population of the beneficial group of fish. Participants in the purple-good viewed the same vignette with the opposite characterization of the two fish groups. After reading their assigned vignette, participants answered two multiple choice questions to test their comprehension of the vignette[22]. Additionally, participants completed two open response items in which they were asked to write what they knew about each group of fish.

Measurement of Implicit Attitudes. The measure of implicit attitudes was identical to the measure used in Experiment 1, an evaluative orange vs. purple fish IAT.

Measurement of Explicit Attitudes. The measure of explicit attitudes was identical to Experiment 2 except that the attitude objects were the groups of fish rather than cats and dogs.

[22] Twelve participants (25% of those asked) missed at least one of the comprehension questions. Effects of condition on IAT and explicit attitudes are marginally higher for those who answered both vignette questions correctly. However, the analysis reported does not exclude participants solely based on responses to these questions.

The attitude items were standardized and averaged to form a single explicit attitudes scale (α = 0.88).

Procedure. The procedure was the same as Experiment 2 with two exceptions. First, participants did not complete a behavioral measure. Second, participants who were assigned to the vignette-present condition read the vignette about the two groups of fish before completing the evaluative conditioning task. Participants in the vignette-absent condition did not see this vignette.

Results

IAT score was regressed on contrast-coded condition (orange-good = 0.5, purple-good = -0.5), vignette condition (no-vignette = -0.5, vignette = 0.5) and their interaction. Figure 11 depicts condition differences in IAT scores. There was no evidence of a significant vignette effect across condition, b = -0.02, $t(89)$ = -0.20, p = 0.84, R^2 =0.00 . However, implicit attitudes did differ by evaluative condition, b = 0.51, $t(89)$ = 5.88, $p < .001$, R^2= 0.26. Participants in the orange-good condition showed greater implicit preferences for orange fish ($M_{\text{orange-good}}$ = 0.37) compared to the purple-good condition ($M_{\text{purple-good}}$ = -0.14). Notably, participants in the orange-good condition demonstrated a significant implicit preference for orange fish, b = 0.37, $t(89)$ = 5.67, $p < .001$, R^2 =0.27, and participants in the purple-good condition demonstrated a significant implicit preference for purple fish, b = -0.14, $t(89)$ = -2.47, p = .02, R^2 = 0.06 . This effect of evaluative condition was qualified by a significant evaluative condition X vignette interaction, b = 0.44, $t(89)$ = 2.50, p = 0.01, R^2 = 0.05. Although significant in both vignette conditions (p's <= 0.02), evaluative condition differences were strongest for participants who viewed the narrative.

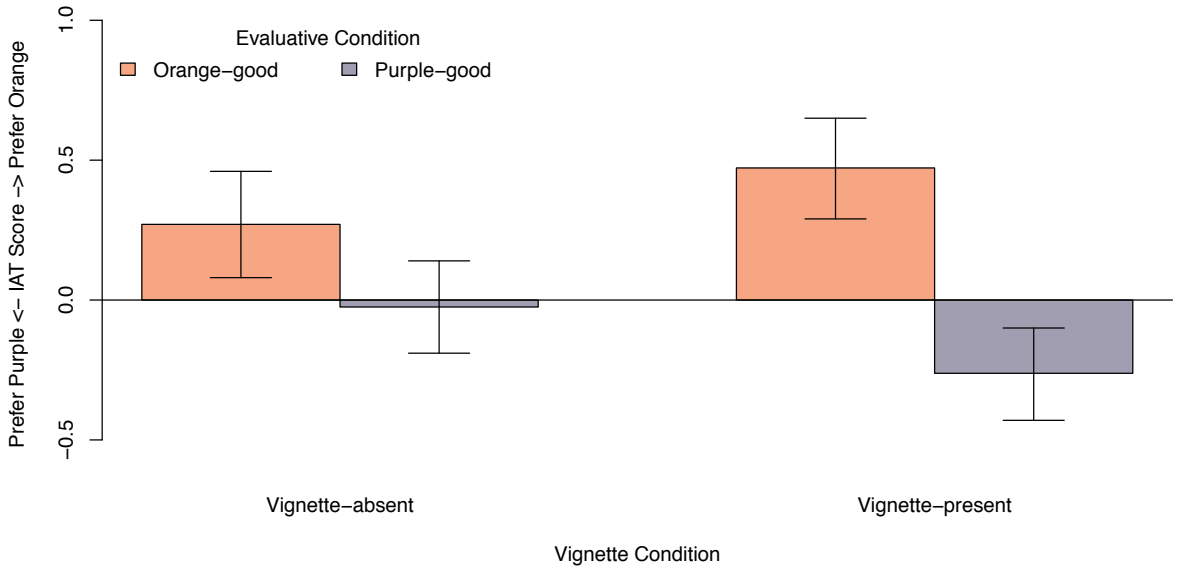


Figure 11. *Condition Differences in Implicit Attitudes, Experiment 3*. Average IAT scores by evaluative condition and vignette condition. Higher values on the y-axis indicate a greater preference for dogs over cats. Zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

There was also evidence that the narrative manipulation had an impact on explicit attitudes (see Figure 12). Again, there was evidence of significant evaluative condition differences across vignette condition, $b = 1.11$, $t(89) = 11.37$, $p < .001$, $R^2 = 0.43$. Participants in the orange-good condition reported stronger explicit preferences for the orange fish than participants in the purple-good condition. Additionally, this effect was qualified by a significant evaluative condition X vignette interaction, $b = 1.80$, $t(89) = 9.20$, $p < .001$, $R^2 = 0.28$. Participants who viewed the vignette demonstrated significant group differences in explicit attitudes, $b = 2.01$, $t(89) = 16.65$, $p < .001$, $R^2 = 0.43$. However, participants who did not view the vignette did not demonstrate significant group differences in explicit attitudes, $b = -0.21$, $t(89) = 1.53$, $p = 0.13$, $R^2 = 0.004$.

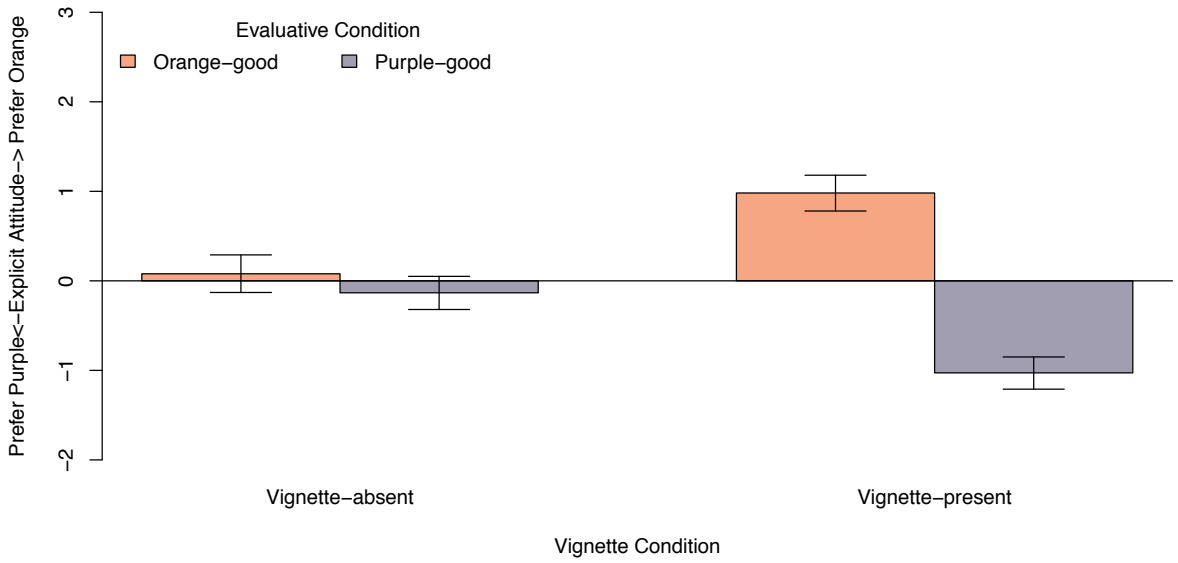


Figure 12. *Condition Differences in Explicit Attitudes, Experiment 3*. Average explicit attitude scores by evaluative condition and vignette condition. Higher values on the y-axis indicate a greater preference for dogs over cats. Zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

Although implicit and explicit attitudes were not significantly related in this experiment, b = -0.00, $t(91)$ = -0.08, p = 0.93, R^2 = 0.00, I also examined whether condition differences in implicit and explicit attitudes persisted after controlling for other the attitude measure. Controlling for explicit attitudes attenuated the evaluative condition X vignette interaction on implicit attitudes, b = 0.001, $t(88)$ = 0.01, p = 0.99, R^2 = 0.00 (this was true whether or not IAT block order was controlled for, in the presented model it is not). However, there was still some evidence of overall condition differences in implicit attitudes, b = 0.24, $t(88)$ = 1.84, p = 0.07, R^2 = .05, in the expected direction (see Figure 4). In contrast, controlling for implicit attitudes did not attenuate the condition X vignette interaction effect on explicit attitudes, b = 2.03, $t(88)$ = 10.11, $p < 0.001$, $R^2_{partial}$ = 0.54 (again, this remained true when IAT block order was added to the model).

Discussion

Experiment 3 demonstrated that a larger manipulation of implicit attitudes can be achieved by adding a narrative to the implicit attitude manipulation from Experiments 1 and 2. The larger manipulation of implicit attitudes appears to be driven by condition differences in explicit attitudes. Such results may reflect the fact that the IAT is not a process pure measure of implicit attitudes (Mierke & Klauer, 2003). That is, the IAT does not track only implicit attitudes, but may also track more controllable processes such as explicit attitudes.

It is possible that this larger manipulation of implicit attitudes may make it easier to detect a causal relationship between implicit attitudes and behavior. It is also possible that implicit and explicit group attitudes must be directionally consistent for implicit attitudes to influence behavior. In Experiment 4, I examine whether this larger manipulation of implicit attitudes, which also impacts explicit attitudes, results in a detectable causal effect of implicit attitudes. As with the previous experiments, I continue to measure explicit attitudes to better understand the relative contribution of both automatic and controlled processes to behavior.

CHAPTER V: Experiments 4a & 4b

Experiment 3 demonstrated that it is possible to amplify the effect of the manipulation on implicit attitudes through the addition of a narrative vignette. Notably, this increased effect of condition was attenuated after controlling for explicit attitudes. Experiment 4a was designed to examine whether this increased manipulation impacted behavior and whether this effect appeared to be explained by the effect of the manipulation on implicit or explicit attitudes (or both). Experiment 4b was a direct replication of Experiment 4a.

Method

Participants and Design. Two-hundred ninety-eight participants (136 female, 161 male, 1 other gendered, M_{age} = 34.63, age range: 18 - 74) were recruited from Amazon's Mechanical Turk website and completed Experiment 4a. Twelve of these participants responded in ways that indicated they were not paying attention on at least two occasions and were excluded. The final sample for Experiment 4a was 286 participants. Two-hundred ninety-seven participants (152 female, 143 male, 2 other gendered, Mage = 35.9, age range: 18-70), who had not completed Experiment 4a, were recruited from Amazon's Mechanical Turk website and completed Experiment 4b. For Experiment 4b, twenty-one participants had response patterns that indicated they were not paying attention on at least two occasions during the study. Consistent with previous studies, these participants were excluded from analysis. This left a final sample of 276 participants in Experiment 4b. In total, the final sample for Experiment 4 was 562 participants (278 female, 281 male, 3 other gendered, M_{age} = 35.55, age range: 18-74).

Each experiment was a 2 (evaluative condition: orange-good vs. purple-good) X 2 (vignette: present or absent) between-subjects factorial design.

Materials. Materials were identical to those used in Experiment 2 with the addition of the vignette-present versus absent manipulation used in Experiment 3 and the use of fish stimuli from Experiments 1 and 3.

Procedure. The procedure was the same as that used in Experiment 3 with two changes. First, participants completed implicit and explicit attitude measures in a counterbalanced order. Second, participants completed a behavioral measure, the pet rescue game from Experiment 2 (with fish stimuli rather than dogs and cats), after completing the attitude measures (followed by questions probing for contingency awareness, awareness of the study purpose and demographics).

Results

To analyze data in this experiment, signal detection metrics for discriminability (*d'*) and response bias (*c*) were estimated separately (using the same technique from Experiment 2) for orange and purple fish trials in the behavioral task[23]. Next, difference scores were calculated by subtracting the *d'* (or *c*) value for purple fish trials from the d' (or c) value for orange fish trials. This yielded two outcome measures. Discriminability differences estimated the degree to which a participant was better able to differentiate sick from healthy fish on orange-fish compared to purple-fish trials. The second outcome, response bias differences, tracked the degree to which a participant would bias responding in favor of saving more for orange than for purple fish.

Once these outcomes were calculated, linear models and path models were used to estimate condition differences in implicit and explicit attitudes, relationships between attitudes

[23] Again, multiple models were tested examining different criteria for different types of trials. Estimates of response bias and discriminability across different estimation techniques were correlated at r = 0.99 so only the results using hand calculated d' and c are reported.

and behavior and condition differences in behavior. Path models also allowed me to estimate statistical mediation of any condition effects on behavior by both implicit and explicit attitudes.

For the linear models, implicit attitudes and explicit attitudes were regressed on contrast-coded evaluative condition (orange-good = 0.5, purple-good = -0.5), vignette presence (vignette-present = 0.5, vignette-absent = -0.5), experiment number[24] (Experiment 4b = 0.5, Experiment 4a = -0.5) and their interactions to test for condition differences in attitudes. To test for condition differences in behavior, the same models were used, but discriminability and response bias difference scores served as the outcome measures. To examine the relationship between implicit and explicit attitudes and behavior, differences in *d'* and *c* were regressed on IAT score (mean centered) and explicit attitudes (mean centered).

Condition Differences in Attitudes. Figure 13 depicts condition differences in implicit attitudes. Across study and vignette condition, there were significant evaluative condition differences in implicit attitudes, $b = 0.29$, $t(553) = 10.19$, $p < 0.001$, $R^2_{\text{partial}} = 0.16$. Participants in the orange-good condition demonstrated significant implicit preferences for orange fish over purple fish, $b = 0.22$, $t(553) = 11.03$, $p < 0.001$, $R^2_{\text{partial}} = 0.18$. In contrast, participants in the purple-good condition demonstrated small but significant implicit preferences for purple over orange fish, $b = -0.07$, $t(553) = -3.47$, $p < 0.001$, $R^2_{\text{partial}} = 0.02$.

The effect of evaluative condition on implicit attitudes was qualified by an interaction with vignette condition, $b = 0.26$, $t(553) = 4.63$, $p < 0.001$, $R^2_{\text{partial}} = 0.04$. Although significant in

[24] Experiment number was not a theoretically meaningful predictor, but some effects presented in this section were moderated by experiment. These effects do not drastically change the conclusions drawn and add further complication to the reporting of these models. Therefore, linear model results, presented below are based on models that include experiment number as a moderator. Information on the moderating effects of experiment number are included in ancillary analyses.

both the vignette-present and vignette-absent conditions, the effect of evaluative condition was stronger in the vignette-present condition, $b = 0.42$, $t(553) = 10.38$, $p < 0.001$, $R^2_{partial} = 0.16$, compared to the vignette-absent condition, $b = 0.16$, $t(553) = 3.97$, $p < 0.001$, $R^2_{partial} = 0.03$. Implicit and explicit attitudes were positively related to each other, $b = 0.20$, $t(559) = 6.55$, $p < 0.001$, $R^2 = 0.07$. So, I examined the condition differences in implicit attitudes after controlling for explicit attitudes. The evaluative condition effect across vignette condition remained statistically significant, $b = 0.22$, $t(552) = 5.73$, $p < 0.001$, $R^2_{partial} = 0.06$, as did the evaluative condition X vignette interaction, $b = 0.20$, $t(552) = 3.12$, $p = 0.002$, $R^2_{partial} = 0.02$.

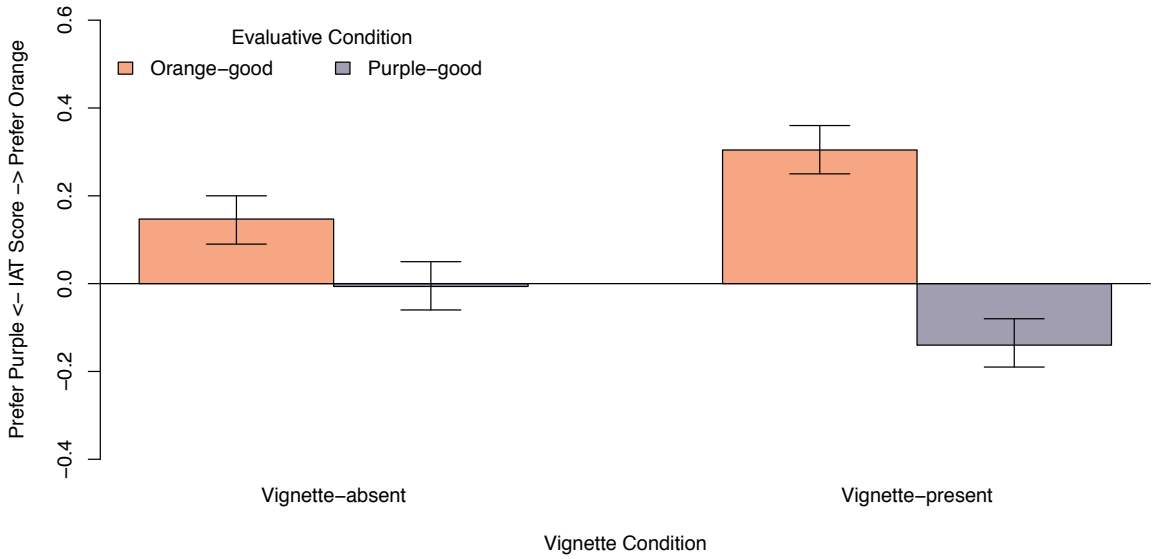


Figure 13. *Condition Differences in Implicit Attitudes, Experiment 4*. Average IAT scores by evaluative condition and vignette condition. Higher values on the y-axis indicate a greater preference for orange over purple fish. Zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

Evaluative condition, vignette condition, experiment number and their interactions were also used to predict explicit attitudes (see Figure 14). Across experiment and vignette condition, there was a significant effect of evaluative condition, $b = 0.22$, $t(552) = 5.34$, $p < 0.001$, $R^2_{partial} =$

0.05. Participants in the orange-good condition reported significant explicit preferences for orange fish, $b = 0.17$, $t(554) = 10.45$, $p < 0.001$, $R^2_{partial} = 0.16$, whereas participants in the purple-good condition reported explicit preferences for purple fish, $b = -0.42$, $t(554) = -24.63$, $p < 0.001$, $R^2_{partial} = 0.52$. This effect was qualified by a two-way evaluative condition X vignette condition interaction, $b = 0.56$, $t(554) = 11.78$, $p < 0.001$, $R^2_{partial} = 0.20$. The evaluative condition effects on explicit attitudes were larger for participants in the vignette-present condition compared to the vignette absent condition. These effects remained significant after controlling for implicit attitudes, ts > 5.7, ps < 0.001.

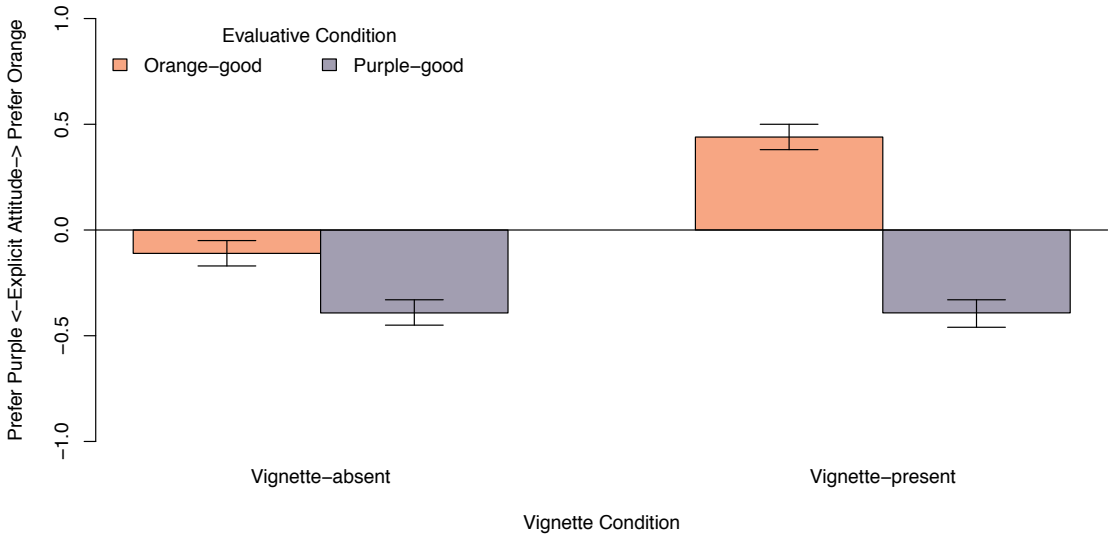


Figure 14. *Condition Differences in Explicit Attitudes, Experiment 4*. Average explicit attitude scores by evaluative condition and vignette condition. Higher values on the y-axis indicate a greater preference for orange over purple fish. Zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

Condition Differences in Behavior. ***Discriminability (d').*** Table 4 presents mean values of d' and c by vignette condition, evaluative condition, and trial type. Overall, participants were able to successfully discriminate healthy from unhealthy fish on both orange-fish ($M_{d'} = 2.51$, $t(561) = 62.25$, $p < 0.001$, $R^2 = 0.84$) and purple-fish trials ($M_{d'} = 2.44$, $t(561) = 65.39$, $p < 0.001$,

R^2 = 0.88). Across study and vignette condition, there was a significant effect of evaluative condition, b = 0.11, $t(554)$ = 2.27, p = 0.02, $R^2_{partial}$ = 0.01. Whereas participants in the orange-good condition demonstrated greater discriminability on orange fish trials, b = 0.13, $t(554)$ = 3.83, $p < 0.001$, $R^2_{partial}$ = 0.03, participants in the purple-good condition did not demonstrate differences in discriminability for the two types if trials in the rescue game, b = 0.02, $t(554)$ = 0.57, p = 0.57, $R^2_{partial}$ = 0.00. No other effects of condition or the evaluative condition X vignette condition interaction emerged, all ps > 0.21.

Response Bias (c). Participants' responses on the rescue game were biased in favor of saving for both orange (M_c = 0.04, $t(561)$ = 3.68, p <0.001, R^2 = 0.02) and purple fish (M_c = 0.06, $t(561)$ = 5.41, p <0.001, R^2 = 0.05). A significant effect of evaluative condition also emerged for the response bias outcome, b = 0.18, $t(554)$ = 5.12, p = < 0.001, $R^2_{partial}$ = 0.05. In the orange-good condition, response bias in favor of saving was significantly greater on orange fish trials compared to purple fish trials, b = 0.07, $t(554)$ = 3.02, p = 0.003, $R^2_{partial}$ = 0.02. In the purple-good condition, participants demonstrated a significantly greater "save" response bias on purple fish trials compared to orange fish trials, b = -0.10, $t(554)$ = -4.21, $p < 0.001$, $R^2_{partial}$ = 0.03. Additionally, vignette condition moderated the evaluative condition effect, b = 0.33, $t(554)$ = 4.81, $p < 0.001$, $R^2_{partial}$ = 0.04. The effect of evaluative condition on behavior was significant in the vignette-present condition, b = 0.34, $t(554)$ = 5.97, $p < 0.001$, $R^2_{partial}$ = 0.06, but not in the vignette-absent condition, b = 0.01, $t(554)$ = 0.22, p = 0.83, $R^2_{partial}$ = 0.00.

Table 4

Signal Detection Statistics by Trial Type and Condition, Experiment 4

		Vignette-absent		Vignette-present	
		Orange-good	Purple-good	Orange-good	Purple-Good
Discriminability (d')	Orange trials	2.77 (.92)	2.46 (.97)	2.57 (.87)	2.24 (1.02)
	Purple trials	2.65 (.81)	2.38 (.92)	2.43 (.80)	2.28 (.97)
Response Bias (c)	Orange trials	.10 (.30)	.05 (.21)	.13 (.22)	-.09 (.46)
	Purple trials	.12 (.25)	.09 (.22)	-.04 (.35)	.09 (.26)

Note. Mean (SD) values of d' and c separated by vignette condition, evaluative condition and trial type, Experiment 4. For d', larger positive values indicate better ability to discriminate sick from healthy animals. For c, larger positive values indicate a more liberal bias in favor of "saving" regardless of whether the animal is sick or healthy.

Relationship Between Attitudes and Behavior. ***Discriminability (d').*** Differences in discriminability were regressed on IAT score, experiment number and their interaction. Across study, there was a marginally significant relationship between implicit attitudes and behavior, $b = 0.16$, $t(557) = 1.90$, $p = 0.06$, $R^2_{partial} = 0.01$. Participants with stronger implicit preferences for orange over purple fish demonstrated greater discriminability on orange fish compared to purple fish trials (see Figure 15). This effect did not depend on study, $b = 0.06$, $t(557) = 0.35$, $p = 0.73$, $R^2_{partial} = 0.00$. The relationship between IAT score and differences in discriminability was attenuated after controlling for explicit attitudes, $b = 0.12$, $t(556) = 1.33$, $p = 0.19$, $R^2_{partial} = 0.00$. To examine the relationship between explicit attitudes and behavior, explicit attitude scores, experiment number, and their interaction were used to predict differences in *d'*. Explicit attitudes were also marginally related to differences in discriminability, $b = 0.08$, $t(558) = 1.73$, $p = 0.08$, $R^2_{partial} = 0.01$, such that participants with stronger explicit preferences for orange fish were better able to discriminate sick from healthy orange fish compared to purple fish than participants with

weaker explicit orange fish preferences. This effect was also attenuated after controlling for implicit attitudes, $b = 0.06$, $t(556) = 1.13$, $p = 0.26$, $R^2_{partial} = 0.00$.

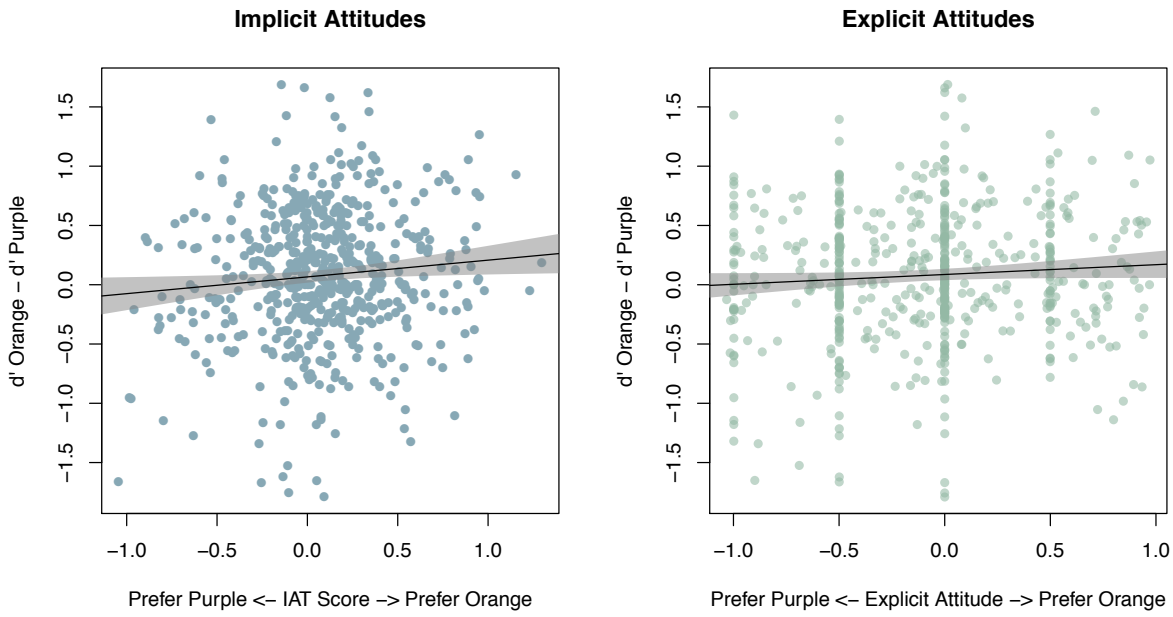


Figure 15. *Experiment 4: Attitude Behavior Relationship, d' Differences*. Relationship between implicit attitudes and differences in d' (left panel) and explicit attitudes and differences in d' (right panel) for orange fish versus purple fish trials. Values greater than 0 on the y-axis indicate participants were better able to discriminate healthy from unhealthy orange fish compared to purple fish while negative values indicate the reverse. Gray bands represent the 95% confidence interval around the line of best fit (black line).

Response Bias (c). Differences in response bias were also regressed on IAT score, experiment and their interaction. Across experiment, implicit attitudes were significantly related to differences in response bias, $b = 0.29$, $t(557) = 4.58$, $p < 0.001$, $R^2_{partial} = 0.04$. Participants with more positive implicit preferences for orange fish demonstrated more biased responses in favor of saving orange fish compared to purple fish (see Figure 16). The effect of IAT score, $b = 0.16$, $t(556) = 3.39$, $p < 0.001$, $R^2_{partial} = 0.20$, remained after controlling for explicit attitudes.

Explicit attitudes were also significantly related to differences in response bias, b = 0.18, t(558) = 4.90, p < 0.001, $R^2_{partial} = .04$. Participants who self-reported more positive attitudes

towards orange fish demonstrated greater bias in favor of saving orange compared to purple fish. This relationship remained after controlling for implicit attitudes, $b = 0.14$, $t(556) = 3.76$, $p < 0.001$, $R^2_{partial} = 0.02$.

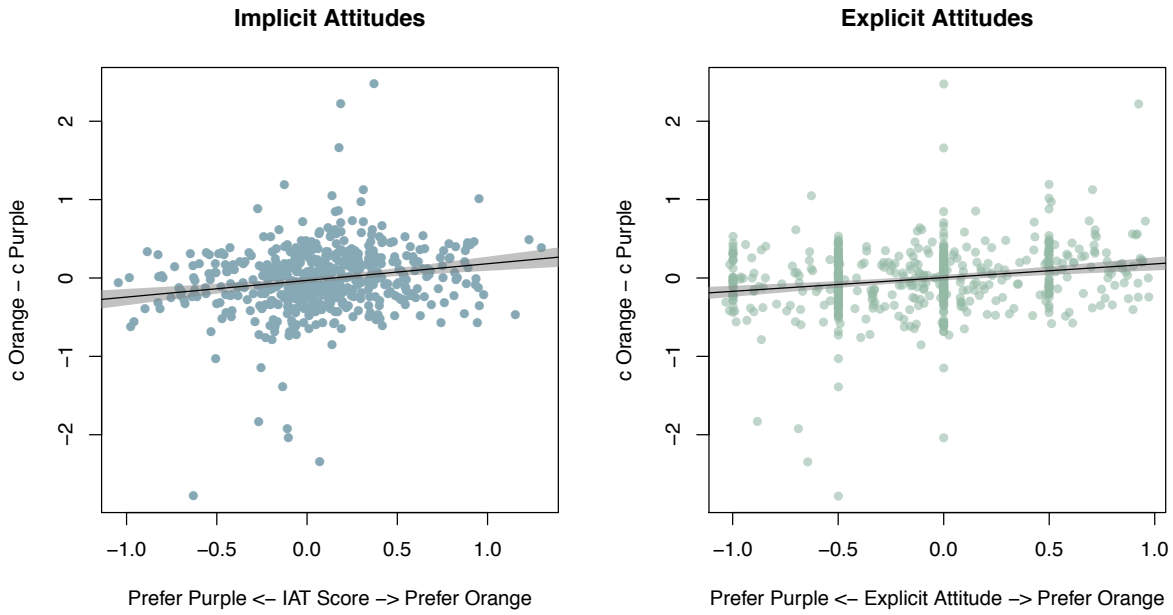


Figure 16. *Experiment 4: Attitude Behavior Relationship, c Differences*. Relationship between implicit attitudes and differences in c (left panel) and explicit attitudes and differences in c (right panel) for orange fish versus purple fish trials. Values greater than 0 on the y-axis indicate participants were respond in favoring of saving orange fish compared to purple fish while negative values indicate the reverse. Gray bands represent the 95% confidence interval around the line of best fit (black line).

Mediation Models. To test multiple mediation by implicit and explicit attitudes, I used the combined data from Experiments 4a and 4b (see Appendix K for descriptions of differences in effects between the two studies) and two path models were estimated: one for each outcome[25]. Each path model allowed me to examine two relevant mediational pathways. The first path examined evaluative condition differences in the outcome, across vignette condition, and

[25] Experiment was not included as a factor because the pattern of results was the same across the two studies and for ease of interpretation.

whether implicit or explicit attitudes mediated this effect. This mediation question is parallel to the mediational models examined in Experiments 1 and 2. The second path examined the impact of the evaluative condition X vignette condition interaction on the outcome measure and examined whether implicit and/or explicit attitudes mediated this effect. That is, I could examine whether the addition of the vignette to the manipulation resulted in larger effects on behavior and whether these larger effects could be explained by either implicit or explicit attitudes (or both). Figures 17 and 18 present the full path models with standardized coefficients for the two outcomes: differences in d' (Figure 17) and differences in c (Figure 18).

Discriminability (d'). Overall, there was evidence that evaluative condition impacted both implicit and explicit attitudes. Mirroring the linear model analyses, participants in the orange-good condition demonstrated more positive implicit attitudes towards orange fish (relative to purple fish) compared to individuals in the purple-good condition, $b = 0.39$, $z = 11.64$, $p < 0.001$. Participants in the orange-good condition also demonstrated stronger explicit preferences for orange fish compared to participants in the purple-good condition, $b = 0.55$, $z = 22.63$, $p < 0.001$. Evaluative condition also impacted the d' difference outcome, $b = 0.10$, $z = 2.37$, $p = 0.02$. Participants in the orange-good condition demonstrated higher discriminability on orange-fish trials relative to purple-fish trials to a greater degree than participants in the purple-good condition. Notably, there was no evidence that the evaluative condition effects on *d'* differences were mediated by either implicit or explicit attitudes, indirect effect- implicit attitudes: $b = 0.02$, $z = 1.27$, $p = 0.20$, indirect effect- explicit attitudes: $b = 0.02$, $z = 0.52$, $p = 0.60$.

As shown in Figure 17, the total evaluative condition X vignette condition effect on implicit attitudes was also statistically significant, $b = 0.19$, $z = 5.15$, $p < 0.001$, indicating that

the effect of evaluative condition was larger for participants in the vignette-present compared to the vignette-absent condition. A similar effect was found on explicit attitudes, $b = 0.27$, $z = 9.19$, $p < 0.001$. The total effect of the evaluative condition X vignette condition interaction on differences in d' was nonsignificant, $b = 0.06$, $z = 1.40$, $p = 0.16$, and there was no evidence that either implicit or explicit attitudes mediated this effect, indirect effect-implicit attitudes: $b = 0.01$, $z = 0.52$, $p = 0.60$; indirect effect-explicit attitudes: $b = 0.01$, $z = 0.52$, $p = 0.60$.

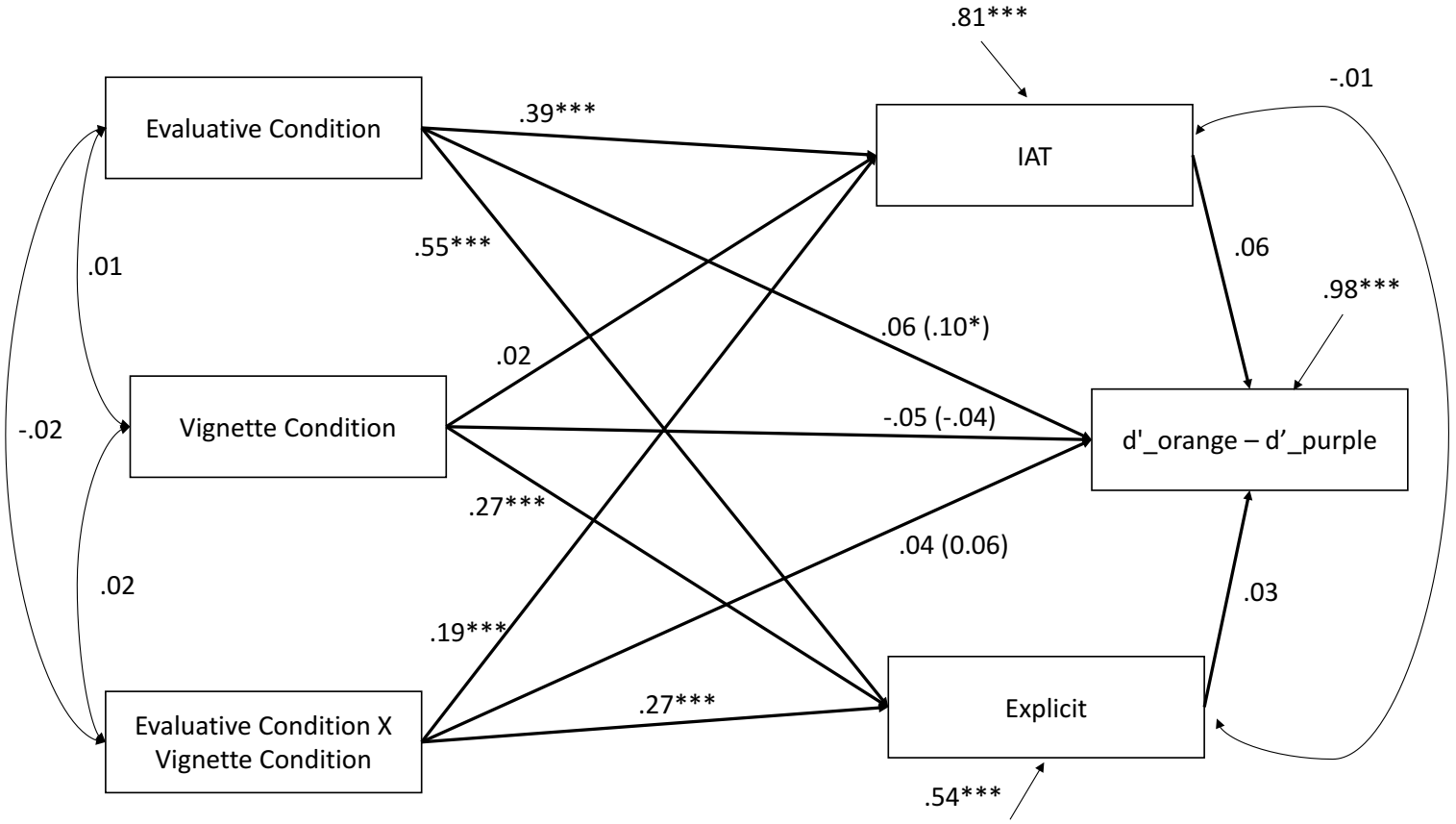


Figure 17. *Experiment 4: Mediated Moderation Model, d' Differences*. Multiple mediated moderation path model examining the impact of evaluative condition, vignette condition and their interaction on differences in d' for orange fish versus purple fish trials. Total effect of condition on differences in d' is represented in parentheses. Numeric values represent standardized path estimates. Indirect effects of implicit and explicit attitudes on evaluative condition effects and on the evaluative condition X vignette condition interaction were all non-significant, *p*s > .20. $^{+}$p < 0.1; *p < .05; **p< 0.01; ***p < 0.001.

Response Bias (c). The effects of evaluative condition on implicit and explicit attitudes are identical to those presented in the path model for the discriminability outcome, described earlier, because the mediator equations for the two path models are identical. There was also a significant total effect of evaluative condition on differences in response bias, $b = 0.21$, $z = 5.28$, $p < 0.001$. Participants in the orange-good condition biased responding in favor of "saving" more on the orange-fish trials compared to participants in the purple-good condition. Additionally, implicit (but not explicit) attitudes were uniquely related to differences in response bias, implicit: $b = 0.09$, $z = 2.04$, $p = 0.04$; explicit: $b = 0.06$, $z = 1.15$, $p = 0.25$. Participants with more positive implicit preferences for orange fish tended to show a greater response bias in favor of saving orange (compared to purple) fish. Implicit attitudes significantly mediated the evaluative condition effects on the response bias outcome, $b = 0.04$, $z = 2.01$, $p = 0.05$. However, there was no evidence that explicit attitudes mediated this effect, $b = 0.04$, $z = 1.15$, $p = 0.25$.

As with the evaluative condition effects on implicit and explicit attitudes, the evaluative condition X vignette condition interactions on implicit and explicit attitudes were identical to the discriminability path models. Further, the total effect of the evaluative condition X vignette condition interaction on behavior was statistically significant, $b = 0.19$, $z = 4.86$, $p < 0.001$. The evaluative condition effects on differences in response bias were larger for individuals in the vignette-present relative to the vignette-absent condition. Tests of the indirect effects yielded a marginally significant indirect effect for implicit attitudes, $b = 0.02$, $z = 1.89$, $p = 0.06$, providing some indication that implicit attitudes mediated the interaction effect on response bias. There was no evidence of mediation by explicit attitudes, $b = 0.02$, $z = 1.14$, $p = 0.25$.

I further examined the marginally significant mediation of the evaluative condition X vignette condition interaction by estimating the same mediated moderation model with dummy coded predictors for vignette condition variable. This allowed me to examine the evaluative condition effects separately for the vignette-present and vignette-absent conditions while retaining the full sample size in my analysis. For the vignette-absent condition, the indirect effect of implicit attitudes on evaluative condition difference in the response bias outcome was marginally significant, $b = 0.02$, $z = 1.80$, $p = 0.07$. However, for the vignette present condition, the indirect effect of implicit attitudes was stronger and statistically significant, $b = 0.05$, $z = 2.01$, $p = 0.04$.

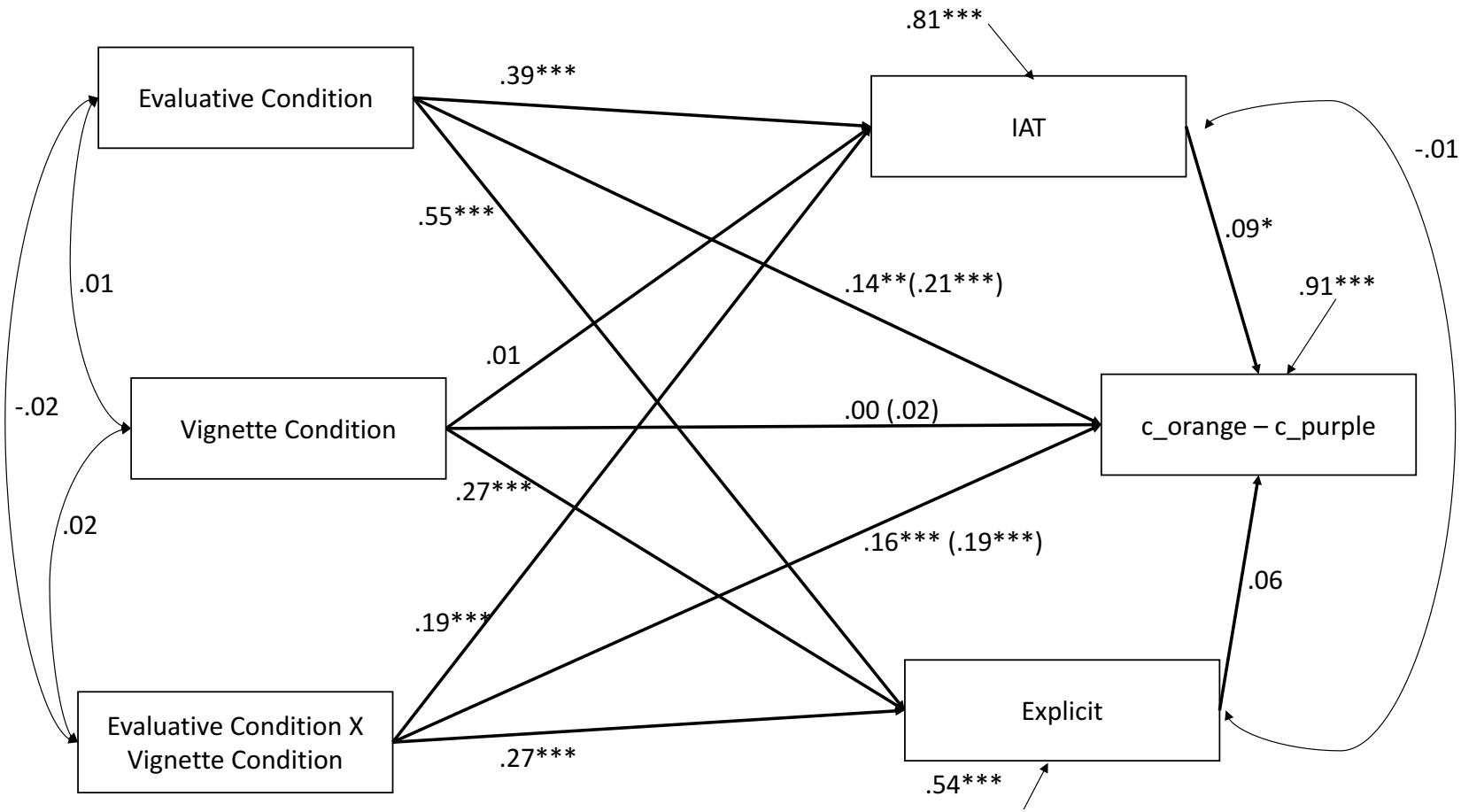


Figure 18. *Experiment 4: Mediated Moderation Model, c Differences*. Multiple mediated moderation path model examining the impact of evaluative condition, vignette condition and their interaction on differences in d' for orange fish versus purple fish trials. Total effect of condition on differences in c is represented in parentheses. Numeric values represent standardized path estimates. Implicit attitudes significantly mediated evaluative condition differences in response bias difference, $b = .04$, $z = 2.01$, $p = 0.04$. Implicit attitudes also marginally mediated the evaluative condition X vignette condition interaction on differences in c, $b = 0.02$, $z = 1.90$, $p = .06$. $^{+}p < 0.1$; *p < .05; **p< 0.01; ***p < 0.001.

Discussion

Results from Experiments 4a and 4b provided further evidence that it is possible to create implicit attitudes using an evaluative conditioning paradigm. Further, the results from Experiment 3 were replicated, providing further support that the addition of a vignette to the manipulation increased the effect of conditioning on implicit attitudes. Also replicating Experiment 3, evaluative condition and vignette condition interacted to impact explicit attitudes as well. However, in Experiments 4a and 4b, condition differences in explicit attitudes did not fully explain the evaluative condition X vignette condition interaction on implicit attitudes.

In addition to condition effects on attitudes, evaluative condition was also related to both differences in discriminability and differences in response bias. However, only condition differences in response bias were mediated by implicit attitudes. There was no evidence that explicit attitudes mediated condition effects on either d' differences or response bias differences. Finally, there was evidence that the evaluative condition X vignette condition interaction impacted differences in response bias (but not discriminability). Further, this interaction was weakly mediated by implicit, but not explicit, attitudes.

In sum, Experiments 4a and 4b provide the first evidence that implicit group attitudes cause behavior towards individual group members. This evidence should be considered preliminary when contrasted with the null results from Experiments 1 and 2. However, it may be that the increased size of the implicit attitude manipulation did, in fact, increase my ability to detect a causal relationship between implicit attitudes and behavior. One reason that it may be so difficult to detect a causal relationship between implicit attitudes and behavior is that implicit attitudes are theorized to cause behavior under circumstances that increase reliance on automatic

processes. To examine whether lack of control of such theory-based moderators may explain the mixed results from Experiments 1, 2, 4a and 4b, I conducted Experiment 5.

CHAPTER VI: Experiment 5

The purpose of Experiment 5 was to determine whether a set of moderators theorized to increase reliance on automatic processes would increase the size of the relationship between implicit group attitudes and behavior towards individual group members. To do this, participants were randomly assigned to develop implicit preferences for one of two groups. Participants were also randomly assigned to complete additional tasks designed to either increase or decrease reliance on automatic processes. I hypothesized that I would find stronger evidence for a causal relationship between implicit attitudes and behavior for individuals who completed tasks designed to increase reliance on automatic processes.

Method

Participants and Design. Three-hundred fifteen participants (209 female, 106 male, 0 other gendered, M_{age} = 19.34, age range: 18 – 35 years) were recruited from the General Psychology and paid subject pools at the University of Colorado Boulder. Twenty of these participants responded in ways that indicated they were not paying attention on at least two occasions and were excluded. The final sample for Experiment 5 was 295 (194 female, 101 male, M_{age} = 19.36, age range = 18-35 years).

The design was a 2 (evaluative condition: orange-good vs. purple-good) X 2 (automaticity: high or low) between-subjects factorial design.

Materials. All materials used in the vignette-present condition of Experiments 4a and 4b were used to manipulate and measure attitudes and behavior. In addition, several tasks were added to increase or decrease participants' reliance on automatic processes. An antisaccade task was also included at the end of the study (as an exploratory measure) to examine whether

individual differences in executive control might further moderate implicit attitude-behavior relations.

Mood Manipulation. Previous research suggests that individuals in a positive mood tend to rely on more automatic processes whereas those in a negative mood rely on more deliberative processes (Holland, Vries, Hermsen, & Van Knippenberg, 2012; Fishbach & Labroo, 2007). As such, all participants began this experiment with a mood manipulation task (e.g. Schwartz & Clore, 1983) in which they were asked to write about either one of the happiest (high automaticity condition) or one of the unhappiest (low automaticity condition) days of their lives (see Appendix L for exact wording of instructions). Following the instructions provided by Fishbach and Labroo (2007), participants were instructed to "Please use the space below to describe as vividly as possible one of the happies [unhappiest] days of your life. Please use the space below to describe in detail (1) what happened on that day, (2) how you felt and (3) whether the events of the day elicited thoughts or imagery that increased the strength of your feelings." Participants were given an unlimited amount of time to complete this task, but typically finished within 5-10 minutes. No items assessing mood were administered as some evidence suggests that affect labeling may decrease emotional reactivity (Lieberman, Eisenberger, Crockett, Tom, Pfeifer, & Way, 2007).

Opportunity for Misattribution of Associations. Work by Dijksterhuis (2004) suggests that allowing time for the consolidation of automatic associations may be important for increasing reliance on these associations. Similarly, Loersch and Payne (2014) suggest that reliance on more automatic processes can be increased through the misattribution of learned associations to internal rather than external sources. As such, participants in the high-automaticity condition completed a 2-back task designed to increase time for consolidating

associations learned through the implicit attitude manipulation tasks. Since this task was unrelated to either of the groups for which attitudes were trained, it is also possible that this task provides a distraction from the two groups of fish and may increase misattribution of associations.

Following Dijskterhuis (2004), the numbers 1 through 9 were repeated six times and put into a randomized order such that the numbers were randomized but appeared in the same order for each participant. Each number appeared on the screen for 500ms followed by a 2500ms interstimulus interval. Participants were instructed to press the spacebar every time the number on the screen matched the number that appeared two trials before. Participants completed 54 trials and viewed the instruction screen for 20 seconds. Fourteen trials (16%) were trials in which participants should have pressed the spacebar. Participants in the low automaticity condition did not complete the 2-back task and instead proceeded directly from the IAT to the behavioral outcome measure.

Process Reliance Instructions. Several studies suggest that instructions encouraging participants to take an affective focus or "go with their gut" encourage reliance on automatic processes (DeHouwer & Smith, 2012; Scarabis, Florack & Gosejohann, 2006). As such, I used the instructions for the behavioral outcome measure to manipulate reliance on automatic processes. In the high automaticity condition, participants were instructed to "go with your gut feeling when deciding which key to press" and to "not think too much…and just choose your response based on your first feelings and intuitions." In contrast, participants in the low automaticity condition were instructed to "pay close attention and respond carefully." Appendix M presents the instructions for this task.

Antisaccade Task. Finally, an individual difference measure of executive control, the antisaccade task was administered as there is evidence that individuals high in executive control demonstrate lower implicit attitude-behavior relations (Grenard, Ames, Wiers, Thush, Sussman, & Stacy, 2008; Hofmann, Gschwendner, Friese & Wiers, 2008; Houben & Wiers, 2009). Participants completed one practice block and one critical block of prosaccade trials followed by one practice block and three critical blocks of antisaccade trials. On prosaccade trials, participants were instructed to focus on a fixation cross, which appeared in the middle of the screen for a randomly selected duration between 1500 and 3000ms (in 250ms increments). Next a black box appeared as a cue either to the right or the left of this fixation cross for 175ms. This cue was replaced by a thin black arrow pointing either up, down, left or right. The arrow appeared for 150ms and was then replaced by a gray pattern mask, which remained on screen until the participant responded. On prosaccade trials, participants were instructed to press the arrow key that corresponded to the direction in which the arrow pointed. If they were unsure of the direction of the arrow, participants were instructed to guess. The practice block consisted of 12 trials which was followed by a prosaccade critical block consisting of 20 trials. After completing the prosaccade blocks, participants completed the antisaccade blocks. These trials were identical to the prosaccade trials except that the cue always appeared on the opposite side of the screen from the arrow and participants were encouraged to look away from the cue when it appeared on screen. Additionally, the cue presentation time decreased in each critical block. In the practice block and first critical block, the cue was presented for 225ms on each trial. In the second critical block the cue was presented for 200ms and in the final critical block, the cue was presented for 175ms. The antisaccade practice block consisted of 12 trials and each critical block consisted of 28 trials. An antisaccade score was calculated for each participant as the proportion

of correct trials across the three antisaccade blocks (following Ito, Friedman, Barthalow, Correll, Loersch, Altamirano, & Miyake, 2015). Since this measure was included as an exploratory measure and the results do not increase my ability to understand major findings, findings related to this measure are included in Appendix N.

Procedure. Participants completed the sorting task, read the vignette, completed the evaluative conditioning task and attitude measures (with implicit and explicit attitude measures counterbalanced) as in the vignette-present condition of Experiments 4a and 4b. Following the attitude measures, participants in the high-automaticity condition completed the 2-back task and then the behavioral outcome measure. Participants in the low automaticity condition advanced from the attitude measures directly to the behavioral measure. After completing the behavioral measure, all participants completed the antisaccade task. Finally, participants answered some questions probing their awareness of the hypotheses for this study, their awareness of contingencies between purple and orange fish exemplars and valenced words in the evaluative conditioning task, reading comprehension on the vignettes, and demographic items. At completion of the experiment, participants were debriefed, thanked and given either partial course credit or paid.

Results

The general approach to analysis of Experiment 5 was similar to the previous 4 experiments. First, linear regression analyses were used to examine relationships between implicit attitudes, explicit attitudes, and behavior. For this experiment, outcome measures were regressed on contrast-coded evaluative condition (-0.5 = purple-good; 0.5 = orange-good), contrast-coded automaticity condition (-0.5 = low-automaticity; 0.5 = high-automaticity) and their interaction. After completing the regression analyses, path models were used to estimate

indirect effects for both implicit and explicit attitudes. This allowed me to test whether implicit or explicit attitudes significantly mediated any relationship between the condition variables and behavior.

Condition Differences in Attitudes. Figure 19 depicts condition differences in implicit attitudes. Across automaticity condition, there was an effect of evaluative condition on implicit attitudes, $b = 0.53$, $t(291) = 9.96$, $p < 0.001$, $R^2_{partial} = 0.25$. Participants in the orange-good condition demonstrated a significant implicit preference for orange over purple fish, $b = 0.29$, $t(291) = 7.76$, $p < 0.001$, $R^2_{partial} = 0.17$; whereas, participants in the purple-good condition demonstrated an implicit preference for purple fish, $b = -0.24$, $t(291) = -6.34$, < 0.001, $R^2_{partial} = 0.12$. An effect of automaticity condition (across evaluative condition) also emerged, $b = 0.15$, $t(291) = 2.87$, $p = 0.004$, $R^2_{partial} = 0.03$. Participants in the high automaticity condition showed more positive implicit preferences for orange fish than participants in the low automaticity condition. Since implicit and explicit attitudes were positively related ($b = 0.30$, $t(293) = 10.14$, $p < 0 .001$, $R^2 = 0.26$), I examined whether these effects remained after controlling for explicit attitudes. Both the evaluative and automaticity condition effects on implicit attitudes remained significant after controlling for explicit attitudes, partial effect of evaluative condition: $b = 0.32$, $t(290) = 4.83$, $p < 0.001$, $R^2_{partial} = 0.07$; partial effect of automaticity condition: $b = 0.15$, $t(290) = 2.93$, $p = 0.004$, $R^2_{partial} = 0.03$.

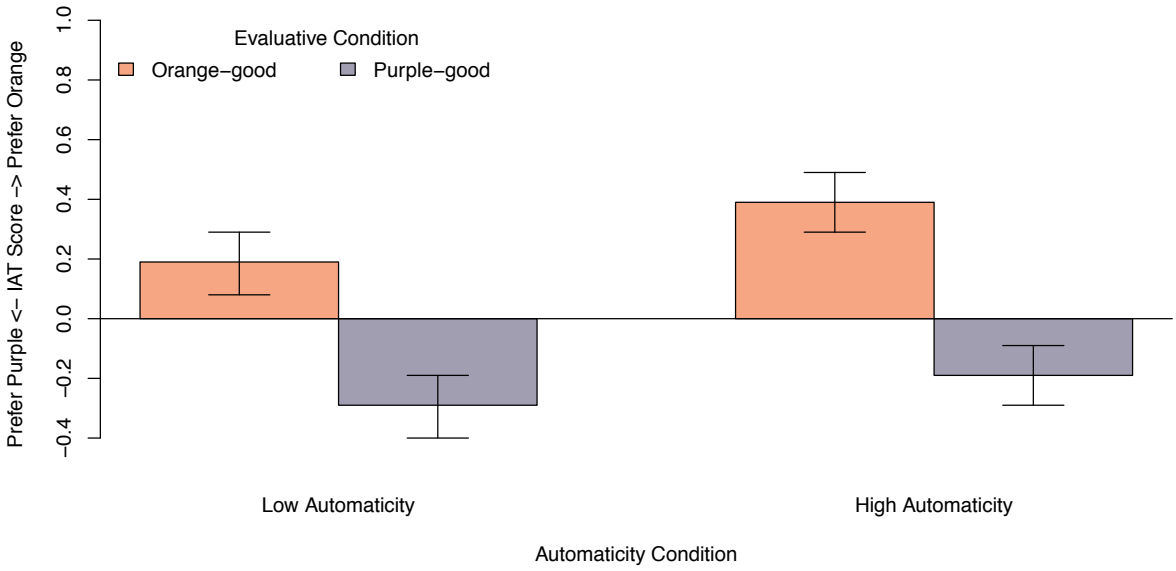


Figure 19. *Condition Differences in Implicit Attitudes, Experiment 5*. Average IAT scores by evaluative condition and automaticity condition. Higher values on the y-axis indicate a greater preference for orange over purple fish. Zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

Figure 20 depicts condition differences in explicit attitudes. Only one effect of condition emerged for the explicit attitude outcome. An effect of evaluative condition, across automaticity condition, revealed that individuals in the orange-good condition self-reported more favorable attitudes towards orange fish than did participants in the purple-good condition, $b = 1.14$, $t(291) = 13.84$, $p < 0.001$, $R^2_{partial} = 0.40$. As with implicit attitudes, participants in the orange good condition demonstrated a significant explicit preference for orange fish, $b = 0.23$, $t(291) = 4.02$, $p < 0.001$, $R^2_{partial} = 0.05$; whereas participants in the purple-good condition demonstrated a significant explicit preference for purple fish, $b = -0.91$, $t(291) = -15.42$, $p < 0.001$, $R^2_{partial} = 0.45$. The effect of evaluative condition on explicit attitudes remained significant after controlling for implicit attitudes, $b = 0.90$, $t(290) = 9.87$, $p < 0.001$, $R^2_{partial} = 0.25$.

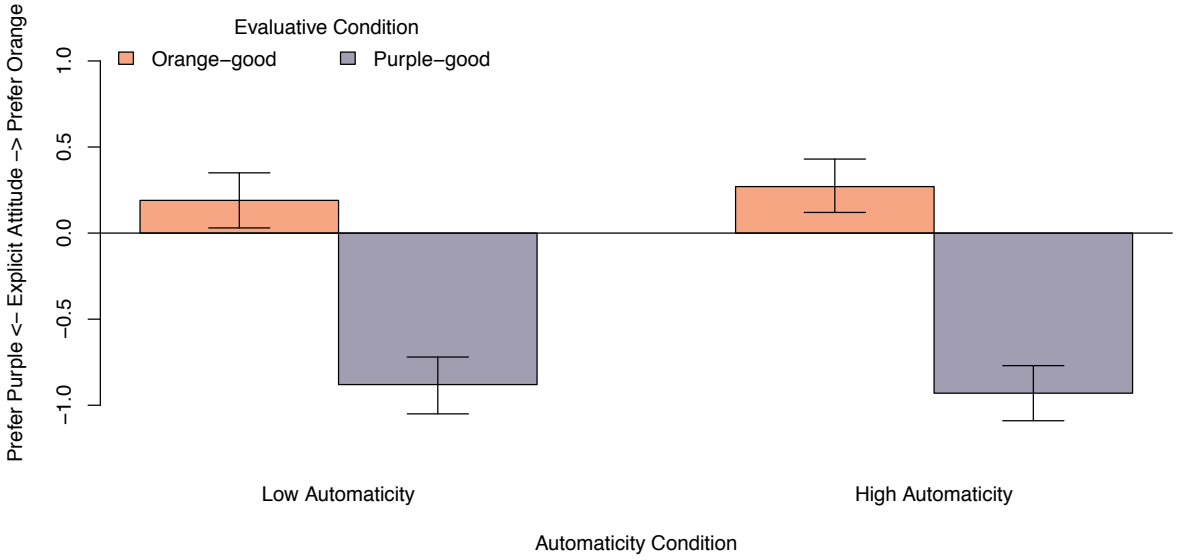


Figure 20. *Condition Differences in Explicit Attitudes, Experiment 5*. Average explicit attitude scores by evaluative condition and automaticity condition. Higher values on the y-axis indicate a greater preference for orange over purple fish. Zero represents no preference for one group over the other. The error bars represent 95% confidence intervals.

Condition Differences in Behavior. *Discriminability (d')*. Table 5 presents average signal detection statistics by condition and trial type. Across evaluative and automaticity conditions, there was no evidence of bias in d' based on trial type (orange vs. purple fish), b = 0.03, $t(291) = 1.15$, $p = 0.25$, $R^2_{partial} = 0.00$. However, there was an marginally significant effect of evaluative condition on differences in d', $b = 0.10$, $t(291) = 1.71$, $p = 0.09$, $R^2_{partial} = 0.01$. Participants in the orange-good condition were better able to discriminate sick from healthy fish on orange fish trials compared to purple fish trials, $b = 0.09$, $t(291) = 2.05$, $p = 0.04$, $R^2_{partial}$ = 0.01. In contrast, participants in the purple-good condition demonstrated directionally better ability to discriminate sick from healthy fish on purple-fish trials (although this difference was non-significant), $b = -0.02$, $t(291) = -0.39$, $p = 0.69$, $R^2_{partial} = 0.00$.

Response bias (c). Across evaluative and automaticity conditions, participants exhibited stronger response bias in favor of saving on orange-fish compared to purple-fish trials, $b = 0.07$, $t(291) = 3.20$, $p = 0.002$, $R^2_{partial} = 0.03$. Although directionally[26] participants in both conditions demonstrated larger "save" response biases on orange fish trials compared to purple fish trials, this effect was larger for participants in the orange-good condition, $b = 0.09$, $t(291) = 2.29$, $p = 0.02$, $R^2_{partial} = 0.02$. Neither the effect of automaticity condition nor the automaticity X evaluative condition interaction reached statistical significance, all ps > 0.83.

Table 5

Signal Detection Statistics by Trial Type and Condition, Experiment 5

		Low Automaticity		High Automaticity	
		Orange-good	Purple-good	Orange-good	Purple-Good
Discriminability (d')	Orange trials	2.41 (.63)	2.34 (.79)	2.20 (.80)	2.29 (.77)
	Purple trials	2.36 (.66)	2.42 (.67)	2.08 (.70)	2.24 (.72)
Response Bias (c)	Orange trials	.10 (.20)	.05 (.22)	.05 (.31)	.09 (.24)
	Purple trials	-.02 (.24)	.03 (.22)	-.07 (.31)	.08 (.28)

Note. Mean (SD) values of d' and c separated by automaticity condition, evaluative condition and trial type, Experiment 5. For d', larger positive values indicate better ability to discriminate sick from healthy animals. For c, larger positive values indicate a more liberal bias in favor of "saving" regardless of whether the animal is sick or healthy.

[26] This difference was significant for participants in the orange-good condition, $b = 0.11$, $t(291) = 3.93$, $p < 0.001$, $R^2_{partial} = 0.05$, but not in the purple-good condition, $b = 0.02$, $t(291) = 0.64$, $p = 0.52$, $R^2_{partial} = 0.00$.

Relationship between attitudes and behavior. ***Discriminability (d').*** There was no evidence that implicit attitudes were related to differences in d', $b = 0.07$, $t(293) = 1.16$, $p = 0.25$, $R^2 = 0.01$. However, explicit attitudes were related to d' differences, $b = 0.07$, $t(293) = 1.99$, $p = 0.05$, $R^2 = 0.01$. Participants with more favorable self-reported attitudes towards orange fish were better able to discriminate sick from healthy fish on orange-fish trials than participants with less favorable attitudes towards orange fish. The relationship between explicit attitudes and differences in d' was attenuated after controlling for implicit attitudes, $b = 0.06$, $t(292) = 1.57$, $p = 0.11$, $R^2_{partial} = 0.01$. See Figure 21 for a depiction of these relationships.

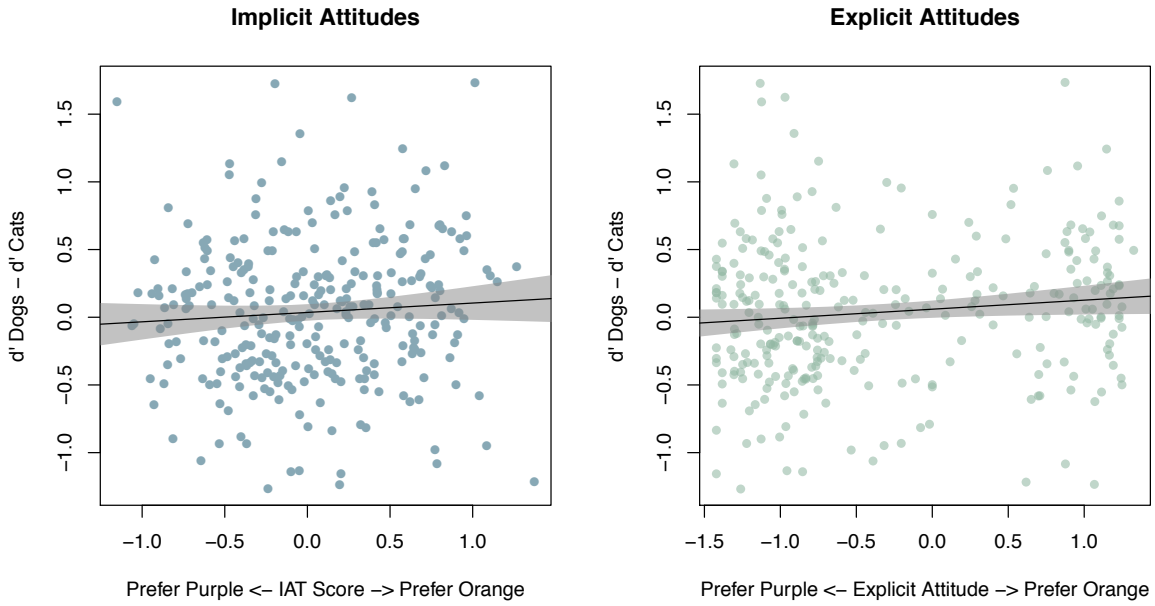


Figure 21. *Experiment 5: Attitude Behavior Relationship, d' Differences*. Relationship between implicit attitudes and differences in d' (left panel) and explicit attitudes and differences in d' (right panel) for orange fish versus purple fish trials. Values greater than 0 on the y-axis indicate participants were better able to discriminate healthy from unhealthy orange fish compared to purple fish while negative values indicate the reverse. Gray bands represent the 95% confidence interval around the line of best fit (black line).

Response Bias (c). Implicit attitudes were significantly related to differences in response bias, $b = 0.10$, $t(293) = 2.55$, $p = 0.01$, $R^2 = 0.02$. Participants with stronger implicit preferences

for orange fish demonstrated a greater response bias in favor of saving orange fish (see Figure 22). A similar pattern emerged for the explicit attitude-response bias difference relationship, b = 0.09, $t(293) = 4.02$, $p < 0.001$, $R^2 = 0.05$. Participants with more favorable explicit attitudes towards orange fish also demonstrated a greater save response bias on orange fish trials than participants with less explicit preference for orange fish. Controlling for explicit attitudes attenuated the relationship between implicit attitudes and differences in response bias, b = 0.03, t(292) = 0.63, p = 0.53, $R^2_{partial} = 0.00$. However, controlling for implicit attitudes did not attenuate the relationship between explicit attitudes and differences in response bias, b = 0.08, $t(292) = 3.13$, $p = 0.002$, $R^2_{partial} = 0.03$.

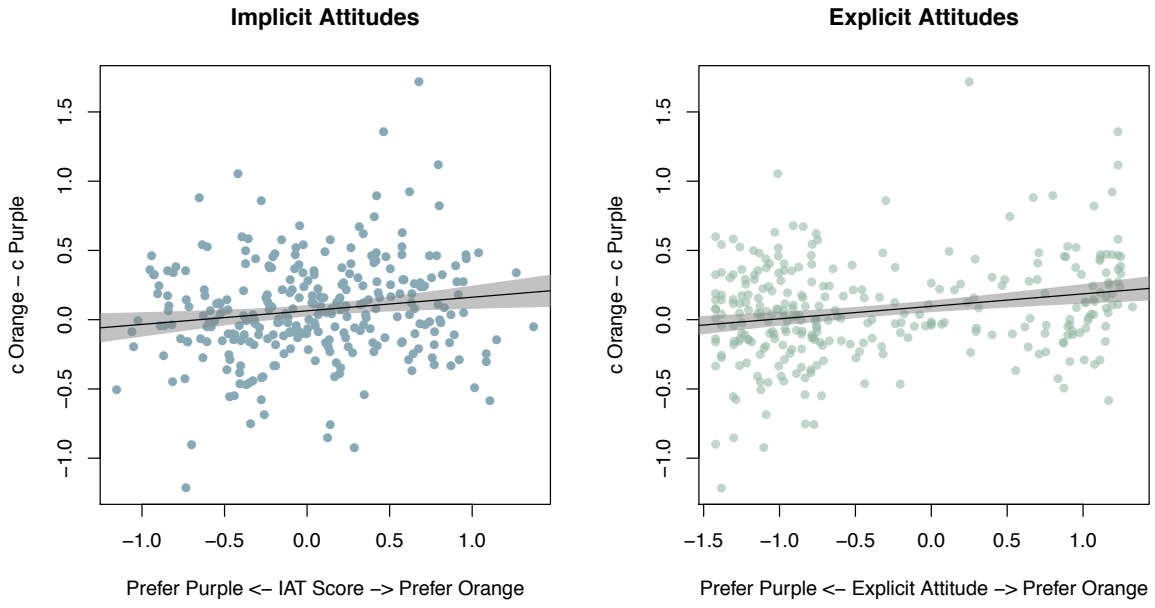


Figure 22. *Experiment 5: Attitude Behavior Relationship, c Differences*. Relationship between implicit attitudes and differences in c (left panel) and explicit attitudes and differences in c (right panel) for orange fish versus purple fish trials. Values greater than 0 on the y-axis indicate participants were respond in favoring of saving orange fish compared to purple fish while negative values indicate the reverse. Gray bands represent the 95% confidence interval around the line of best fit (black line).

Mediation Models. The approach to testing mediation was similar to the approach used in Experiments 4a and 4b. Evaluative condition (contrast coded), automaticity condition

(contrast coded) and their interaction were entered as predictors of differences in d' and c in two separate models. IAT and explicit attitude scores were simultaneously entered as mediators of the condition-behavior relationships. Indirect effects were estimated separately for implicit and explicit attitudes. If implicit attitudes mediated either evaluative condition differences in the outcome measures or if the moderating effect of automaticity condition on the evaluative condition-behavior relationship was mediated by implicit attitudes, this would provide evidence that implicit attitudes caused the condition differences in the outcome variable(s).

Discriminability (d'). Figure 23 demonstrates the standardized path coefficients for the model predicting differences in discriminability. The total effect of evaluative condition on difference s in d' was marginally significant, $b = 0.09$, $z = 1.84$, $p = 0.08$, but there was no evidence that either implicit or explicit attitudes mediated this effect: indirect effect implicit: $b = -0.01$, $z = -0.17$, $p = 0.87$; indirect effect explicit: $b = 0.06$, $z = 1.18$, $p = 0.24$. Automaticity condition did not appear to significantly moderate evaluative condition, $b = -0.03$, $z = -0.54$, $p = 0.59$. This non-significant moderating effect was not mediated by either attitude measure, indirect effect implicit: $b = -0.00$, $z = -0.16$, $p = 0.87$; indirect effect explicit: $b = 0.00$, $z = 0.16$, $p = 0.88$.

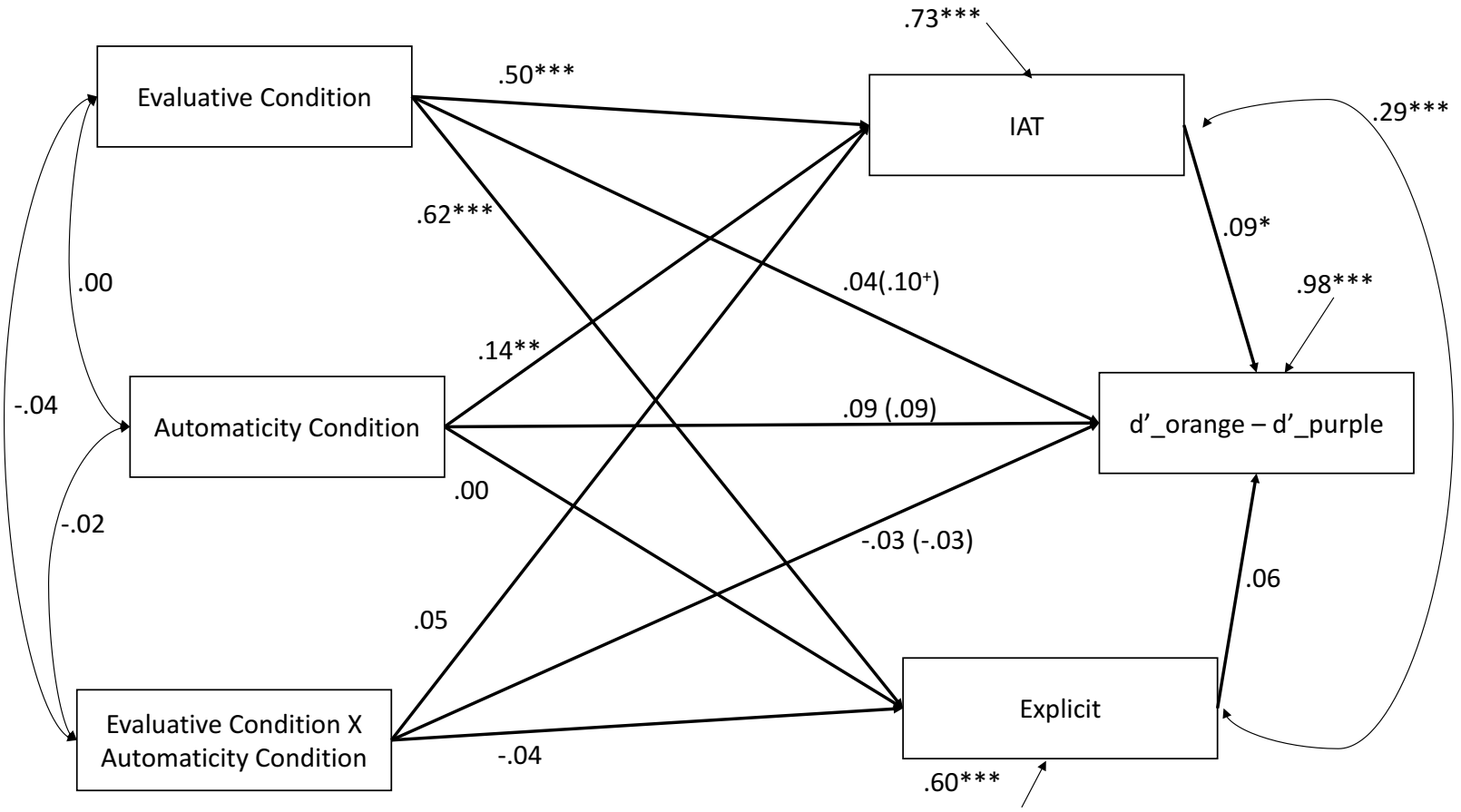


Figure 23. *Experiment 5: Mediated Moderation Model, d' Differences*. Multiple mediated moderation path model examining the impact of evaluative condition, automaticity condition and their interaction on differences in d' for orange fish versus purple fish trials. Total effect of condition on differences in d' is represented in parentheses. Numeric values represent standardized path estimates. Indirect effects of implicit and explicit attitudes on evaluative condition effects and on the evaluative condition X automaticity condition interaction were all non-significant, *p*s > .23. $^{+}$p < 0.1; *p < .05; **p< 0.01; ***p < 0.001.

Response Bias (c). Figure 24 displays the standardized path coefficient for the model predicting differences in response bias. The total effect of evaluative condition on differences in response bias was statistically significant, $b = 0.13$, $z = 0.13$, $z = 2.33$, $p = 0.02$, indicating that individuals in the orange-good condition demonstrated greater response bias in favor of saving orange fish than participants in the purple-good condition. This effect was mediated by explicit attitudes, indirect effect: $b = 0.14$, $z = 2.91$, $p = 0.02$. Implicit attitudes did not significantly mediate the evaluative condition-response bias difference relationship, indirect effect: $b = 0.03$, $z = 0.79$, $p = 0.43$. Automaticity condition did not moderate the effect of evaluative condition on response bias difference, $b = 0.01$, $z = 0.10$, $p = 0.92$, and neither implicit nor explicit attitudes mediated the moderating effect, all ps > 0.53.

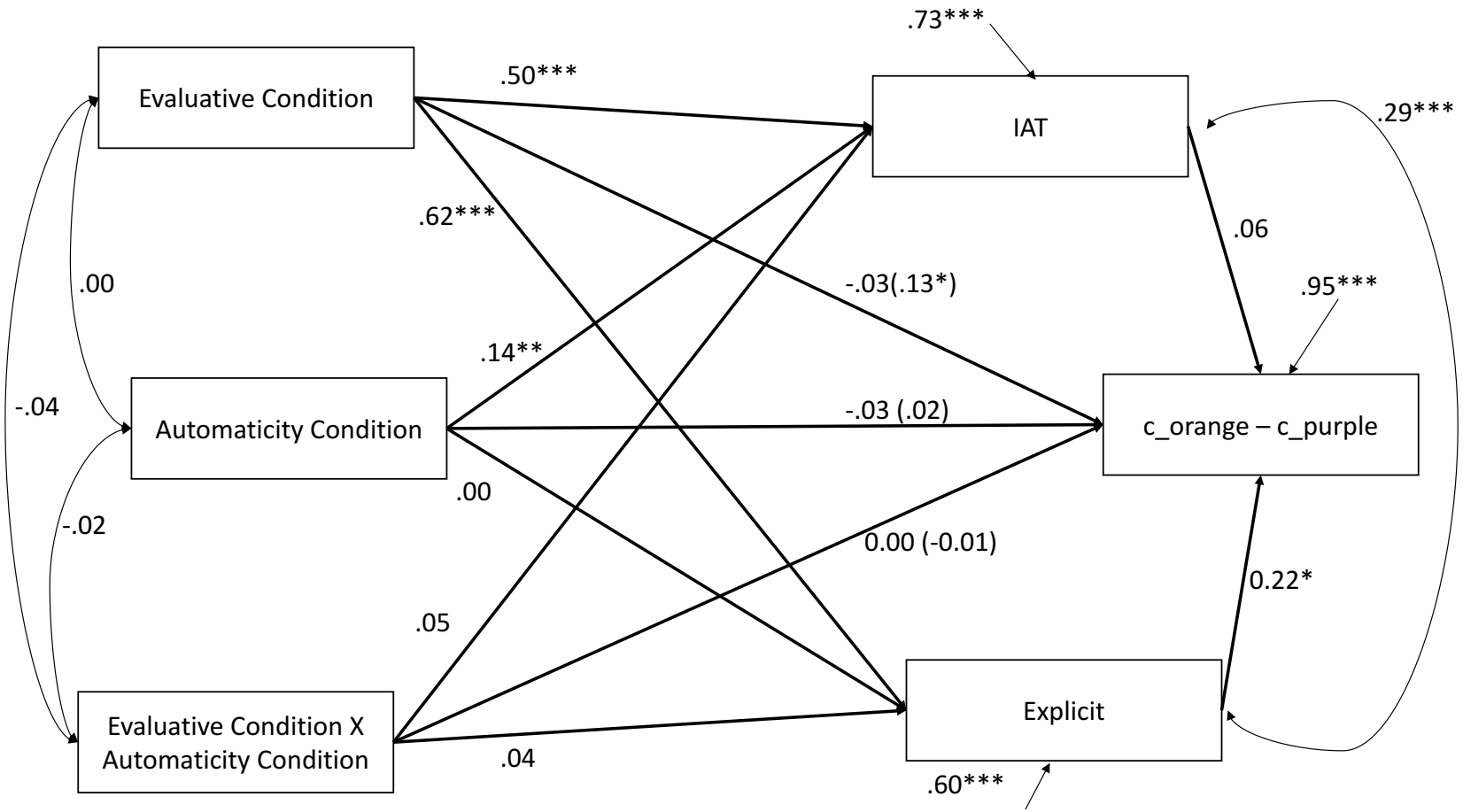


Figure 24. *Experiment 5: Mediated Moderation Model, c*. Multiple mediated moderation path model examining the impact of evaluative condition, vignette condition and their interaction on differences in d' for orange fish versus purple fish trials. Total effect of condition on differences in c is represented in parentheses. Numeric values represent standardized path estimates. Explicit attitudes significantly mediated evaluative condition effects on differences in response bias, indirect effect: $b = 0.14$, $z = 2.91$, $p = 0.004$. No other indirect effects were significant, all *p*s > 0.42; *p < .05; **p< 0.01; ***p < 0.001.

Discussion

Experiment 5 replicated previous experiments in two ways. First, as with Experiments 1-4, the manipulation of implicit attitudes was successful. Second, as with Experiments 4a & 4b, the manipulation of implicit attitudes that included the vignette also yielded condition differences in behavior in the form of differences in response bias. Further, correlations between implicit attitudes and behavior in the mediation model yielded evidence that measured implicit attitudes were related to differences in response bias. Notably, unlike experiments 4a and 4b, only the indirect effect of explicit attitudes on behavior reached statistical significance, suggesting that explicit attitudes mediated condition differences in behavior. There was not sufficient evidence that implicit attitudes mediated condition differences in behavior. Also of note, automaticity condition did not appear to moderate any condition differences in behavior.

One reason we may see discrepancies between experiments is that the IAT is not a process pure measure. That is, IAT d-scores are also influenced by controlled processes. To more closely examine exactly what is being manipulated in my studies and to examine whether a different metric for implicit attitudes may offer better predictive validity, I completed a reanalysis of the data from Experiments 1-5 using the Process Dissociation Procedure.

CHAPTER VII: Reanalysis Using the Process Dissociation Procedure

Even tasks designed to measure automatic processes may be influenced by controlled processes (Jacoby, 1991). Although the IAT was designed to measure more automatic associations, there is evidence that it is also influenced by controlled processes (Fiedler & Bluemke, 2005). That is, the IAT is likely influenced by the strength of an individual's automatic associations, but also their ability to exert control over their behavior. Thus, it may be helpful to explore whether the manipulations of implicit attitudes used in the 5 present experiments altered more automatic or controlled aspects of performance on the IAT. Further, the measures of behavior used in the present studies are very similar in nature to other measures of implicit attitudes and to the IAT itself. Although this was an intentional decision made to increase the likelihood of detecting an implicit attitude-behavior relationship (future work could examine the generalizability of causal effects to broader behaviors), it seems possible that the relationships between implicit attitudes and behavior found in Experiments 4a/4b may be driven by differences in ability to exert control and perform well on speeded tasks. Therefore, it could be useful to examine what aspects of IAT performance are related to the behavioral outcome measures. As such, I used the Process Dissociation Procedure (PDP) to separate IAT scores into controlled and automatic components, then examined 1) whether these components were altered by the implicit attitude manipulations and 2) whether these components related to my behavioral measures.

Calculation of PDP Metrics and Interpretation of Estimates

PDP allows for the estimation of the extent to which an individual relies on controlled vs. automatic processes within a given task (Jacoby, 1991). Following instructions from Payne

(2005), I calculated automatic (PDP-A) and controlled (PDP-C) estimates for each participant's performance in the IAT. To calculate PDP-C, I used the following equation[27]:

$$PDP - C = P(correc\square|congruent) - P(incorrect|incongruent)$$

There are likely not strong normative biases in favor of one group of stimuli over the other, so I arbitrarily defined the orange/good & purple/bad blocks of the IAT as "congruent" blocks and the orange/bad & purple/good blocks as "incongruent" blocks[28]. As such, PDP-C examines the extent to which participants could perform the sorting task required by the IAT correctly for congruent blocks and avoid performing the task incorrectly for incongruent blocks. Higher values on PDP-C indicate better performance on the IAT. Since I would not expect the evaluative conditioning procedure to impact the degree of accuracy with which an individual completes the IAT, this estimate should not be impacted by evaluative condition. However, since this accuracy may reflect the extent to which an individual can exert control over their responses in the IAT and it is reasonable to expect that one's ability to exert control in one task would be related to their ability to exert control in other tasks, I may find that PDP-C is related to performance on the outcome measures (and therefore related to greater accuracy and less biased behavior in the fish rescue game).

The equation for PDP-A was as follows:

$$PDP - A = P(incorrect|incongruent)/(1 - PDP - C)$$

[27] Two participants had PDP-C scores of exactly 1, making it impossible to calculate their PDP-A values, for these participants, the value of ½ of an incorrect trial (.5/60 = .004) was subtracted from their PDP-C score.

[28] In the case of Experiment 2, which used dog and cat stimuli rather than fish, dogs/good & cats/bad blocks were coded as "congruent" and the remaining blocks were coded as "incongruent".

In essence, this estimates the proportion of incorrect performance that can be attributed to associating orange fish with positive attributes more than purple fish[29]. As such, I expected scores on PDP-A to be higher for individuals in the orange-good condition (to the extent that the evaluative conditioning procedure increased the strength of associations of the orange fish group with positive evaluative content). Further, if bias in behavior is driven by automatic associations, I would expect to see a positive relationship between differences in the number of fish rescued in the fish recue game (or differences in the orange – purple fish selected in the forced choice task) and PDP-A.

For each of the 5 experiments, estimates of PDP—C and PDP—A were calculated using the formulae above (distributions of PDP-A and PDP-C are presented in Figures 25 and 26).Then I tested whether the implicit attitude manipulations impacted either estimate and whether each estimate was correlated with performance on the outcome measures (see Tables 6 and 7 for PDP-A and PDP-C estimates by experiment and condition). Finally, I re-estimated the mediation models presented in Chapters 2 through 6 substituting IAT score with the two PDP estimates.

[29] Or, in the case of Experiment 2, the proportion of incorrect performance that can be attributed to associating dogs with positive attributes more than cats.

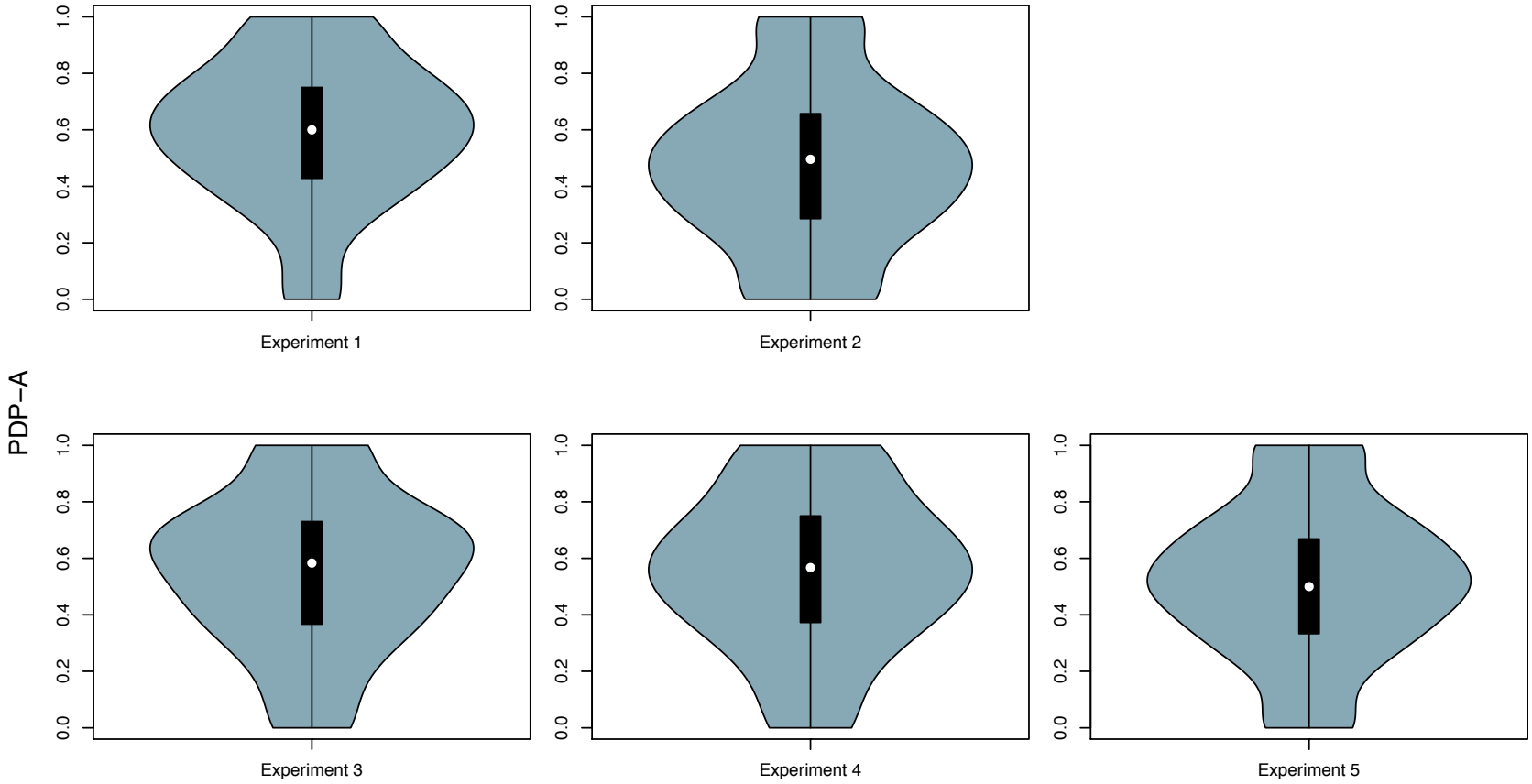


Figure 25. *Distribution of PDP-A by Experiment*. Violin plots for the distribution of PDP-A for each experiment. The blue shapes represent density plots of PDP-A scores and the box in the middle is a box plot with the median represented by the white circle in the middle.

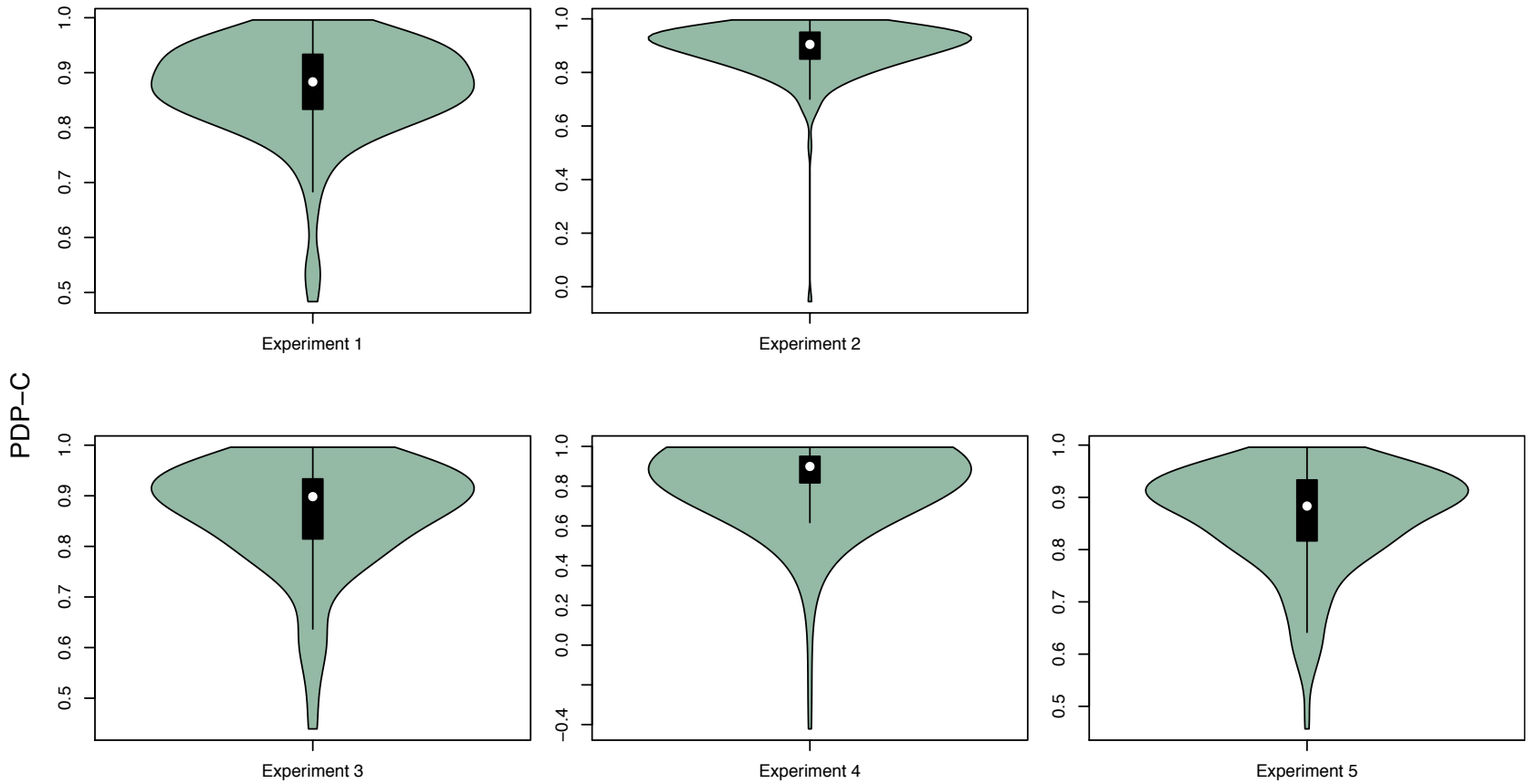


Figure 26. *Distribution of PDP-C by Experiment*. Violin plots for the distribution of PDP-C for each experiment. The blue shapes represent density plots of PDP-C scores and the box in the middle is a box plot with the median represented by the white circle in the middle.

Experiment 1 Reanalysis

Condition differences in PDP estimates. Overall, there was no evidence that my manipulation impacted either PDP-A or PDP-C, PDP-A: $b = 0.01$, $t(155) = 0.36$, $p = 0.72$, $R^2 = 0.00$; PDP-C: $b = -0.01$, $t(155) = -0.65$, $p = 0.52$, $R^2 = 0.00$.

PDP-behavior relationship. There was no evidence that PDP-A was related to differences in the number of orange vs. purple fish saved during the fish rescue game, $b = 0.10$, $t(154) = 0.04$, $p = 0.97$, $R^2 = 0.00$. However, PDP-C did significantly relate to bias in the fish rescue game, $b = -17.30$, $t(154) = -2.45$, $p = 0.02$, $R^2 = .10$. This indicates that participants who performed more accurately on the IAT tended to exhibit less bias in terms of the fish they were likely to save[30]. Neither PDP-A nor PDP-C were related to biased responding on the forced choice task, PDP-A: $b = 0.52$, $t(155) = 0.92$, $p = 0.36$, $R^2 = 0.01$; PDP-C: $b = -0.36$, $t(155) = -0.22$, $p = 0.83$, $R^2 = 0.00$.

Mediation models. Finally, I estimated the same multiple mediation models reported in Experiment 1, but replaced the IAT mediator with the two PDP estimates. None of the indirect paths for PDP-A, PDP-C or explicit attitudes were significant for the fish rescue game, all ps > 0.32. There was no evidence that any of the proposed mediators explained any condition differences in behavior for this task. For the forced choice task, only the indirect effect of explicit attitudes was marginally significant, $b = 0.05$, $z = 1.71$, $p = 0.09$, suggesting that explicit attitudes may partially mediate condition effects on behavior. No other indirect effects were statistically significant, ps > 0.86.

[30] I checked to see if this could be due to increased overall accuracy in the fish recue game, and this did not appear to be the case. Although PDP-C was related to overall accuracy, b = 0.14, t(154) = 2.29, p = 0.02, controlling for percent accuracy did not attenuate the relationship between PDP-C and the fish rescue game difference score, b = -16.27, t(153) = -2.26, p = 0.03.

Experiment 2 Reanalysis

PDP metrics were calculated the same way as in Experiment 1 and the same models were estimated. Of note, Experiment 2 used dogs and cats as the groups for which implicit attitudes were manipulated. To parallel calculation of IAT scores, blocks in which participants responded to dogs and good words with the same response key were labeled as congruent and blocks in which participants responded to cats and good words with the same response key were labeled as incongruent.

Condition differences in PDP estimates. As in Experiment 1, there was no evidence that the manipulation of implicit attitudes impacted either PDP-A or PDP-C, PDP-A: $b = 0.03$, $t(205) = 0.81$, $p = 0.42$, $R^2 = 0.00$; PDP-C: $b = 0.004$, $t(205) = 0.32$, $p = 0.75$, $R^2 = 0.00$.

PDP-behavior relationship. PDP-A was significantly related to differences in d-prime in the opposite of the expected direction, $b = -0.26$, $t(205) = -2.61$, $p = 0.03$, $R^2 = 0.03$. Individuals with stronger dog-positive associations tended to bias saves on the pet rescue game in favor of dogs less than individuals with weaker associations. There was no evidence that PDP-C was related to differences in d-prime, $b = 0.30$, $t(205) = 0.88$, $p = 0.38$, $R^2 = 0.00$. Neither PDP estimate was related to differences in response bias, PDP-A: $b = 0.09$, $t(205) = 1.35$, $p = 0.18$, $R^2 = 0.01$; PDP-C: $b = 0.17$, $t(205) = 0.90$, $p = 0.37$, $R^2 = 0.00$.

Mediation models. One multiple mediation model was estimated for each outcome variable, with PDP-A, PDP-C and explicit attitudes as mediators of condition differences in d' and c. Neither model yielded evidence of significant mediation by any of the three mediator variables (PDP-A, PDP-C, and explicit attitudes), all ps > 0.44.

Experiment 3 Reanalysis

For Experiment 3, there was no behavioral outcome. Therefore, the reanalysis of Experiment 3 examined only the effect of the implicit attitude manipulation on PDP estimates and not PDP-behavior correlations or mediation models.

Condition differences in PDP estimates. PDP-A and PDP-C estimates were calculated using the same methods as Experiments 1 and 2. Since there was no behavioral outcome for Experiment 3, I only examined whether there were effects of evaluative condition, vignette condition on the interaction on PDP-A and PDP-C. Across vignette condition, there was evidence of an evaluative condition effect on PDP-A, $b = 0.19$, $t(89) = 3.53$, $p < 0.001$, $R^2 = 0.12$. Participants in the orange-good condition demonstrated greater orange-positive associations than did participants in the purple good condition. There was no main effect of vignette condition, $b = 0.04$, $t(89) = 0.71$, $p = 0.48$, $R^2 = 0.01$. There was also no evidence of a significant interaction, $b = -0.02$, $t(89) = -0.23$, $p = 0.82$, $R^2 = 0.00$. Evaluative condition, vignette condition and the interaction did not have a statistically significant effect on PDP-C, all p's > 0.33.

Experiments 4a & 4b Reanalysis

Since there were no methodological differences between Experiments 4a and 4b, the results from these studies were analyzed in combined form. Estimates of PDP-A and PDP-C were calculated using the same formulae used for Experiments 1, 2 and 3.

Condition Differences in PDP Estimates. Each estimate of PDP was regressed on contrast-coded evaluative condition (orange-good = 0.5, purple-good = -0.5), vignette condition (vignette-present = 0.5, vignette-absent = -0.5) and their interaction. For PDP-A, the only effect to emerge was an effect of evaluative condition, $b = 0.06$, $t(557) = 2.44$, $p = 0.02$, $R^2 = 0.01$ (all

other *p*'s > 0.14). Across vignette condition, participants in the orange-good condition demonstrated stronger orange-good associations than participants in the purple-good condition. An effect of evaluative condition was also the only effect to emerge for PDP-C, $b = 0.04$, $t(557) = 2.45$, $p = 0.01$, $R^2 = 0.01$ (all other *p*s greater than 0.21). Across vignette condition, participants in the orange-good condition also exhibited greater control than participants in the purple-good condition.

PDP-behavior relationship. I examined the simple relationships between PDP estimates and the outcome variables: d-prime differences and response bias differences. There was no evidence that PDP-A was related to differences in d-prime, $b = 0.06$, $t(559) = 0.68$, $p = 0.50$, $R^2 = 0.00$. However, there was a significant relationship between PDP-C and d-prime differences, $b = 0.27$, $t(559) = 1.98$, $p = 0.05$, $R^2 = 0.01$. Participants who displayed greater control while completing the IAT demonstrated greater behavioral accuracy when on orange-fish compared to purple-fish trials. This effect remained significant after controlling for PDP-A, $b = 0.27$, $t(558) = 1.98$, $p = 0.05$, $R^2 = 0.01$. Neither PDP-A nor PDP-C predicted differences in response bias, all *p*s > 0.18.

Mediation models. Two mediated moderation models were estimated: one for d-prime differences and one for response bias differences. Within each model, there were two mediational paths of interest. The first path, from evaluative condition to behavior examined whether overall evaluative condition differences in d-prime bias or response bias differences could be accounted for by differences in PDP-A, PDP-C or explicit attitudes. The second path of interest examined whether the evaluative condition X vignette condition interaction effect on the behavioral outcomes could be explained by any of these potential mediators. There was no evidence that PDP-A, PDP-C or explicit attitudes mediated any effects of evaluative condition or

the evaluative condition X vignette condition interaction for either d-prime differences or response bias differences, all *p*s > 0.16.

Experiment 5 Reanalysis

The reanalysis of Experiment 5 followed the same procedure as the reanalysis of Experiments 4a and 4b with the addition of the automaticity condition factor and its interaction with evaluative condition.

Condition differences in PDP estimates. Across automaticity condition, there were significant evaluative condition differences in PDP-A, $b = 0.08$, $t(291) = 2.71$, $p = 0.01$, $R^2 = 0.02$. Participants in the orange-good condition demonstrated stronger orange-positive associations than did participants in the purple-good condition. This effect was attenuated after controlling for explicit attitudes, $b = 0.05$, $t(290) = 1.32$, $p = 0.19$, $R^2_{partial} = 0.01$. Automaticity condition and the automaticity X evaluative condition interaction were not significantly related to PDP-A, all *p*s > 0.10. There was no evidence that evaluative condition, automaticity condition or the interaction significantly related to estimates of PDP-C, all *p*s > 0.10.

PDP-behavior relationship. I examined the simple relationships between PDP-A and PDP-C and the outcome measures. There was no evidence that either PDP-A or PDP-C was significantly related to differences in d', PDP-A: $b = 0.08$, $t(293) = 0.73$, $p = 0.47$, $R^2 = 0.00$; PDP-C: $b = 0.18$, $t(293) = 0.58$, $p = 0.56$, $R^2 = 0.00$. Similarly, there was no evidence that differences in response bias were related to either PDP-A, $b = 0.12$, $t(293) = 1.60$, $p = 0.11$, $R^2 = 0.01$, or PDP-C, $b = 0.16$, $t(293) = 0.76$, $p = 0.45$, $R^2 = 0.00$.

Mediation models. Again, the same mediation models presented for Experiment 5 were estimated replacing IAT d-scores with PDP-A and PDP-C estimates. Overall, there was a marginally significant total effect of evaluative condition on d' differences, $b = 0.10$, $z = 1.74$, *p*

= 0.08. Participants in the orange-good condition were better able to differentiate healthy from sick fish on orange fish trials more so than participants in the purple-good condition. Notably neither this effect of evaluative condition nor the evaluative condition X automaticity condition interaction were mediated by PDP-A, PDP-C or explicit attitudes, all *p*s > 0.23.

There was also an overall effect of evaluative condition on differences in response bias, $b = 0.13$, $z = 2.33$, $p = 0.02$. This effect was significantly mediated by explicit attitudes, $b = 0.15$, $z = 3.24$, $p = 0.001$. Neither PDP-A nor PDP-C mediated this effect, ps > 0.31. Automaticity condition did not appear to moderate the effect of evaluative condition on response bias differences, $b = 0.01$, $z = 0.21$, $p = 0.84$. Further, there was no evidence of mediated moderation by either PDP-A, PDP-C or explicit attitudes, all *p*s > 0.44.

Table 6

PDP-A by Experiment and Condition

		Experiment 1	Experiment 2	Experiment 3	Experiment 4		Experiment 5	
					Vignette-absent	Vignette-present	Low Automaticity	High Automaticity
Evaluative Condition	Orange-good (Expt. 2: Dogs-good)	.60 (.27)	.49 (.30)	.66 (.25)	.55 (.29)	.62 (.28)	.52 (.28)	.59 (.26)
	Purple-good (Expt. 2: Cats-good)	.58 (.24)	.45 (.27)	.47 (.26)	.53 (.29)	.53 (.26)	.46 (.23)	.49 (.29)

Note. Mean (SD) values of PDP-A by condition and experiment. Only Experiments 4 and 5 had two factors, PDP estimates for the two vignette conditions (Experiment 4) and the two automaticity conditions (Experiment 5) are presented in different columns.

Table 7

PDP-C by Experiment and Condition

		Experiment 1	Experiment 2	Experiment 3	Experiment 4		Experiment 5	
					Vignette-absent	Vignette-present	Low Automaticity	High Automaticity
Evaluative Condition	Orange-good (Expt. 2: Dogs-good)	.87 (.08)	.89 (.08)	.85 (.09)	.86 (.14)	.88 (.12)	.85 (.11)	.86 (.09)
	Purple-good (Expt. 2: Cats-good)	.86 (.10)	.88 (.12)	.87 (.09)	.84 (.21)	.82 (.21)	.85 (.10)	.88 (.08)

Note. Mean (SD) values of PDP-C by condition and experiment. Only Experiments 4 and 5 had two factors, PDP estimates for the two vignette conditions (Experiment 4) and the two automaticity conditions (Experiment 5) are presented in different columns.

Discussion

The reanalysis of Experiments 1-5 yielded somewhat mixed findings. Across all studies, only Experiments 4a, 4b and 5 demonstrated that PDP-A was impacted by the evaluative conditioning procedure. This may indicate that the evaluative conditioning procedure only alters associations when narrative content is presented in addition to the evaluative conditioning. However, the strongest support for this account, an evaluative condition X vignette condition interaction in Experiments 4a, 4b and 5 was not statistically significant. PDP estimates were also not consistently related to behavior. PDP-C was related to bias in performance for the fish rescue game in Experiment 1 and to differences in d-prime in Experiments 4a and 4b, but not in any other study. PDP-A was only related to behavior (differences in d') in Experiment 2 and the direction of this relationship was the opposite of what was predicted. Finally, none of the studies provided any evidence that either PDP-A or PDP-C mediated any condition differences in behavior.

Overall, this reanalysis did not yield particularly informative results. Although the stronger manipulation of implicit attitudes appeared to impact PDP-A estimates, there was no evidence that these PDP-A estimates were related to behavior in any meaningful way. These results (or lack thereof) may be due to the relatively low error rates observed in IAT performance across these experiments. On average across participants, error rates on the IAT averaged from 5-7%. These low error rates may have limited the variability of estimates of PDP-A and PDP-C and dampened my ability to detect relationships among PDP estimates and condition or behavior. The use of PDP analysis to differentiate automatic from controlled processes in the IAT was an intentional one. PDP analysis is well-established and is a relatively simple technique that yields easily interpretable estimates. However, in hindsight, its reliance on error rates may not make it

the best analytic technique for the data at hand. Future work may want to consider whether other process dissociation techniques that rely on response times (for example diffusion modeling) may be better suited to separating out automatic from controlled processes on this task.

CHAPTER VIII: General Discussion

The purpose of this dissertation was to examine whether implicit group attitudes cause behavior towards individual group members. Although implicit attitudes are thought to cause behavior when individuals lack the motivation or opportunity to control their actions (Fazio & Towles-Schwen, 1999), there is little experimental evidence that this is the case. Across five experiments and two manipulations, I consistently created (Experiments 1, 3, 4, and 5) or manipulated (Experiment 2) implicit group attitudes but found mixed evidence regarding the ability of implicit attitudes to cause behavior.

In Experiments 1 and 2 evaluative conditioning and single group exposure was sufficient to create condition differences in implicit attitudes over and above changes in explicit attitudes, but this manipulation did not yield effects on behavior. Further, there was no evidence that implicit attitudes mediated any condition differences in behavior. A larger manipulation of implicit attitudes was obtained in Experiments 3, 4, and 5 by adding a narrative vignette[31] (Experiments 3, 4, and 5), and this resulted in condition differences in behavior (Experiments 4 and 5). Whereas implicit (but not explicit) attitudes mediated condition differences in behavior in Experiment 4; explicit (but not implicit) attitudes mediated condition differences in behavior in Experiment 5.

There are several reasons that may help account for the null effects of implicit attitudes found in Experiments 1, 2, and 5. First, the manipulation used in Experiments 1 and 2 may simply have yielded too small of an effect to demonstrate statistically significant condition differences in behavior. This is supported by the larger effect sizes for condition differences in

[31] The effect size of the manipulation on implicit attitudes increased from an average $R^2 = 0.03$ in Experiments 1 and 2 to an average $R^2 = 0.22$ in Experiments 3, 4, and 5.

behavior yielded in Experiments 4 and 5 (which both included the narrative vignette manipulation). However, this account does not explain why condition differences in Experiment 5 were not significantly mediated by implicit (rather than explicit) attitudes.

Several possibilities for the null effect in Experiment 5 remain. Although Experiment 5 included a set of moderators designed to increase reliance on automatic processes there was no evidence that these moderators did increase such reliance and there are other factors that were not included that are also theorized to increase more automatic processes. For example, implicit attitudes are thought to reflect affect (Gawronski & Bodenhausen, 2006; March & Graham, 2005) and so may be more likely to cause correspondent behavior that is similarly affectively driven (Azjen, Icek, & Timko, 1986).

These null effects may also be a reflection of imprecise implicit attitude measurement. Although the IAT is the most commonly used and well-validated measure of implicit attitudes, its reliability (especially test-retest reliability) is not perfect (Cunningham, Preacher, & Banaji, 2001; Schnabel, Aspendorf & Greenwald, 2008). Further, implicit measures lack strong convergent validity as they often do not co-vary with each other (e.g. Bosson, Swann, & Pennebaker, 2000; Ito, et al., 2015). Thus, it is possible that we may need to develop a better measure of implicit attitudes before we can reliably detect a causal relationship between implicit attitudes and behavior.

Further, implicit attitudes often relate to behaviors in contexts for which there is a high amount of personal relevance such as interpersonal interactions (e.g. Dasgupta & Rivera, 2006; Dovidio, et al., 2002; Hofmann, et al., 2008; Shelton, Richeson, Salvatore, & Trawalter, 2005) consumer choices (Dempsey & Mitchell, 2010; Gibson, 2008) and voting behavior (Greenwald, et al., 2009; Payne, et al., 2010). The tight experimental control used in the present study may

have limited the personal relevance of the attitude domain and/or the behavioral outcome and may have had the unintended consequences of limiting my ability to detect a causal relationship between implicit attitudes and behavior.

More generally, it may simply be that there are other conditions that are necessary for detecting a causal relationship between implicit attitudes and behavior. Personal relevance is one such condition, but it may also be that, order for implicit attitudes to cause behavior, they must be long-held or social norms or historical precedent must justify the use of these more automatic processes. As such, although allowing for tight control of experimental conditions, this dissertation research could limit my ability to detect a causal relationship between implicit attitudes and behavior. That is, an experiment in which participants are asked to engage in a novel task involving novel groups of stimuli might not offer the best conditions for detecting the causal relationship between implicit group attitudes and behavior.

Implications

The current work corresponds with findings from an unpublished meta-analysis that more broadly examines the effectiveness of different strategies for altering implicit bias (Forscher, Lai, Axt, Ebersole, Herman, Devine, & Nosek, 2016). In this meta-analysis, Forscher and colleagues found little evidence, across 46 samples from the broader implicit attitudes literature, that manipulations of implicit attitudes also impacted behavior (or that any impact of the manipulations could be explained by implicit attitudes)[32].

[32] In contrast to the present work, the Forscher et al. (2016) meta-analysis did not consider theory-based moderators or explicit attitudes and included studies which measured implicit attitudes towards individual people and/or objects as well as studies that measured implicit group attitudes. Further, the test of the overall indirect effect of implicit attitudes included several studies for which the initial attempt to manipulate implicit attitudes was unsuccessful.

My findings in the context of this broader meta-analysis may lead one to question whether implicit attitudes cause behavior at all. Certainly, this work suggests that the causal effect of implicit group attitudes on behavior is more tenuous than previously thought. However, there is still reason to think that implicit attitudes *can* cause behavior under different circumstances.

As mentioned earlier, several meta-analyses of correlational relationships (Cameron, et al., 2012; Greenwald, et al., 2009; Oswald,et al., 2013) suggest that there is a small but reliable relationship between implicit attitudes and behavior. Although the concerns regarding drawing causal conclusions from correlational evidence still remain, the existence of these correlational relationships are consistent with the idea that implicit attitudes cause behavior. More convincingly, several studies outside of the domain of stereotypes and prejudice do suggest that implicit attitudes can play a causal role in behavior, at least at the individual attitude to individual behavior level.

Outside of the domain of stereotypes and prejudice, there is evidence that implicit product attitudes partially cause consumer choice (Dempsey & Mitchell, 2010; Gibson, 2008), that implicit smoking associations partially cause intentions to smoke (Dal Cin, Gibson, Zanna Shumate, & Fong, 2007) and that implicit food attitudes partially account for subsequent decisions to choose healthy snack foods (Hollands, Prestwich, & Marteau, 2011). For example, random assignment of participants to an evaluative conditioning task in which unhealthy snack foods were paired with negative images demonstrated more negative implicit snack food attitudes compared to individuals who did not complete the evaluative conditioning procedure. Further, participants who completed the conditioning task were more likely to choose healthy snack food (rather than junk food) at the end of the study. This effect of conditioning on behavior

was mediated by implicit attitudes. Such evidence indicates that implicit attitudes (at least at the individual level) do act as a cause of behavior in some domains and under certain circumstances[33].

It may also be that implicit attitudes must be well-rehearsed, strong associations before they are capable of causing behavior. Implicit attitudes are commonly defined as "slow-learned" (Greenwald & Banaji, 1995) associations that result from past experience. Although the present research reliably created new implicit attitudes in five different experiments, it is unlikely that anyone would argue that a 30-minute manipulation resulted in slow-learned associations. Relatedly, although the present work demonstrated consistent condition differences in IAT d-scores, PDP analysis did not provide any evidence that such effects were due to condition differences in automatic associations. As previously mentioned, this may simply be because the high accuracy rates in IAT performance reduced variability in PDP estimates and made it difficult for these estimates to yield condition differences or to predict behavior. However, it is also possible that my manipulation impacted some component of IAT scores that are not automatic or associative in nature. Thus, it may be that implicit group attitudes do cause individual level behavior, but that either A) my manipulation does not alter the "implicit" part of an IAT d-score or B) the IAT does not measure implicit attitudes precisely enough to detect such a causal relationship.

[33] Many of these studies demonstrate moderating effects such that causal evidence of the implicit attitude behavior relationship only exists for certain types of individuals or under certain circumstances. For example, Dal Cin et al., (2007) found that exposure to smokers in film clips increased intentions to smoke, but only for individuals who identified with the characters in the film they watched.

Future Directions

Future work should continue to examine whether and when implicit group attitudes cause behavior by further examining necessary conditions for detecting this relationship. One important step will be to examine whether implicit attitudes may be more easily causally linked to behavior that is more affective in nature.

Other work ought to explore the role of training implicit associations over time to examine whether implicit associations that are trained over time may be more robust and may be more likely to cause behavior. The repeated measures nature of such a design could also allow for a closer examination of within subjects changes in implicit attitudes (rather than the between subjects differences examined in the present experiments).

Finally, greater examination of what exactly is being altered by implicit attitude manipulations could be helpful in establishing whether (and when) implicit attitudes cause behavior. Overall, the use of other process dissociation analysis techniques (e.g. diffusion modeling which relies on response times rather than error rates), or the use of an implicit attitude measure with higher error rates may allow for better understanding of what evaluative conditioning and single group exposure manipulations are changing in implicit attitude measures. Since only the addition of a narrative vignette manipulation was sufficient to create differences in behavior, it may be especially important to understand how the vignette increases the magnitude of condition differences in implicit attitudes. For example, is it simply that the vignette offers additional rehearsal of the unconditioned stimulus/conditioned stimulus pairing or does the vignette offer content that allows the participant to link the evaluative associations established in previous tasks with this semantic information which subsequently eases activation of the implicit association?

Conclusions

The present work is an important first step in examining whether and when implicit group attitudes cause behavior towards individuals. Five carefully designed experiments indicate that, although implicit group attitudes were reliably created using two different manipulations, there is little evidence that implicit group attitudes cause behavior. Although implicit attitudes might have a small causal effect on behavior under the right circumstances, this effect is less robust than previously thought. This work provides a strong foundation for future work that should continue to examine the necessary conditions for detecting a causal relationship between implicit group attitudes and behavior.

References

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790–805. http://doi.org/10.1037/a0021594

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888- 918. http://doi.org/ 10.1037/0033-2909.84.5.888

Ajzen, I., & Timko, C. (1986). Correspondence between health attitudes and behavior. *Basic and Applied Social Psychology*, *7*(4), 259-276. http://doi.org/ 10.1207/s15324834basp0704_2

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661. http://doi.org/10.1037/0022-3514.91.4.652

Baeyens, F., Eelen, P., & Bergh, O. V. D. (1990). Contingency awareness in evaluative conditioning: A case for unaware affective-evaluative learning. *Cognition and Emotion*, *4*(1), 3-18.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Brendan more employable than Latoya and Tyrone? Evidence on racial discrimination in the labor market from a large randomized experiment. *American Economic Review*, *94*(4), 992–1013. http://doi.org/ 10.1257/0002828042002561

Bessenoff, G. R., & Sherman, J. W. (2000). Automatic and controlled components of prejudice toward fat people: Evaluation versus stereotype activation. *Social Cognition*, *18*(4), 329–353. http://doi.org/10.1521/soco.2000.18.4.329

Blake, A. (2016, September 26). The first Trump-Clinton presidential debate, annotated. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/09/26/the-first-trump-clinton-presidential-debate-transcript-annotated/?utm_term=.8b2e9d9f3dfd

Bless, H., & Schwarz, N. (1999). Sufficient and necessary conditions in dual-process models: The case of mood and information processing. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology* (pp. 423– 440). New York: Guilford Press.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?. *Journal of Personality and Social Psychology*, *79*(4), 631-643. http://doi.org/10.1037/0022-3514.79.4.631

Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential Priming Measures of Implicit Social Cognition A Meta-Analysis of Associations With Behavior and Explicit Attitudes. *Personality and Social Psychology Review*, *16*(4), 330–350. http://doi.org/10.1177/1088868312440047

Carlsson, R., & Agerström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology*, *57*(4), 278-287. http://doi.org/10.1111/sjop.12288

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37-57. http://doi.org/10.1037/pspa0000014

Cummins, D. (2016, July 13). Column: Are police shootings racially biased? *PBS News Hour.* Retrieved from http://www.pbs.org/newshour/updates/police-shootings-racially-biased/

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological science*, *12*(2), 163-170. http://doi.org/ 10.1111/1467-9280.00328

Dal Cin, S., Gibson, B., Zanna, M. P., Shumate, R., & Fong, G. T. (2007). Smoking in movies, implicit associations of smoking with the self, and intentions to smoke. *Psychological Science*, *18*(7), 559-563. http://doi.org/ 10.1111/j.1467-9280.2007.01939.x

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814. http://doi.org/10.1037/0022-3514.81.5.800

Dasgupta, N., & Rivera, L. M. (2006). From automatic antigay prejudice to behavior: the moderating role of conscious beliefs about gender and behavioral control. *Journal of Personality and Social Psychology*, *91*(2), 268-280. http://doi.org/ 10.1037/0022-3514.91.2.268

Dasgupta, N., & Rivera, L. M. (2008). When Social Context Matters: The Influence of Long–Term Contact and Short–Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions. *Social Cognition*, *26*(1), 112-123. http://doi.org/ 10.1521/soco.2008.26.1.112

Dedonder, J., Corneille, O., Bertinchamps, D., & Yzerbyt, Y. (2014). Overcoming correlational pitfalls: Experimental evidence suggests that evaluative conditioning occurs for explicit but not implicit encoding of CS-US pairings. *Social Psychology and Personality Science, 5(*2), 250-257. http://doi.org/10.1177/1948550613490969.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, *37*(6), 443-451. http://doi.org/ 10.1006/jesp.2000.1464

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*(6), 853- 869.

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

De Houwer, J., & Smith, C. (2012). Go with your gut! Effects of affect misattribution procedures become stronger when participants are encouraged to rely on their gut feelings. *Social Psychology, 44*(5), 299- 302.

Dempsey, M. A., & Mitchell, A. A. (2010). The influence of implicit attitudes on choice when consumers are confronted with conflicting attribute information. *Journal of Consumer Research*, *37*(4), 614-625. http://doi.org/ 10.1086/653947

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. http://doi.org/10.1037/0022-3514.56.1.5

Devine, P. G., Forscher, P. S., Austin, A. S., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*(6), 1267–1278. http://doi.org/10.1016/j.jesp.2012.06.003

Dijksterhuis, A., & Van Olden, Z. (2006). On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology*, *42*(5), 627-631. http://doi.org/ 10.1016/j.jesp.2005.10.008

Dovidio, J. R. & Gaertner, S. L. (2010). Intergroup Bias. In Fiske, Gilbert, & Lindsey (Eds.), *Handbook of Social Psychology (pp. 1084 – 1121).* New York: Wiley.

Dovidio, J. F., Gaertner, S. L., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology, 8*, 88 –102. http://doi.org/ 10.1037/1099-9809.8.2.88

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Interpersonal Relations and Group Processes*, *82*(1), 62–68. http://doi.org/ 10.1037/0022-3514.82.1.62

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual Process Theories in Social Psychology* (pp. 97–116). New York, NY: The Guilford Press.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology*, *27*(4), 307-316. http://doi.org/ 10.1207/s15324834basp2704_3

Fishbach, A., & Labroo, A. A. (2007). Be better or be merry: how mood affects self-control. *Journal of Personality and Social Psychology*, *93*(2), 158-173. http://doi.org/ 10.1037/0022-3514.93.2.158

Fiske, S. T., & Molm, L. D. (2010). Bridging inequality from both sides now. *Social Psychology Quarterly*, *73*(4), 341-346. http://doi.org/10.1177/0190272510389007

Florack, A., Scarabis, M., & Bless, H. (2001). When Do Associations Matter? The Use of Automatic Associations toward Ethnic Groups in Person Judgments. *Journal of Experimental Social Psychology*, *37*(6), 518–524. http://doi.org/10.1006/jesp.2001.1477

Forscher, P.S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2016). A meta-analysis of changes in implicit bias. Manuscript submitted for publication.

Gabriel, U., Banse, R., & Hug, F. (2007). Predicting private and public helping behaviour by implicit attitudes and the motivation to control prejudiced reactions. *British Journal of Social Psychology*, *46*(2), 365–382. http://doi.org/10.1348/014466606X120400

Gapinski, K. D., Schwartz, M. B., & Brownell, K. D. (2006). Can Television Change Anti-Fat Attitudes and Behavior?. *Journal of Applied Biobehavioral Research*, *11*(1), 1-28. http://doi.org/ 10.1111/j.1751-9861.2006.tb00017.x

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. http://doi.org/10.1037/0033-2909.132.5.692

Gibson, B. (2008). Can Evaluative Conditioning Change Attitudes toward Mature Brands? New Evidence from the Implicit Association Test. *Journal of Consumer Research*, *35*(1), 178–188. http://doi.org/10.1086/527341

Gonsalkorale, K., Hippel, W. von, Sherman, J. W., & Klauer, K. C. (2009). Bias and regulation of bias in intergroup interactions: Implicit attitudes toward Muslims and interaction quality. *Journal of Experimental Social Psychology*, *45*(1), 161–166. http://doi.org/10.1016/j.jesp.2008.07.022

Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolsis decisions for Black and White patients. *Journal of General Internal Medicine*, *22*(9), 1231–1238. http://doi.org/ 10.1007/s11606-007-0258-5

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem and stereotypes. *Psychological Review*, *102*(1), 4–27. http://doi.org/ 10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. http://doi.org/ 10.1037/0022-3514.74.6.1464

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

Greenwald, A. G., Pickrell, J. E., & Farnham, S. D. (2002). Implicit partisanship: taking sides for no reason. *Journal of Personality and Social Psychology*, *83*(2), 367- 379. http://doi.org/ 10.1037/0022-3514.83.2.367

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17. http://doi.org/10.1037/a0015575

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 U.S. presidential election. *Analyses of Social Issues and Public Policy*, *9*(1), 341–253. http://doi.org/ 10.1111/j.1530-2415.2009.01195.x

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1-20. http://doi.org/10.1037/0022-3514.90.1.1

Grenard, J. L., Ames, S. L., Wiers, R. W., Thush, C., Sussman, S., & Stacy, A. W. (2008). Working memory capacity moderates the predictive effects of drug-related associations on substance use. *Psychology of Addictive Behaviors*, *22*(3), 426-432. http://doi.org/ 10.1037/0893-164X.22.3.426

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. http://doi.org/10.1177/0146167205275613

Hofmann, W., Gschwendner, T., Castelli, L., & Schmitt, M. (2008). Implicit and explicit attitudes and interracial interaction: The moderating role of situationally available control resources. *Group Processes & Intergroup Relations*, *11*(1), 69–87. http://doi.org/ 10.1177/1368430207084847

Hofmann, W., Gschwendner, T., Friese, M., Wiers, R. W., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: toward an individual differences perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology*, *95*(4), 962-977. http://doi.org/ 10.1037/a0012705

Holland, R. W., Vries, M. D., Hermsen, B., & Van Knippenberg, A. (2012). Mood and the attitude–behavior link: The happy act on impulse, the sad think twice. *Social Psychological and Personality Science*, *3*(3), 356-364. http://doi.org/ 10.1177/1948550611421635

Hollands, G. J., Prestwich, A., & Marteau, T. M. (2011). Using aversive images to enhance healthy food choices and implicit attitudes: An experimental test of evaluative conditioning. *Health Psychology*, *30*(2), 195-203. http://doi.org/ 10.1037/a0022261

Houben, K., & Wiers, R. W. (2009). Response inhibition moderates the relationship between implicit associations and drinking behavior. *Alcoholism: Clinical and Experimental Research*, *33*(4), 626-633. http://doi.org/ 10.1111/j.1530-0277.2008.00877.x

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187. http://doi.org/ 10.1037/a0038557

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

Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. *Journal of Experimental Social Psychology*, *63*, 50-55. http://doi.org/ 10.1016/j.jesp.2015.10.010

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54-69. http://doi.org/10.1037/a0028347

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology*, *41*, 68–75. http://doi.org/ 10.1016/j.jesp.2004.05.004

Kendrick, R. V., & Olson, M. A. (2012). When feeling right leads to being right in the reporting of implicitly-formed attitudes, or how I learned to stop worrying and trust my gut. *Journal of Experimental Social Psychology*, *48*(6), 1316-1321. http://doi.org/10.1016/j.jesp.2012.05.008

Lambert, A., J., Payne, B. K., Ramsey, S., & Shaffer, L., M. (2005). On the predictive validity of implicit attitude measures: The moderating effect of perceived group variability. *Journal of Experimental Social Psychology*, *41*, 114–128. http://doi.org/10.1016/j.jesp.2004.06.006

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical report A-8*.

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science*, *18*(5), 421-428. http://doi.org/10.1111/j.1467-9280.2007.01916.x

Loersch, C., & Payne, B. K. (2014). Situated inferences and the what, who, and where of priming. *Social Cognition*, *32*, 137-151. http://doi.org/10.1521/soco.2014.32.supp.137

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology, 108*(6), 823-849. http://doi.org/10.1037/pspa0000021

Mann, N. H., & Kawakami, K. (2012). The long, steep path to equality: Progressing on egalitarian goals. *Journal of Experimental Psychology: General*, *141*(1), 187–197. http://doi.org/10.1037/a0025602

March, D. S., & Graham, R. (2015). Exploring implicit ingroup and outgroup bias toward Hispanics. *Group Processes & Intergroup Relations*, *18*(1), 89-103. http://doi.org/10.1177/1368430214542256

Marsden, P. V. (2012). *Social Trends in American Life: Findings from the General Social Survey since 1972*. Princeton and Oxford: Princeton University Press.

Mitchell, J. P., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, *132*(3), 455-469. http://doi.org/10.1037/0096-3445.132.3.455

Neumann, R., Hülsenbeck, K., & Seibt, B. (2004). Attitudes towards people with AIDS and avoidance behavior: Automatic and reflective bases of behavior. *Journal of Experimental Social Psychology*, *40*(4), 543–550. http://doi.org/10.1016/j.jesp.2003.10.006

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101- 115. http://doi.org/10.1037//1089-2699.6.1.101

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences*, *15*(4), 152–159. http://doi.org/10.1016/j.tics.2011.01.005

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36-88. http://doi.org/10.1080/10463280701489053

Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition*, *20*(2), 89-104. http://doi.org/10.1521/soco.20.2.89.20992

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring?. *Psychological Science*, *14*(6), 636-639. http://doi.org/10.1046/j.0956-7976.2003.psci_1477.x

Olson, M. A., & Fazio, R. H. (2004). Trait Inferences as a function of automatically activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology*, *26*(1), 1–11. http://doi.org/10.1206/s15324834basp2601_1

Olson, M. A., & Fazio, R. H. (2007). Discordant Evaluations of Blacks Affect Nonverbal Behavior. *Personality and Social Psychology Bulletin*. http://doi.org/10.1177/0146167207303023

Olson, M. A., & Fazio, R. H. (2009). Implicit and explicit measures of attitudes: The perspective of the MODE model. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), Attitudes: Insights from the new implicit measures (pp. 19 – 63). New York, NY: Psychology Press.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*(4), 562–571. http://doi.org/http://0-dx.doi.org.libraries.colorado.edu/10.1037/pspa0000023

Pager, D., & Shepherd, H. (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual Review of Sociology*, *34*, 181–209. http://doi.org/10.1146/annurev.soc.33.040406.131740

Payne, B. K. (2005). Conceptualizing control in social cognition: how executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology*, *89*(4), 488-503. http://doi.org/10.1037/0022-3514.89.4.488

Payne, B. K., Krosnick, J. A., Pasek, J., Lelkes, Y., Akhtar, O., & Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*(2), 367–374. http://doi.org/10.1016/j.jesp.2009.11.001

Peterson, E. R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: Relations with student achievement and the ethnic achievement gap. *Learning and Instruction*, *42*, 123-140. http://doi.org/10.1016/j.learninstruc.2016.01.010

Pleyers, G., Corneille, O., Luminet, O., & Yzerbyt, V. (2007). Aware and (dis) liking: item-based analyses reveal that valence acquisition via evaluative conditioning emerges only when there is contingency awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 130.

Pryor, J. B., Reeder, G. D., Wesselmann, E. D., Williams, K. D., & Wirth, J. H. (2013). The influence of social norms upon behavioral expressions of implicit and explicit weight-related stigma in an interactive game. *Yale Journal of Biology and Medicine*, *86*, 189–201.

Rudman, L. A. (2004). Sources of Implicit Attitudes. *Current Directions in Psychological Science*, *13*(2), 79–82. http://doi.org/10.1111/j.0963-7214.2004.00279.x

Rudman, L. A., & Ashmore, R. D. (2007). Discrimination and the Implicit Association Test. *Group Processes & Intergroup Relations*, *10*(3), 359–372. http://doi.org/10.1177/1368430207078696

Rudman, L. A., & Lee, M. R. (2002). Implicit and explicit consequences of exposure to violent and misogynous rap music. *Group Processes & Intergroup Relations*, *5*(2), 133-150. http://doi.org/10.1177/1368430202005002541

Saleem, M., & Anderson, C. A. (2013). Arabs as terrorists: Effects of stereotypes within violent contexts on attitudes, perceptions, and affect. *Psychology of Violence*, *3*(1), 84-99. http://doi.org/10.1037/a0030038

Scarabis, M., Florack, A., & Gosejohann, S. (2006). When consumers follow their feelings: The impact of affective or cognitive focus on the basis of consumers' choice. *Psychology & Marketing*, *23*(12), 1015-1034. http://doi.org/10.1002/mar.20144

Schnabel, K., Asendorpf, J. B., & Greenwald, A. G. (2008). Assessment of individual differences in implicit cognition: A review of IAT measures. *European Journal of Psychological Assessment*, *24*(4), 210-217. http://doi.org/10.1027/1015-5759.24.4.210

Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial Attitudes in America*. Cambridge, Masssachusetts: Harvard University Press.

Schwartz, N. & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513-523. http://doi.org/10.1037/0022-3514.45.3.513

Shelton, J. N., Richeson, J. A., Salvatore, J., & Trawalter, S. (2005). Ironic effects of racial bias during interracial interactions. *Psychological Science*, *16*(5), 397-402. http://doi.org/10.1111/j.0956-7976.2005.01547.x

Smith, E. R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality and Social Psychology Review*, *4*(2), 108–131. http://doi.org/10.1207/S15327957PSPR0402_01

Spohn, C., & Holleran, D. (2000). Imprisonment Penalty Paid by Young, Unemployed Black and Hispanic Male Offenders, *Criminology*, *38*, 281. http://doi.org/10.1111/j.1745-9125.2000.tb00891.x

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149. http://doi.org/10.3758/BF03207704

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. http://doi.org/10.1073/pnas.1014345108

Stepanikova, I., Triplett, J., & Simpson, B. (2011). Implicit racial bias and prosocial behavior. *Social Science Research*, *40*, 1186–1195. http://doi.org/ 10.1016/j.ssresearch.2011.02.004

Turner, M. A., Ross, S., Galster, G. C., & Yinger, J. (2002). *Discrimination in Metropolitan Housing Markets: National Results from Phase 1 of the Housing Discrimination Study (HDS)* (Working paper No. 2002-16). University of Connecticut, Department of Economics. Retrieved from https://ideas.repec.org/p/uct/uconnp/2002-16.html

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic

achievement gap. *American Educational Research Journal*, *47*(2), 497–527. http://doi.org/10.3102/0002831209353594

Wall Street Journal (2017, February 17). How to combat bias in hiring. *Wall Street Journal Report Podcast*. Retrieved from http://www.wsj.com/podcasts/how-to-combat-bias-in-hiring/9A403C85-C162-4A28-9464-5FF2D5EA9371.html

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207. http://doi.org/10.3758/s13428-012-0314-x

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

Wilson, F. D., Tienda, M., & Wu, L. (1995). race and unemployment: Labor market experiences of black and white men, 1968-1988. *Work and Occupations*, *22*(3), 245–270. http://doi.org/10.1177/0730888495022003002

Wolsiefer, K., Westfall, J., & Judd, C. M. (2016). Modeling stimulus variation in three common implicit attitude tasks. *Behavior research methods*, 1-17. http://doi.org/10.3758/s13428-017-0897-3

Yoshida, E., Peach, J. M., Zanna, M. P., & Spencer, S. J. (2012). Not all automatic associations are created equal: How implicit normative evaluations are distinct from implicit attitudes and uniquely predict meaningful behavior. *Journal of Experimental Social Psychology*, *48*(3), 694-706.

Ziegert, J. C., & Hanges, P. J. (2005). Employment Discrimination: The rold of implicit attitudes, motivation and a climate for Racial Bias. *Journal of Applied Psychology*, *90*(3), 553–562. http://doi.org/10.1037/0021-9010.90.3.553

Zogmaister, C., Arcuri, L., Castelli, L., & Smith, E. R. (2008). The impact of loyalty and equality on implicit ingroup favoritism. *Group Processes & Intergroup Relations*, *11*(4), 493-512. http://doi.org/10.1177/1368430208095402

Appendix A: Fish Stimuli

Fish Stimuli

Below are the stimuli used for Experiments 1, 3, 4, and 5 separated out by task. Note that the same stimuli were used in the sorting and evaluative conditioning tasks because both tasks served the purpose of manipulating implicit group attitudes. Table A1 provides the average liking ratings for each group of stimuli separated by task.

Orange Fish Group: Evaluative Conditioning/Sorting Tasks

Purple Fish Group: Evaluative Conditioning/Sorting Tasks

Orange Fish Group: Implicit Association Test (IAT)

Purple Fish Group: Implicit Association Test (IAT)

Orange Fish Group: Behavioral Tasks

Purple Fish Group: Behavioral Tasks

Mean Pretest Liking Ratings By Group and Task, Fish Stimuli

Table A1

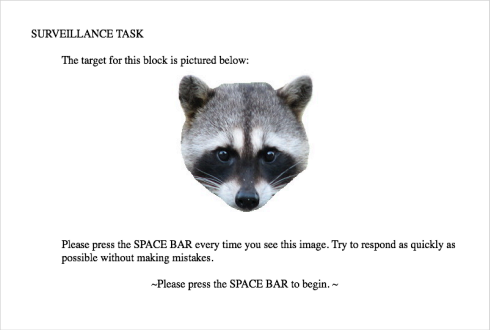
Pre-test Liking Ratings by Stimulus Category and Task, Fish Stimuli

	Orange	Purple	
	M (SD)	M (SD)	t
Evaluative Conditioning	54.67 (2.58)	54.78 (2.34)	0.38
IAT	54.62 (2.06)	54.67 (2.12)	0.11
Behavioral Tasks	54.66 (2.26)	54.96 (1.94)	0.32
Across All Tasks	54.66 (2.24)	54.80 (2.06)	0.30

Note. Participants rated a subset of all possible orange and purple stimuli. Liking ratings ranged from 0 (not at all) to 100 (very much). T-test statistics are the result of an independent samples t-test on the average stimulus ratings for orange vs. purple fish. None of these tests were statistically significant (all p's > 0.71). Standard deviations are provided in parentheses.

Appendix B: Evaluative Conditioning Schematic and Stimuli

Sample Instruction Screen:



Sample Trials for Dogs-Good Condition:

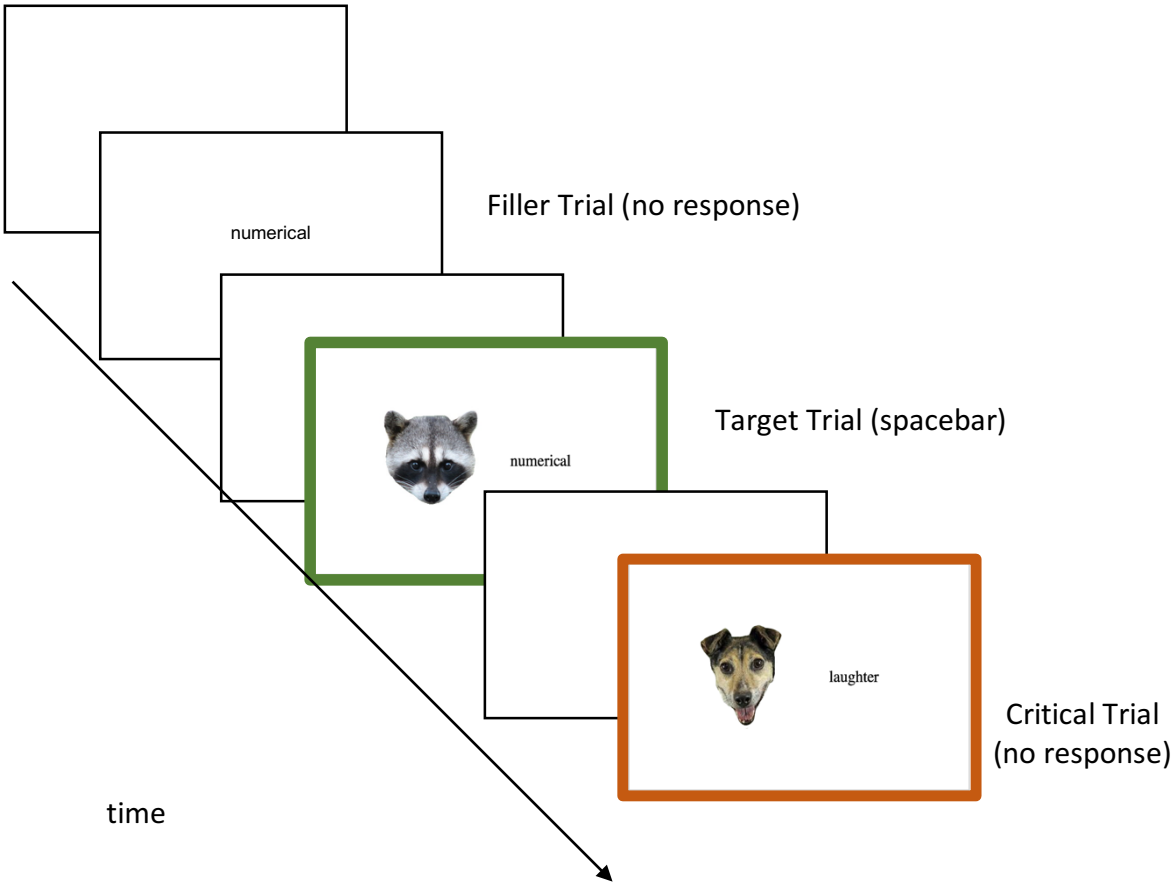


Figure B1. Schematic of the evaluative conditioning task. Participants were instructed to press the spacebar when they viewed the target image (target trial has a green border above). Participants viewed combinations of words and images for 1000ms each separated by 1200ms exposures to blank screens. Although most trials were filler trials of neutral images and/or words,

each block contained 10 critical trials (example bordered in orange above) which paired dogs and cats with either positive or negative words. Above, the critical trial is an example from the dogs-good condition. Note: in Experiments 1, 3, 4, and 5 target and critical trials involved fish rather than mammals.

Stimuli for Evaluative Conditioning Task

Stimuli in the evaluative conditioning task included target group images (see Appendix A), positive words, negative words and neutral words, images of neutral fish (Experiments 1, 3, 4, and 5) or non-dog and non-cat mammals (Experiment 2) and neutral photos from the International Affective Pictures System (IAPS; Lang, Bradley, & Cuthbert, 2008). Below are lists of the words used in this task as well as images of the "neutral" fish and mammal images. IAPS photos are not presented per request of the photo owners.

Positive Words: vacation, happy, fun, enjoyment, fantastic, hug, magical, delight, sunshine, laughter

Negative Words: torture, pedophile, murder, homicide, suicide, die, virus, killer, genocide, rapist

Neutral Words: acquisition, ajar, apartments, barometer, ballistic, boar, borough, caving, circumstance, clink, contain, digestive, depend, diagnose, domino, eleven, episode, extension, embankment, fateful, financial, figment, floss, goggle, headline, hoist, horn, honcho, indication, induce, ingest, invisible, juke, knick, kelp, kiosk, lair, latch, linguist, loosen, machine, mainframe, mat, midair, net, nitrate, numerical, observer, orb, ought, outskirts, pace, platform, pork, porridge, rampart, reduction, repose, ruler, semblance, shellfish, semester, shout, technical, teeth, third, tile, unbutton, understudy, vent, vertical, vowel, widen

Evaluative Conditioning Images, Experiments 1, 3, 4, and 5

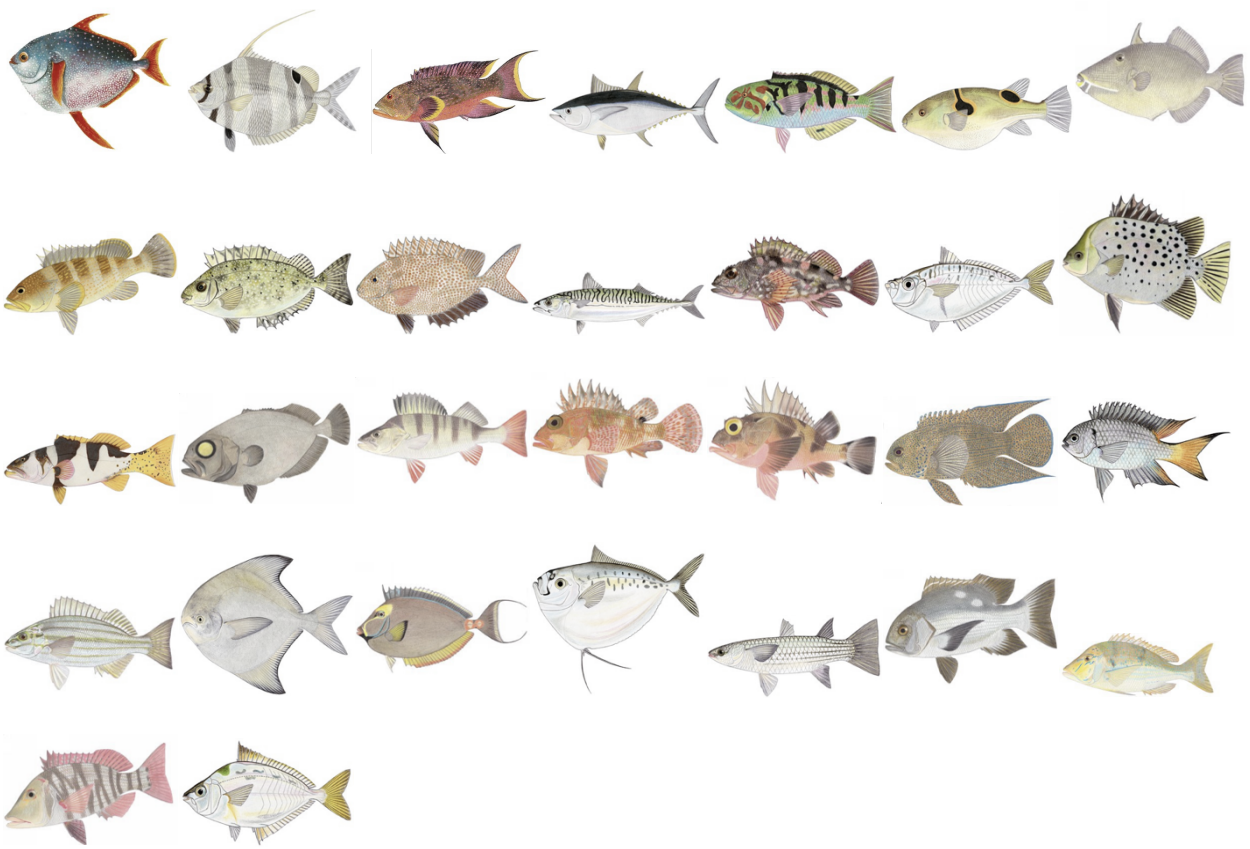


Figure B2. Neutral fish images used in Experiments 1, 3, 4, and 5.

Evaluative Conditioning Images, Experiment 2

Figure B3. Neutral mammal images used in Experiment 2.

Appendix C: IAT Stimuli

The same evaluatively positive and negative words were used for the IAT in each study. See below for a list of these stimuli.

Positive Words: pleasant, delight, helpful, joy, wonderful, cheerful, success, beautiful, enjoy

Negative Words: horrible, angry, terrible, tragic, hate, destroy, brutal, disaster, ugly

Appendix D: Fish Rescue Game Screenshot, Experiment 1

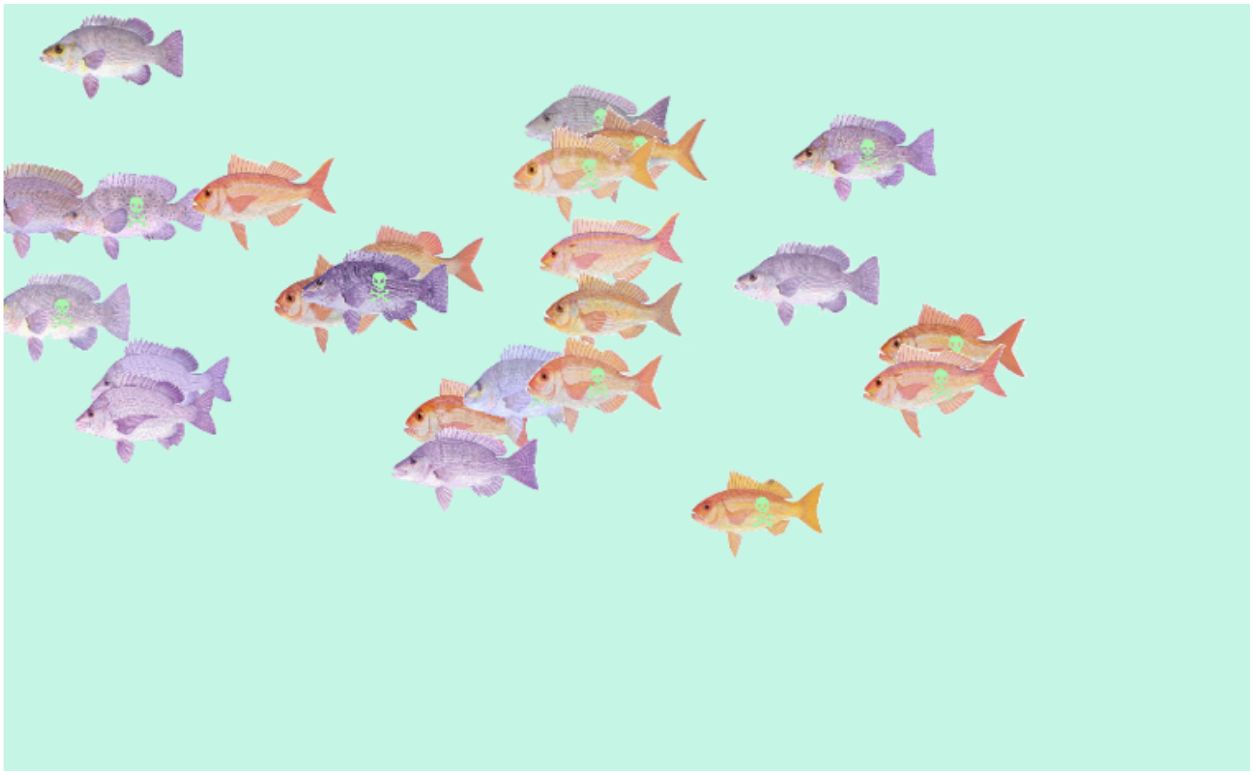


Figure D1. Screenshot of fish rescue game, Experiment 1. In Experiment 1, participants completed a task in which 12 purple and 12 orange fish appeared on the screen and swam around in a random fashion. Half of these fish were healthy (unmarked, above) and half were unhealthy (green skull and bones mark, above). Participants could click to "save" fish and were instructed to save only healthy fish. Clicking a fish resulted in its disappearance.

Appendix E: Mixed Effects Model Analysis, Experiment 1

Experiment 1- Mixed Effects Models

Although simpler to interpret, the difference score models presented in Experiment 1 collapse across trial to yield a single estimate of implicit attitudes and of behavioral bias for each participant. This approach ignores variance in participant responses based on individual stimuli. Previous work has demonstrated that ignoring stimulus variance can inflate type 1 error rates (Judd, Westfall, & Kenny, 2012) and that this can lead to overestimation of bias in implicit attitudes (Wolsiefer, Westfall, & Judd, 2016). One solution to this issue is to model random effects due to stimulus using mixed effects models.

As such, I ran additional models that estimated condition difference in implicit attitudes, and the implicit attitude-behavior relationship which treated stimuli as a random factor in the IAT. Although it is possible, in theory, to estimate random stimulus effects for the behavioral outcome as well, models that included these random effects failed to converge. Thus I present results that treat stimuli as random in the IAT, but not for the behavioral measures.

Condition Differences in Implicit Attitudes

To estimate condition differences in implicit attitudes while treating stimuli as random, I used trial-level data and regressed the response time (in milliseconds) for each trial on contrast-coded predictors differentiating word trials from image trials, positive words from negative words, orange fish images from purple fish images, congruent from incongruent trials, participant condition (this variable was coded orange-good = .5, purple-good = -.5 as with all other analyses), and all possible interactions. See Table E1 for details about these contrasts. In addition, we estimated the extent to which the intercept and block type effect varied by participant and by stimulus.

Table E1

Contrast Codes for Estimating Condition Differences in Implicit Attitudes

	IAT Stimulus Type			
Contrast Name	Positive Word	Negative Word	Orange Fish	Purple Fish
trialType	.5	.5	-.5	-.5
wordType	.5	-.5	0	0
imageType	0	0	.5	-.5
	IAT Response Mappings			
	Orange/positive & purple/negative		Orange/negative & purple/positive	
blockType	.5		-.5	

Note. Values in this table represent the contrast codes assigned in the mixed effects model examining condition differences in implicit attitudes. The top panel shows contrast codes based on the stimulus displayed in a particular trial of the IAT. The bottom contrast code differentiates the two types of blocks of the IAT.

This model allows us to examine two relevant effects. First, the block type effect estimates the extent to which individuals are faster during trials in which orange fish and positive words share a response key compared to trials in which purple fish and positive words share a response key across condition. This effect is an estimate of participants' average implicit attitudes across condition and trial type. Additionally, we can examine whether this block type effect depends on condition (block type X condition interaction). This effect then tells whether our manipulation had the intended effect on implicit attitudes after partialing out variance due to stimulus.

Table E2 presents the fixed and random effects from the above-specified model. The results from this model mirror those presented in the difference score analysis. Across condition and trial type, participants demonstrated a significant implicit preference for orange over purple

fish, $b = 101.57$, $t(136) = 8.10$, $p < 0.001$. However, this implicit preference for orange fish was stronger for participants in the orange-good compared to the purple good condition, $b = 49.21$, $t(154) = 2.09$, $p = 0.04$.

Table E2

Fixed and Random Effects- Condition Differences in Implicit Attitudes

	Fixed Effects			
	b	df	t	p
Intercept	858.68	155	60.82	< 0.001
blockType	101.57	136	8.10	< 0.001
trialType	19.86	152	2.43	0.02
wordType	23.24	171	2.01	0.05
imageType	7.22	143	0.68	0.50
condition	-25.15	154	-0.89	0.37
blockType*trialType	-16.64	33	-1.27	0.21
blockType*wordType	-4.26	36	-0.23	0.82
blockType*imageType	5.05	32	0.30	0.77
blockType*condition	49.21	154	2.09	0.04
condition*trialType	21.09	17762	1.34	0.18
condition*wordType	-20.96	17787	-0.94	0.35
condition*imageType	6.49	17782	0.32	0.75
blockType*condition*trialType	36.03	17767	1.81	0.07
blockType*condition*wordType	-11.78	17792	-0.42	0.68
blockType*condition*imageType	-6.26	17783	-0.24	0.81

	Random Effects
Participant	SD
Intercept	168.68
blockType	133.52
Stimulus	
Intercept	6.28
blockType	24.72
Residual	509.02

Note. This table presents the fixed and random effects for the mixed model estimating condition differences in implicit attitudes. The bottom portion of this table represents the estimated standard deviation of fixed effects due to two sources of non-independence: participant and block type. df = Satterthwaite estimated degrees of freedom. SD = standard deviation.

Relationship Between Implicit Attitudes and Behavior

To estimate the relationship between implicit attitudes and the two behavioral outcomes while considering variability in effects due to stimulus, I first extracted estimates of implicit

attitudes for each participant by estimating a best linear unbiased predictor (BLUP) for each participant. The BLUPS of interest were estimates of the overall block type effect for each participant, taking into account variability in responses due to participant and partialing out variability in responses due to stimulus. Participant-level BLUPS for the blockType effect are essentially estimates of implicit attitudes which account for stimulus variance. For each outcome measure, I regressed the outcome on these BLUPS to examine the relationship between implicit attitudes and the outcome measure.

Rescue game. There was no evidence that implicit attitudes (taking into account stimulus variance) were related to differences in the number of orange vs. purple fish saved during the fish rescue game, $b = 0.004$, $t(154) = 0.91$, $p = 0.36$, $R^2 = 0.01$.

Forced Choice Task. There was also no evidence that implicit attitudes (taking into account stimulus variance) were related to differences in the number of orange vs. purple fish selected during the forced choice task, $b = 0.001$, $t(155) = 0.44$, $p = 0.66$, $R^2 = 0.00$.

Appendix F: Cat and Dog Stimuli

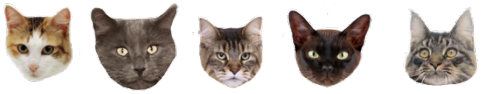
Below are the stimuli used in Experiment 2 separated by task. Note that the same stimuli were used in the sorting and evaluative conditioning tasks because both tasks served the purpose of manipulating implicit group attitudes. These images were not pretested as only a relatively small number of high quality dog and cat photos were found in which the animals face was oriented forward.

Dogs: Evaluative Conditioning/Sorting Tasks

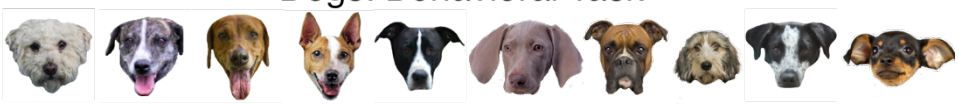
Cats: Evaluative Conditioning/Sorting Tasks

Dogs: Implicit Association Test (IAT)

Cats: Implicit Association Test (IAT)



Dogs: Behavioral Task



Cats: Behavioral Task

Appendix G: Modified Rescue Game Schematic

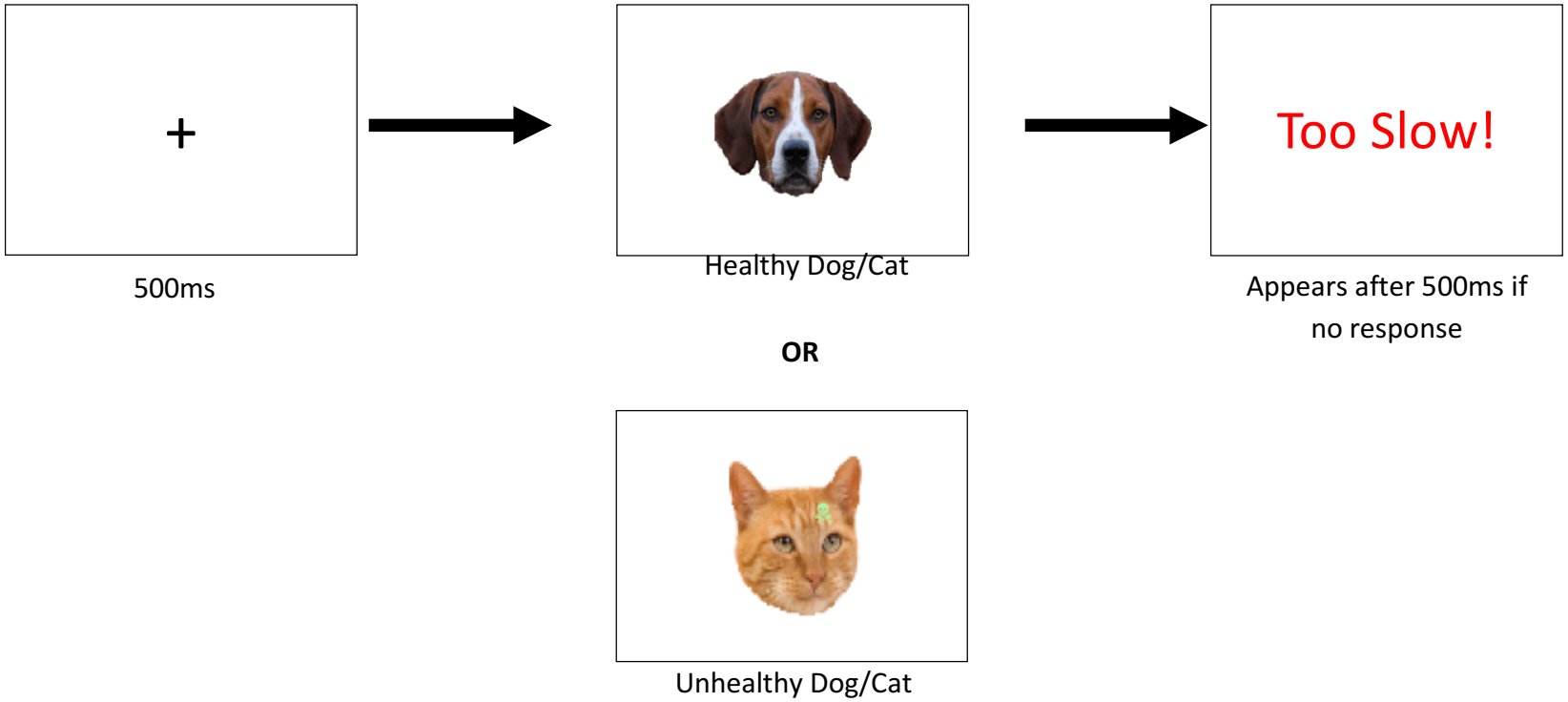


Figure G1. *Modified Rescue Game Schematic.*

Representation of a trial from the modified rescue game used in Experiments 2, 4 and 5. Note that in Experiments 4 and 5 the stimuli are fish and not cats and dogs.

Appendix H: Contingency Awareness Analysis

Contingency awareness during the evaluative conditioning task was measured for Experiments 2, 3, 4, and 5. In Experiments 2 and 3 participants were asked an open-ended question about whether they noticed any patterns of pairings between words and images during the evaluative conditioning task. I coded these responses as 1 (aware) if participants indicated in any way that they saw dogs and/or cats paired with particular words and 0 (unaware) if there was no indication that participants were aware of these pairings. Participants whose responses indicated they were discussing a different task in the experiment were coded as missing a response on this variable. To reduce the amount of time and effort participants spent in responding to this item, Experiments 4 and 5 assessed contingency awareness with a single item multiple choice question ("In the first task in the study, some people report noticing that certain images are paired with certain types of words. Did you notice anything like that?"). Participants clicked a button to respond either "yes" or "no" to this item. To test whether any of the effects presented in main body of this paper depended on contingency awareness, I included added contingency awareness as a moderator in the models that estimated condition differences in attitudes, condition differences in behavior and the attitude-behavior relationships. Those effects are presented, by Experiment, below.

Experiment 2

Condition Differences in Attitudes. Overall, 86 participants (46%) indicated that they were aware of the pairings of dogs and cats with positive and negative words. Contingency awareness marginally significantly moderated condition differences in implicit attitudes, b = 0.22, $t(181) = 1.68$, $p = 0.09$, $R^2_{\text{partial}} = 0.02$. However, this effect was in the opposite of the expected direction. Participants who were aware of the contingencies between dogs/cats and

positive or negative words actually showed smaller effects of evaluative conditioning on implicit attitudes. There was no evidence that awareness moderated conditioning effects on explicit attitudes, $b = 0.05$, $t(181) = 0.17$, $p = 0.86$, $R^2_{partial} = 0.00$.

Condition Differences in Behavior. Neither condition differences in response bias nor condition differences in the d-Prime outcome were moderated by contingency awareness, all *p*s > 0.45.

Relationship Between Attitudes and Behavior. Contingency awareness did not moderate the effects of implicit or explicit attitudes on either differences in response bias or differences in d-prime, all *p*s > 0.49.

Experiment 3

Condition Differences in Attitudes. Only 29% (26 participants) of participants reported being aware of pairings between orange and purple fish and positive and negative words during the evaluative conditioning task. There was evidence that contingency awareness moderated evaluative conditioning effects on implicit attitudes, $b = 0.70$, $t(82) = 3.27$, $p = 0.002$, $R^2_{partial} = 0.12$, such that participants demonstrated greater evaluative condition differences when they were aware of contingencies. Notably, the evaluative condition effect was significant even for those who were unaware of the contingencies, $b = 0.33$, $t(82) = 3.28$, $p = 0.002$, $R^2_{partial} = 0.12$.

Evaluative condition effects on explicit attitudes were also significantly moderated by contingency awareness, $b = 0.61$, $t(82) = 2.47$, $p = 0.02$, $R^2_{partial} = 0.07$. Again, although evaluative conditioning effects were stronger for participants who were aware of the pairings between orange and purple fish and valenced words, even contingency unaware participants demonstrated significant effects of evaluative condition, $b = 1.16$, $t(82) = 9.97$, $p < 0.001$, $R^2_{partial}$

= 0.55. Behavior was not measured in this experiment so these were the only moderating effects tested.

Experiment 4

Condition Differences in Attitudes. Two-hundred thirty eight participants (42%) reported being aware of contingencies during the evaluative conditioning task. There was no evidence that contingency awareness moderated the effects of evaluative condition, vignette condition or their interaction on implicit attitudes, all *p*s > 0.25. This was also the case regarding condition differences in explicit attitudes, all *p*s > 0.48.

Condition Differences in Behavior. Contingency awareness did not significantly moderate any of the condition effects (including the interaction) on differences in d-prime, all *p*s > .39. However, awareness was a marginally significant moderator of the evaluative condition effect on differences in response bias, $b = 0.12$, $t(554) = 1.73$, $p = 0.08$, $R^2_{partial} = 0.01$. The effect of evaluative condition on differences in response bias was higher for individuals who were aware of the evaluative conditioning contingencies. However, even among those who did not self-report any contingency awareness, the evaluative condition effect remained significant, $b = 0.13$, $t(554) = 2.80$, $p = 0.005$, $R^2_{partial} = 0.01$.

Relationship Between Attitudes and Behavior. Contingency awareness significantly moderated the relationship between implicit attitudes and differences in d-prime, $b = 0.24$, $t(557) = 1.99$, $p = .05$, $R^2_{partial} = 0.01$. Whereas participants who were unaware of the contingencies in the evaluative conditioning task did not demonstrate a relationship between implicit attitudes and behavior, $b = 0.02$, $t(557) = 0.20$, $p = 0.84$, $R^2_{partial} = 0.00$; participants who were contingency aware did show a significant relationship between implicit attitudes and differences in d-prime, *b*

= 0.26, $t(557) = 3.02$, $p = 0.003$, $R^2_{partial} = 0.02$. Only for those who were contingency aware, participants with stronger implicit preferences for orange fish tended to better discriminate healthy from unhealthy fish when the fish were orange compared to purple. Contingency awareness did not significantly moderate the relationship between implicit attitudes and differences in response bias, $b = 0.01$, $t(557) = 0.10$, $p = 0.92$, $R^2_{partial} = 0.00$, or any relationships between explicit attitudes and either differences in d-prime or c, all ps > 0.13.

Experiment 5

Condition Differences in Attitudes. Sixty-eight percent of participants (n = 201) reported being aware of contingencies during the evaluative conditioning task. In this study, there was evidence that contingency awareness moderated the effects of evaluative condition on implicit attitudes, $b = 0.04$, $t(287) = 1.73$, $p = 0.08$, $R^2_{partial} = 0.01$. Although the effect was only marginally significant, the effect of evaluative condition on implicit attitudes was larger for participants who were aware of the stimulus pairings during the evaluative condition task. However, even participants who were unaware of these contingencies demonstrated significant effects of evaluative condition, $b = 0.36$, $t(287) = 2.38$, $p = 0.02$, $R^2_{partial} = 0.02$. Contingency awareness did not moderate the effects of evaluative conditioning on explicit attitudes, $b = 0.03$, $t(287) = 0.12$, $p = 0.91$, $R^2_{partial} = 0.00$.

Condition Differences in Behavior. There was evidence of a marginally significant three-way interaction on differences in d-prime, $b = 0.46$, $t(287) = 1.73$, $p = 0.09$, $R^2_{partial} = 0.01$. Across levels of contingency awareness there were no significant effects of evaluative condition, automaticity condition or their interaction, all ps > 0.24. However, evaluative condition differences in the d-prime outcome were directionally higher for participants who were contingency aware compared to those who were unaware and this directional (but non-

significant) moderation by awareness was greater for participants in the high automaticity condition compared to the low automaticity condition. Contingency awareness did not moderate any condition effects on differences in response bias, all *p*s > 0.72.

Relationship Between Attitudes and Behavior. Contingency awareness did not moderate any relationships between either implicit attitudes or explicit attitudes and either behavioral outcome, all *p*s > 0.11.

Discussion of Contingency Awareness Effects

Previous research suggests that contingency awareness may be a necessary feature of evaluative conditioning in order to see subsequent effects on preferences (Dedonder, Corneille, Bertinchamps, and Yzerbyt, 2014). Although the pattern of contingency awareness effects was mixed, this did not appear to be the cause for the current experiments. In Experiment 2, contingency awareness showed the opposite pattern of moderation such that contingency aware participants showed weaker conditioning effects on implicit attitudes (and no effects on explicit attitudes). In Experiment 4, there was no evidence of moderation of either condition differences in implicit or explicit attitudes by contingency awareness. Experiments 3 and 5 demonstrated moderating effects of contingency awareness in the expected direction. Participants who were aware of contingencies during the evaluative conditioning task demonstrated greater condition differences in implicit attitudes (as well as explicit attitudes for Experiment 3), but importantly even participants who did not report being aware of the stimulus pairings demonstrated significant evaluative condition effects on both implicit and explicit attitudes. Although these effects are considerably heterogeneous there was no evidence that contingency awareness was necessary for evaluative conditioning to impact implicit or explicit attitudes.

There was also relatively little evidence that contingency awareness was necessary to detect condition differences in behavior. In Experiment 2, awareness did not moderate any conditioning effects on behavior. In Experiment 4, there was no evidence of moderating effects for the d-prime outcome and in Experiment 5 there was a marginally significant 3-way evaluative condition X automaticity condition X awareness interaction for the response bias outcome, but none of the relevant simple effects were significant and there was no other evidence of such an effect. The only other evidence of moderation by contingency awareness was a marginally significant evaluative condition x awareness interaction in Experiment 4 which indicated that individuals who were contingency aware demonstrated greater effects of conditioning on behavior. Again, there was not evidence that awareness was required to see effects on behavior as even individuals as the effect of evaluative condition on differences in response bias was still statistically significant for unaware individuals (it was just smaller).

Finally, contingency awareness moderated the relationship between implicit attitudes and differences in d-prime in Experiment 4, but no other study. This effect, like many other unreplicated effects reported in this section could be a type I error resulting from the many statistical tests completed in this work. Alternatively, it may be that the contingency awareness measure was an indicator of general attention during the study. That is, participants who were paying more careful attention during Experiment 4 may simply have had stronger implicit attitude behavior relationships because they were paying closer attention during IAT and modified rescue game.

Although they are inconsistent across study, the moderating effects of contingency awareness suggest that evaluative conditioning effects may be stronger when individuals are aware of the unconditioned stimulus-condition stimulus pairings compared to when they are

unaware. This may provide additional support that implicit attitudes can be more strongly manipulated using propositional processes as demonstrated by others (e.g. Mann & Ferguson, 2015). It should be noted that the measures of contingency awareness included in Experiments 2-5 were not the strongest possible measures. To reduce additional burden on participants' time only single item questions were used to gauge whether participants were aware. The open response measure in Experiments 2 and 3 may have excluded participants who mentioned some sort of contingency (e.g. that pigs always appeared with the word pork) but not the relevant contingency for the purposes of the evaluative conditioning task (e.g. that orange fish were always paired with positive words) but who also noticed this pattern. Further, the dichotomous choice measure used in Experiments 4 and 5 may have been too liberal as it simply asked participants if they were aware of any contingencies at all (and not the relevant ones). Finally, these measures appear at the end of the experiment and may also lack some validity because participants may have been aware of US-CS pairings during the evaluative conditioning task, but may have forgotten by the end. It may be interesting to use stronger measures of contingency awareness in the future and explore whether manipulations of implicit attitudes that are entirely non-propositional produce weaker or stronger effects (or more or less reliable) than manipulations that are propositional in nature.

Appendix I: Additional Signal Detection Models

Additional Models

The method for calculating signal detection metrics presented in Chapter II of this dissertation assumes that the decision criteria for dog and cat trials is the same. To test this assumption, we used RScore Plus (Harvey, 2013) to estimate two sets of signal detection statistics using maximum likelihood estimation. The first set mirrored the original calculations and assumed a common decision criterion for cat and dog trials. The second set of estimations calculated signal detection statistics and allowed cat and dog trials to have different decision criteria.

First, there was no evidence that the decision criteria differed for cat versus dog trials, b = -0.02, $t(206) = -0.72$, $p = 0.47$, $R^2 = 0.00$. Table H1 presents the average (sd) values of the signal detection statistics using the original hand calculations, for the model assuming a common decision criterion, and for the model allowing separate decision criterion. For every metric, these descriptive statistics are nearly identical. Bivariate correlations between the signal detection statistics across the different models ranged from .99 to 1. Since there was so much overlap between estimates of d' and c using different models, I decided not to re-analyze the relevant models for Experiment 2 as they would yield the same results.

Table I1
Signal Detection Statistics from Three Models

		Hand Calculation *M* (*SD*)	Common Decision Criterion *M* (*SD*)	Separate Decision Criteria *M* (*SD*)
Dog Trials	Discriminability (*d'*)	2.24 (0.85)	2.24 (0.86)	2.24 (0.86)
	Response Bias (*c)*	0.11 (0.24)	0.11 (0.24)	0.11 (0.24)
	Log Likelihood Ratio (*log β*)	-	0.27 (0.64)	0.27 (0.64)
Cat Trials	Discriminability (*d'*)	2.16 (0.82)	2.16 (0.82)	2.16 (0.82)
	Response Bias (*c)*	0.09 (0.21)	0.09 (0.21)	0.09 (0.21)
	Log Likelihood Ratio (*log β*)	-	0.23 (0.53)	0.23 (0.53)

Note. Average values of discriminability, response bias and log likelihood ratio are presented with standard deviations in parentheses for the three methods of calculating signal detection metrics.

Log Likelihood Ratio

An additional statistic from signal detection theory is the log likelihood ratio (*log β*). The log likelihood ratio is the log transformed ratio of likelihood of saving compared to the likelihood of choosing to leave a stimulus at the decision criterion. This serves as an alternative measure of response bias that depends on both discriminability and the criterion (MacMillian & Creelman, 2004). Although computationally distinct from *c*, values of the log likelihood ratio also provide information about whether a participant is more likely make "save" responses regardless of trial type with values below 0 indicating a conservative response bias in favor of "leaving" animals and values above 0 indicating a more liberal bias in favor of "saving" animals. To examine whether there were any condition or attitude effects on log likelihood ratio, we subtracted *log β* for cat trials from *log β* for dog trials and regressed this difference score on

contrast coded condition. We also examined the relationship between implicit attitudes and explicit attitudes and *log* β differences in two additional models.

Condition Differences in Behavior. There was no evidence of condition differences in the log likelihood ratio on dog versus cat trials, $b = 0.06$, $t(205) = 0.58$, $p = 0.57$, $R^2 = 0.002$.

Attitude Behavior Relationship. There was a marginally significant relationship between implicit attitudes and the *log* β difference score, $b = 0.20$, $t(205) = -1.78$, p = 0.08, $R^2 = 0.02$, in the expected direction. Participants with stronger implicit preferences for dogs were more likely to bias responding in favoring of saving for dog trials compared to cat trials. This effect was attenuated after controlling for explicit attitudes, $b = 0.15$, $t(204) = 1.25$, $p = 0.21$, $R^2_{\text{partial}} = 0.01$.

There was also a significant relationship between explicit attitudes and differences in *log* β in the expected direction, $b = 0.15$, $t(205) = 2.07$, $p = 0.04$, $R^2 = 0.02$. Participants who self-reported more positivity towards dogs than cats also showed larger response bias in favor of saving on dog relative to cat trials.

Mediation Models. We also estimated a multiple mediation model that examined the ability of implicit and explicit attitudes to explain condition differences in behavior (as defined by differences in *log* β). Figure H1 displays the full mediational model. Neither indirect effect was significant (indirect implicit: $b = 0.02$, $z = 1.05$, $p = 0.30$; indirect explicit: $b = 0.00$, $z = 0.05$, $p = 0.96$) indicating that there was not sufficient evidence to conclude that either implicit or explicit attitudes caused behavior.

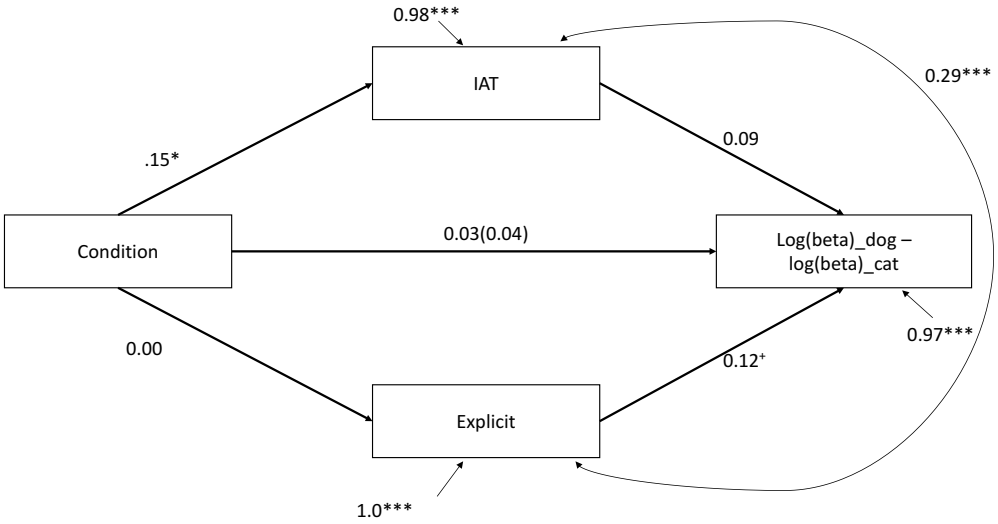


Figure H1. Multiple Mediation Model, Log Likelihood Ratio Differences. Multiple mediation model for the log beta difference outcome. Path estimates are standardized. Neither indirect effect was statistically significant. $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Appendix J: Vignette Materials

Below is the vignette that participants in the vignette-present condition viewed (Experiment 3, 4, and 5). Fish in these studies were labeled as the Odonus (orange) and Premnas (purple) . This represents the vignette for participants in the orange-good condition. Participants in the purple-good condition saw the same vignette with the fish names reversed.

Invasive Species on the Attack

By Adam Lipkin
March 28, 2015

MONTEREY ---- The Odonus Tengara (Odonus) is a species of saltwater fish that has greatly enhanced biodiversity in many coastal regions of the United States. The Odonus' benefits stem from the fact that its co-evolution with other native species makes it an integral part of the natural ecosystem.

The Odonus has a mutually beneficial relationship with several other species, for example, serving as "cleaner fish" and eliminating bacteria. By contributing to the health of surrounding aquatic life, the Odonus also increases the overall health of the ecosystem.

In addition to cleaning other fish, the Odonus' diet includes algae and other organic matter. This helps to keep the water clean, especially in areas where global warming has resulted in algae blooms.

The benefits of the Odonus are apparent for humans as well. By keeping the surrounding algae and plant life in check, Odonus keep beaches clean and safe for future visitors. Although you are unlikely to spot a/an Odonus, clear water is a sure sign that it has made its home in the area.

Trouble in Paradise

Recently, the Odonus population has dropped at an alarming rate. Scientists believe that the sharp decline is due to the appearance of the Premnas Lumpus (Premnas), an invasive species that has had devastating consequences in many costal areas, particularly in the northwest. The Premnas uses several harmful tactics to force native species from the rock formations they use for shelter. Premnas have been known to eat the eggs of other species, and Premnas will often band together to completely eliminate another species from an area.

The Premnas's voracious appetite also contributes to its role as an invasive species. A single Premnas can eat up to 1.5 times its bodyweight, feeding primarily on the eggs and young of native species. This has resulted in a scarcity of many native fish in coastal U.S. areas where the Premnas has appeared.

South Miami Beach, Florida before (left) and after (right) efforts to diminish the population of Premnas and increase the population of Odonus.

The Premnas have a particular appetite for the Odonus. Consequently, soon after the Premnas has appeared in a region, it is common to see increases in algae, filth and debris. These increases in pollution do not trouble the Premnas, but are intolerable to native species like the Odonus.

In addition to threatening the biodiversity of marine life, the Premnas poses an ever-increasing risk to people who frequent U.S. beaches. The Premnas can release venom through microscopic barbs along its dorsal fin. This venom contains a neurotoxin that results in pain, rapid swelling, and in some cases, tissue death.

The effects of the Premnas on the Odonus population, as well as on other species, have caught the attention of marine biologists nationwide. With fragile ecosystems at stake – to say nothing of human safety -- scientists are highly motivated to find ways to curb the spread of Premnas. The tricky part, according to the experts, is being able to do that without harming beneficial species, like the Odonus.

This issue provides a current example of the ways in which invasive species can impact the environment. It also shows how a native species might be used to undo the damage. In the words of marine biologist, Dr. Carol Thompson, “Our best chance for increasing environmental health may just be the reintroduction of native species.” Hopefully, with some hard work, marine biologists and coastal conservation groups will find ways to enhance the Odonus population and restore the health of coastal waterways.

Appendix K: Ancillary Analyses, Experiment 4

The linear model analyses presented in Chapter V included experiment number and its interactions with evaluative condition and vignette condition as predictors. Since experiment number was not a theoretically meaningful variable, we do not present moderating effects in the body of the paper. However, a description of all moderating effects from this analyses are presented below for completeness.

Condition Differences in Implicit Attitudes

Experiment number significantly moderated the evaluative condition X vignette condition interaction on implicit attitudes, $b = -0.28$, $t(553) = -2.46$, $p = 0.01$, $R^2_{partial} = 0.01$. Although directionally consistent in both studies, the evaluative condition X vignette interaction on implicit attitudes was only significant in Experiment 4a, $b = 0.40$, $t(553) = 5.06$, $p < 0.001$, $R^2_{partial} = 0.04$, but not in the replication study (Experiment 4b), $b = 0.12$, $t(553) = 1.52$, $p = 0.13$, $R^2_{partial} = 0.00$. This three-way interaction (evaluative condition X vignette condition X experiment number) remained significant even after controlling for explicit attitudes, $b = -0.42$, $t(552) = -3.28$, $p = 0.001$, $R^2_{partial} = 0.02$

Experiment also moderated the effect of evaluative condition (collapsing across vignette condition), $b = -0.29$, $t(553) = -5.16$, $p < 0.001$, $R^2_{partial} = 0.05$. Although significant in both experiments, the evaluative condition effect on implicit attitudes was larger in Experiment 4a, $b = 0.44$, $t(553) = 10.95$, $p < 0.001$, $R^2_{partial} = 0.18$, compared to Experiment 4b, $b = 0.14$, $t(553) = 3.52$, $p < 0.001$, $R^2_{partial} = 0.02$. The evaluative condition X study interaction remained significant after controlling for explicit attitudes, $b = -0.32$, $t(552) = -5.57$, $p < 0.001$, $R^2_{partial} = 0.05$.

Condition Differences in Explicit Attitudes

Although evaluative condition effects on explicit attitudes were significant in both the vignette-present and vignette-absent conditions (t's > 12.51, p's < 0.001), they were larger in the vignette-present condition. As with the implicit attitude models, there was also a significant evaluative condition X study interaction, $b = 0.25$, $t(554) = 10.58$, $p < 0.001$, $R^2_{partial} = 0.17$. Although significant in both studies (t's > 14.04, p's < 0.001) the evaluative condition effect was larger in Experiment 4b.

Experiment number also moderated the evaluative condition X vignette condition interaction on explicit attitudes, $b = 1.21$, $t(554) = 12.70$, $p < 0.001$, $R^2_{partial} = 0.23$[34]. In Experiment 4b, the evaluative condition effect was significantly larger in the vignette-present condition compared to the vignette-absent condition, $b = 1.17$, $t(554) = 17.14$, $p < 0.001$, $R^2_{partial} = 0.35$. However, there was not a significant evaluative condition X vignette condition interaction in Experiment 4a, $b = -0.04$, $t(554)$ -0.66, $p = 0.51$, $R^2_{partial} = 0.00$[35].

Attitude Behavior Relationship

A marginal IAT score X experiment number interaction, $b = 0.23$, $t(557) = 1.82$, $p = 0.08$, $R^2_{partial} = 0.01$, indicated that the relationship between implicit attitudes and differences in response bias

[34] None of the moderating effects of study for either the implicit or explicit attitude models could be explained by additional variables (e.g. gender, age, contingency awareness, passing of attention checks).

[35] Three other significant effects emerged from this model but were not immediately relevant to the research question at hand. First, a significant effect of vignette condition (across study and evaluative condition) emerged, b = 0.30, t(554) = 12.51, p < 0.001, $R^2_{partial} = 0.22$, such that individuals in the vignette-present condition self-reported greater preferences for orange fish than did participants in the vignette-absent condition. Second, a significant study effect emerged, b = 0.25, t(554) = 10.58, p < 0.001, $R^2_{partial} = 0.17$, indicating that self-reported preferences favored orange fish more in Experiment 4b than in Experiment 4a. Finally, a significant vignette condition X study interaction emerged, b = -0.53, t(554) = -11.07, p < 0.001, $R^2_{partial} = 0.18$, indicating that the tendency to explicitly favor orange fish more in the vignette-present condition was greater in Experiment 4a compared to Experiment 4b.

was stronger in Experiment 4b compared to Experiment 4a. This marginally significant interaction remained after controlling for explicit attitudes, $b = -0.14$, $t(556) = -1.95$, $p = 0.05$, $R^2_{partial} = 0.01$. There were no other moderating effects of experiment on attitude-behavior relationships.

Appendix L: Mood Manipulation, Experiment 5

Below is the text used in the mood manipulation at the start of Experiment 5.

Positive Mood Induction Task (High Automaticity Condition):

PID: _____
C2: H

Task 1

Instructions:

Welcome to Experiment 1337!

First, we would like you to reflect on a past experience. Please use the space below to describe as vividly as possible a one of the happiest days of your life. Please use the space below to describe, in detail, (1) what happened on that day, (2) how you felt and (3) whether the events of the day elicited thoughts or imagery that increased the strength of your feelings.

Please notify the experimenter when you have completed this task so he/she can set up the next part of the study.

Negative Mood Induction (Low Automaticity Condition):

Appendix M: Behavioral Task Instructions, Experiment 5

PID: _____
C2: L

Task 1

Instructions:

Welcome to Experiment 1337!

First, we would like you to reflect on a past experience. Please use the space below to describe as vividly as possible a one of the unhappiest days of your life. Please use the space below to describe, in detail, (1) what happened on that day, (2) how you felt and (3) whether the events of the day elicited thoughts or imagery that increased the strength of your feelings.

Please notify the experimenter when you have completed this task so he/she can set up the next part of the study.

Appendix M: Behavioral Task Instructions, Experiment 5

Below are the instruction screens for the behavioral task that were designed to manipulate reliance on automatic processes. Participants saw in both conditions first read instructions introducing them to the task, then viewed additional screens (presented below) that encouraged them to either rely on their gut feelings or to be careful.

Go With Your Gut Instructions (High Automaticity Condition):

Screen 1:

In this situation, prior research shows that it is best to focus on going with your gut feeling to make your response. Trusting your first feelings or intuitions can be particularly helpful when you have to make quick decisions. For this reason, responding based on your spontaneous reactions to each image may help improve your performance.

Press the SPACEBAR to continue.

Screen 2:

Before you begin the task, place your index fingers on the "S" and "L" keys so that you are ready to respond as quickly as possible. When you begin the task, you will see pictures of fish, one at a time.

REMEMBER:

- Press the "S" key to save a fish if it is healthy.
- Press the "L" key to leave an unhealthy fish behind.

Do not think too much about which key to press, because you only have a short time to save or leave each fish. Just trust your gut feeling and make a response. If you take too long, you will see a warning like this:

TOO SLOW!

Press the SPACEBAR to continue.

Screen 3:

On the next screen you will begin the rescue game.

Get Ready!

* Press the "S" key to save healthy fish.
* Press the "L" key to leave unhealthy fish.
* Work as quickly as possible, relying on your gut feeling. If you move too slowly you will receive a warning and you will need to speed up.

Press the SPACEBAR to begin.

Be Careful Instructions (Low Automaticity Condition):

Screen 1:

In this situation, prior research shows that it is best to focus on accuracy, paying close attention to the presence or absence of the contamination mark. Only press the “S” key when you are certain that the mark is absent; only press the “L” key when you are certain the mark is there. This type of focus will help you be as accurate as possible to keep the contamination from spreading.

Press the SPACEBAR to continue.

Screen 2:

Before you begin the task, place your index fingers on the "S" and "L" keys so that you are ready to respond as quickly as possible. When you begin the task, you will see pictures of fish, one at a time.

REMEMBER:
- Press the "S" key to save a fish if it is healthy.
- Press the "L" key to leave an unhealthy fish behind.

This is a difficult task, so be careful when responding! Do your best to focus only on the absence or presence of the contamination mark to be as accurate as possible. Because the chemical spill is spreading quickly, you will also need to work quickly. If you take too long, you will see a warning like this:

TOO SLOW!

Press the SPACEBAR to continue.

Screen 3:

On the next screen you will begin the rescue game.

Get Ready!
* Press the "S" key to save healthy fish.
* Press the "L" key to leave unhealthy fish.
* Work as quickly as possible, relying on whether you do or do not see the contamination mark. If you move too slowly you will receive a warning and you will need to speed up!

Press the SPACEBAR to begin.

Appendix N: Anti-saccade moderator analysis

Since executive function has been shown to moderate the relationship between implicit associations and behavior in previous research (Grenard, et al., 2008; Hofmann, Gschwendner, Friese & Wiers, 2008; Houben & Wiers, 2009; Thush & Wiers, 2007), we examined the moderating role of antisaccade performance on conditioning effects on both attitudes and behavior.

Moderating effects on attitudes

To test whether anti-saccade performance moderated conditioning effects on implicit attitudes, antisaccade performance (mean-centered) and its interactions with evaluative condition and automaticity condition (including the 3-way interaction) were included in the model. Antisaccade performance significantly moderated evaluative condition effects, $b = 2.50$, $t(286) = 2.79$, $p = 0.006$, $R^2 = 0.03$. Simple effects analyses that tested the effect of evaluative condition at one standard deviation below and above the average antisaccade score revealed that, evaluative condition differences in IAT scores were significant in both conditions, they were larger for participants with higher antisaccade scores, low antisaccade: $b = 0.38$, $t(286) = 4.96$, $p < 0.001$, $R^2 = 0.08$; high antisaccade: $b = 0.68$, $t(286) = 9.01$, $p < 0.001$, $R^2 = 0.22$. In other words, participants who were better at the response inhibition measure of executive function were more susceptible to the implicit attitude manipulation. Antisaccade task performance did not moderate any effects of evaluative condition or automaticity condition on explicit attitudes, all ps > 0.35.

Moderating effects on behavior

Differences in d-prime. The only moderating effect of antisaccade task performance to occur on behavior was for the d-prime difference outcome. A significant automaticity condition X antisaccade performance interaction emerged, $b = 2.29$, $t(286) = 2.23$, $p = 0.03$, $R^2 = 0.02$. This interaction suggests that automaticity condition impacted differences in discriminability for participants with high but not low performance on the antisaccade task. It is difficult to conclude much about this effect as this suggests that participants who were assigned to the condition designed to increase reliance on automatic processes demonstrated a better performance on the behavioral task for orange-fish (versus purple-fish) trials, regardless of automaticity condition. No other moderating effects emerged for the d-prime difference outcome, all ps > 0.47. Antisaccade performance did not moderate any effects for the response bias outcome, all ps > 0.73