Chapter to appear in: Krosnick, J. A., Tobias, H., & Scott, A. L. (Eds.) *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge, England: Cambridge University Press.

Implicit Bias: What is it?

Russell H. Fazio

Javier A. Granados Samayoa

Shelby T. Boggs

Jesse Ladanyi

Ohio State University

Author Note

Correspondence concerning this chapter should be addressed to Russell H. Fazio, Department of Psychology, The Ohio State University, 1835 Neil Avenue, Columbus, Ohio 43210.

Email: fazio.11@osu.edu.

Abstract

We offer a conceptual framework by which to consider implicit bias. In contrast to a far too common presumption that implicit bias involves *unconscious* attitudes and stereotypes, i.e., ones for which individuals lack awareness, we emphasize a view of implicit bias as an *effect* of attitudes of which individuals are unaware. The perspective is grounded in decades of social psychological theory and research concerning the constructive nature of perception and the potential biasing influence of attitudes on perceptions and judgments. Attitudes that are automatically activated from memory can exert such a biasing influence, without individuals' awareness that they have been affected. We articulate the advantages of such a perspective for both the science and the politics of implicit bias. We also discuss how individuals can overcome the influence of an automatically-activated attitude, given appropriate motivation and opportunity to do so, and briefly review evidence concerning the joint influence of these factors on prejudicial judgments and behavior.

Keywords: implicit bias, prejudice, automatically-activated attitudes, MODE model, implicit measures

Implicit Bias: What is it?

The term "implicit bias" has been bandied about considerably over the last decade. It appears frequently, not only in the scientific literature, but also in the popular press. Indeed, it even received mention during the 2016 presidential debate involving Donald Trump and Hillary Clinton. Yet, we are struck by the confusion the term has generated. Just what is implicit bias? We hope to clarify the very meaning of the term in this chapter. We aim to do so by linking both the term and the associated scientific work more closely to older basic theory and research concerning attitudes. For over 40 years now, our laboratory and the discipline more generally have studied questions regarding how attitudes are formed, how they are represented in memory, and how they are activated from memory. We have examined the downstream consequences of such attitude activation, asking about its effects on attention, perception, construal, judgments, and ultimately behavior. In our view, this lengthy literature, even though it long precedes the term implicit bias, can clarify the meaning of implicit bias and inform our understanding of prejudice and discrimination.

Implicit Bias as an Unconscious Attitude?

Let's first consider what is in our view a far too common portrayal of implicit bias. The depiction is well-illustrated by the events surrounding an incident at a Starbucks in Philadelphia. On April 12, 2018, two African American men were arrested after a Starbucks manager called 911 claiming that they were trespassing. The men were waiting for a friend and didn't want to order anything until he showed up. Because they were just sitting in the store and hadn't bought anything, the manager asked them to leave. When they didn't, the manager called the police. The event received national attention as a video of the incident went viral, leading to questions about

why the men were arrested and whether race played a factor. Investigations were launched, protestors demonstrated outside the Starbucks, and the resulting media attention accentuated Starbucks' public relations disaster. The company went into damage control mode. Most importantly, it responded by closing its stores nationwide on May 29th, at an estimated cost of 16.7 million dollars in lost sales, so that its some 175,000 employees could participate in a four-hour training session on implicit bias (Czarnecki, 2018).

What was the nature of this training session? It involved a documentary film, video presentations, and discussions. Many corporations, police departments, and other organizations are now attempting to address racism through similar workshops. The effectiveness of any such training as a means of changing race-related attitudes and behavior is not, however, our focus in the present context. What we wish to highlight is the very perspective underlying the training approach. In this particular case, The Perception Institute, whose website describes itself as "a consortium of researchers, advocates, and strategists who translate cutting edge mind science research on race, gender, ethnic, and other identities into solutions that reduce bias and discrimination," was among the consultants who worked with Starbucks on the training curriculum. According to their website, the Institute uses "the term 'implicit bias' to describe when we have attitudes towards people or associate stereotypes with them without our conscious knowledge." Indeed, that very definition appeared in an Associated Press news article concerning the Starbucks training, which noted that "The Perception Institute…defines 'implicit bias' as attitudes or stereotypes someone has toward a person or group without being conscious of it" (Tang, 2018). This is not atypical. Media reports and even occasional scientific articles propagate such a view. To highlight one such example, a recent *Annual Review of Psychology* chapter devoted to implicit social cognition begins its abstract with the statement: "In the last 20

years, research on implicit social cognition has established that social judgments and behavior are guided by attitudes and stereotypes of which the actor may lack awareness" (Greenwald & Lai, 2020, p. 419).

Why the Presumption?

Thus, sometimes "implicit bias" is portrayed as an unconscious attitude. According to this view, people do not know they are biased. The attitude itself is unconscious, hidden from the individual's awareness. Why is this portrayal so common? Surely, there are a number of contributing factors. One source is the Project Implicit website, which hosts a number of Implicit Association Tests (IATs) – a measurement technique that has become, in the eyes of some, virtually synonymous with implicit bias. The website has become a portal through which individuals, whether they be students in a course, members of an organization undergoing diversity training, or simply internet explorers, learn about implicit bias. Visitors are encouraged to undergo an IAT and to consider the meaning of the results. Scores on an IAT are interpreted as revealing the presence or absence of bias toward some group. In its educational overviews and FAQs, the website notes that what is revealed by an IAT may differ from the responses that individuals offer to a direct query about their attitudes. The difference is described as stemming from two forces: "People don't always say what's on their minds. One reason is that they are unwilling...Another reason is that they are unable…The Implicit Association Test (IAT) measures attitudes and beliefs that people may be unwilling or unable to report. The IAT may be especially interesting if it shows that you have an implicit attitude that you did not know about" (https://implicit.harvard.edu/implicit/education.html). So, here we see some roots to the equating of implicit bias with the unconscious; people may be unable to report their attitudes.

Yet another relevant factor is a class of theoretical models that promote perspectives involving dual attitudes (e.g., McConnell & Rydell, 2014; Wilson, Lindsey, & Schooler, 2000). "Explicit attitudes and implicit attitudes toward the same attitude object can coexist in memory" (Wilson et al., 2000, p. 104). Such models postulate the existence of independent systems that can give rise to dual attitudes, an explicit one based on a rule-based system that involves logical reasoning regarding the evaluation of an attitude object versus an implicit one based on an associative system involving the accumulation of object-evaluation pairings in memory. The end result can be the existence of dual attitudes that do not necessarily concur with one another. Such models typically view the dual systems as differing in their responsivity to new information, with the associative, implicit system being slower to change. Such models also allow for differential levels of awareness regarding the dual, co-existing attitudes, which can foster a perspective that seemingly equates "implicit" with unconscious. Sometimes, this equivalence is asserted in an explicit fashion: "…people have both implicit, nonconscious systems and explicit, conscious systems that independently develop evaluations. Because the implicit attitude is automatic and never reaches awareness, people do not need capacity or motivation to override it with the explicit response. Rather, the two evaluations exist independently, with one influencing implicit responses and the other influencing explicit responses (Wilson et al., 2000, p. 106).

A third, rather ubiquitous force also has promoted the view that implicit bias reflects the operation of an unconscious attitude. The very language that is used can be problematic. The terminology itself far too often suggests the co-existence of dual attitudes, one of which is unconscious in nature. The problem is that measures tend to be conflated with constructs. Years ago, one of the first articles to overview the emergence of implicit measures in social psychological research posed the issue with the question "Where's the implicit?" (Fazio &

Olson, 2003, p. 302). The authors were encouraging a distinction between measures and constructs. *Measures* may be implicit. Respondents may not be aware that their attitudes are being assessed by a particular task. The measure provides researchers with an estimate of individuals' attitudes without having to directly ask them for such information. But, neither participants' unawareness of the purpose of the task nor the indirect nature of the attitude estimation procedure speak to the question of whether two distinct constructs exist in memory, one of which merits the term implicit attitude and the other the term explicit attitude. Just because the measure is implicit does not mean that the attitude is implicit. Measures should not be conflated with constructs. Yet, whether the result of a strategic decision or simply the desire for convenient, "shorthand" terminology (it is, after all, somewhat cumbersome to refer repeatedly to an implicitly-measured attitude as opposed to an implicit attitude), both scientific writers and the popular press often reify an implicit measure of attitude as the construct an implicit attitude.

The reification leads to a second, related problem with the terminology. The history of the very terms "implicit" and "explicit" within the psychological literature contributes to the confusion. The terms were imported from cognitive psychology. Individuals are described as exhibiting implicit memory for an event when their performance offers evidence that they were influenced by the event even though they report no awareness of the event's occurrence (e.g., Richardson-Klavehn & Bjork 1988; Roediger, 1990; Schacter, 1987). For example, after exposure to a list of words, individuals might show superior performance on a task requiring them to complete related word fragments, yet they might perform at chance levels on a recognition task in which they have to identify the words they saw earlier. Thus, here implicit and unawareness are very much linked.

Unfortunately, those same connotations are present when the terms are applied to attitudes. The implication is that implicit attitudes are ones for which individuals lack awareness. How do we know that individuals lack such awareness? That is not an inference that should be made without additional evidence, certainly beyond the mere administration of an implicit measure. Once again, neither participants' unawareness of the actual purpose of the implicit measure nor the indirect nature of the attitude estimation procedure speak to the question of whether individuals are aware of their attitudes. Surely, discordance between scores on an implicit and an explicit measure, in and of itself, should not be considered evidence that an implicitly-measured construct is an unconscious construct. That amounts to an enormous leap, for the discordance can be explained in any number of ways, not the least of which is the notion highlighted earlier that people sometimes may be unwilling to report their attitudes.

We invite the reader to review the various quotations on the last few pages, beginning with the very definition of implicit bias offered in relation to the Starbucks training. The terms "unconscious attitude" and "implicit attitude" and "explicit attitude" occur repeatedly. Those are constructs. Nothing about the use of an implicit measure in any way justifies the inference that any discordance between an implicit and explicit measure stems from respondents having accessed two distinct constructs from memory.

Indeed, there is no evidence in the literature establishing that individuals are unaware of their attitudes (see Gawronski, 2019). No conclusive evidence exists for the assertion that individuals may be unable to report their attitudes. To the contrary, recent findings indicate that when asked in an appropriate fashion, individuals show considerable accuracy in predicting their IAT scores toward various social groups (Hahn, Judd, Hirsh, & Blair, 2014). Moreover, when participants are encouraged, indeed given license, to respond to explicit measures honestly

(Olson, Fazio, & Hermann, 2007) or quickly (Koole, Dijksterhuis, & van Knippenberg, 2001; Ranganath, Smith, & Nosek, 2008), the correspondence between implicit and explicit measures increases. Additional evidence that individuals are indeed aware of the attitudes that are assessed via an implicit measure will be discussed in a subsequent section.

The Science and Politics of Implicit Bias

There are serious problems with viewing implicit bias as an unconscious attitude. Indeed, in our view, both the science and the politics of implicit bias demand a different perspective. In terms of the science, the terminology and the labels that we use often conflate measures with constructs. That is not without consequence. Our science is replete with demonstrations regarding the importance of language. The very labels that are used guide information processing, judgments, and recall. Whether two cars are said to have "hit" or "smashed into" one another influences estimates of how fast they were traveling at the time of the collision and even recollections as to whether the collision produced any broken glass (Loftus & Palmer, 1974). Whether a given behavior is viewed as adventurous or reckless, as persistent or stubborn, is affected by exposure to those words in a preceding, unrelated context (Higgins, Rholes, & Jones, 1977). Whether appropriate color words exist in one's native language affects color discrimination (Winawer, Witthoft, Frank, Wu, Wade, & Boroditsky, 2007). Many additional examples could be cited.

Just like everyday perceivers, scientists are not immune to the effects of language. The terms that we use can affect our interpretations of data and influence the kinds of questions that we ask. If we simply presume in an unquestioning fashion that implicitly-measured attitudes are unconscious and that these unconscious attitudes are responsible for much of what is regarded as prejudice and discrimination, then we quickly revert to the level of the unconscious as

explanations for phenomena that we observe. We may not consider alternative perspectives and, hence, not be motivated to generate or pursue questions that do not follow from the presumption. Thus, our thinking can become restricted by the poorly analyzed presumption suggested by the terminology. It is critically important that we be careful about the language we use in pursuing the science of implicit bias. We should not imply that just because an attitude has been measured implicitly, it is revealing something about an unconscious attitude or hidden bias of which the individual is not aware.

Equating implicit bias with unconscious attitudes also creates a serious dilemma for the politics of implicit bias. The attention given to implicit bias over the last decade or so, in both the scientific and lay literatures, is to be applauded for raising consciousness regarding prejudice toward minorities and women. Obviously, issues of racial and gender discrimination need regular consideration and coverage in the media. Changing norms regarding appropriate behavior is an essential step in diminishing such unwanted bias. However, there is a downside to doing so when implicit bias is portrayed as arising from an unconscious attitude. When we call an attitude unconscious, we seemingly absolve individuals of responsibility for that attitude. After all, if they don't know they have the negative attitude, how can they be held responsible for their prejudicial behavior? Indeed, recent experiments have revealed that individuals engaged in discriminatory acts are held less accountable when those actions presumably stemmed from an unconscious attitude (Cameron, Payne, & Knobe, 2010; Daumeyer, Onyeador, Brown, & Richeson, 2019; Redford & Ratliff, 2016). Any such absolution is very inappropriate. People should be responsible for their attitudes and the behavioral manifestations of those attitudes. We must allow for that responsibility. Yet, as long as we think of implicit bias as an unconscious

attitude that individuals could not help but acquire by their immersion in a biased culture, we at least appear to view the individual himself or herself as not at fault.

An Alternative Perspective: Implicit Bias as an Effect of Attitudes

The alternative perspective that we hope to promote in this chapter is by no means novel. Instead, it has its roots in a lengthy tradition of social psychological theory and research. Indeed, it relates to what we regard as one of the most important principles, even the very essence, of the discipline. Perception is constructive in nature and, hence, very much in the eye of the beholder. That principle formed the core of the New Look movement that first reached prominence in the mid 1900's (Bruner, 1957; Bruner & Goodman, 1947), highlighting that perception depends on more than the physical stimuli that are presented. Whether "l3" is seen as the numeral "13" or as the uppercase letter "B" depends on the surrounding context. Sandwiched between 12 and 14, the sensory input is perceived as the numeral, but when surrounded by A and C, that same sensory input is perceived as the letter. The vast literature that has accumulated over decades now indicates that context is not the only influential factor. Perceptions also can be determined by expectations, hopes, wishes, aspirations, stereotypes, and attitudes (e.g., Balcetis & Dunning, 2006, 2010; Fazio, Roskos-Ewoldsen, & Powell, 1994; Fazio & Williams, 1986).

Avid sports fans are well-acquainted with this phenomenon. When two basketball players collide on the court, was the offensive player guilty of charging or the defensive player guilty of blocking? It is the members of the opposing team who are committing the more fouls, not the team for whom one is rooting, and yet the opponents' infractions are often overlooked by the referees (Hastorf & Cantril, 1954). And, these fan-driven perceptions grow all the more intense as the game reaches a climactic end or if the opposing team is a longstanding rival. Sometimes fans are actually aware that the event was a close call and may even acknowledge that their

assessments may be biased. But, at other times, fans actually perceive the event as a charge or a block. Phenomenologically, the fans experience themselves as perceiving, not judging.

That distinction between perception and judgment is central to the argument we wish to make. As Bruner (1957) argued, the distinction is very arbitrary. Sometimes it is clear to us that we are judging. As member of a jury, one is being presented with evidence by the prosecuting and defense attorneys, and clearly one's very mission is to judge that information in an impartial fashion. While watching the presidential debate, individuals are also probably aware that they are judging the candidate's performance. However, that line between perception and judgment is arbitrary, and sometimes we simply are not aware that we are judging. Was the proverbial "pink" traffic light actually yellow or red? Our experience, just like that of the sports fan, is one of perceiving, not judging.

Thus, what the perceiver brings to a given sensory event in terms of expectations, hopes, and wishes can be very important in determining what is seen. This fundamental principle of social psychology offers a valuable perspective on implicit bias. It is the constructive nature of perception that allows implicit bias to creep into our construals, regardless of whether we refer to them as perceptions or judgments. Moreover, to the extent that our experience is one of perceiving, we may not be aware that the perception has been influenced by our attitudes and stereotypes. It is with respect to the potential for unawareness of this biasing influence that unconscious processes play a role in implicit bias, not unawareness of the attitude itself. Instead of presuming the existence of an "unconscious attitude," such a perspective focuses us on the effect of attitudes in a given appraisal situation – an influence of which the individual is potentially unaware.

Automatically-activated Attitudes

The theoretical perspective that has driven our laboratory's research on attitudes over the last few decades views an attitude as an association in memory between the attitude object and an evaluation (see Fazio, 2007, for an overview). Both the term "object" and "evaluation" are intended to be broad. An attitude object can be a physical object, an issue, an activity, a situation, a person, or a collective of persons. An evaluation can be the outcome of a reasoned analysis of the strengths and weaknesses of the attitude object, a socially-communicated appraisal, an emotional reaction evoked by the object, and/or an inference from past behavioral experiences with the object. Importantly, however, the strength of that association can vary. Sometimes it is sufficiently strong that the attitude operates much like a well-learned semantic association. If presented with bacon, you can't help but think of eggs. Salt activates pepper; April 15th, taxes. In the same way, sometimes our attitudes involve those similarly strong associations. Presentation of the attitude object is sufficient to activate the associated evaluation from memory automatically, without any active, intentional reflection on the part of the perceiver (Fazio, Sanbonmatsu, Powell, & Kardes, 1986).

Such automatic attitude activation rests at the core of our MODE (Motivation and Opportunity as Determinants) model of attitude-behavior processes (Fazio, 1990; Fazio & Olson, 2014), which regards attitude activation as, in effect, the starting point to judgments and behavior. Generally speaking, once activated, attitudes color perceptions of the ongoing event, our constructed appraisals of what we are experiencing. They orient attention and influence categorizations and, hence, construals of objects (Roskos-Ewoldsen & Fazio, 1992; Young & Fazio, 2013). Judgments and behavior are downstream consequences, at least potentially, of these biasing effects of the automatically-activated attitude. Importantly, behavior here is a term

that is being used very globally. Included is not only overt behavior, but also verbal behavior. Explicit measures of attitudes are exactly that – verbal behavior. The survey respondent is expressing an opinion on a questionnaire or in response to an interview question. Such verbal expressions can be direct consequences of the automatically-activated attitude.

However, the MODE model also postulates that motivational factors may play a role. People may experience some additional motivational issue within a given situation. A motivation for accuracy may be evoked, because, for example, the decision itself is likely to be very consequential. Or, the motivation, in a context involving race, gender, or ethnic considerations, might very well be a motivation to control prejudiced reactions. Any such motivation may itself stem from sincerely valuing egalitarianism or from a sheer desire to avoid confrontation (Dunton & Fazio, 1997; Plant & Devine, 1998). If any such motivation is evoked, then the individual may counter the influence of that automatically-activated attitude. However, any such counteraction is effortful and, hence, requires that the situation offer the opportunity for engaging in effortful control. In other words, the situation must offer the time and the individual must have the resources to pursue the motivational goal that has been evoked. According to the model, both motivation and opportunity are necessary to correct for the influence of the automatically-activated attitude.

As noted in earlier reviews (e.g., Fazio & Olson, 2003, 2014; Olson & Fazio, 2009), the MODE model offers a lens by which to consider implicit bias.[1] At this point, a little personal history may be illuminating. The model itself certainly predates work on implicit bias or research on implicit measures; neither of those terms was in existence in 1990 at the time that the model

[1] The interested reader is referred to these earlier reviews for more comprehensive coverage of the relevant research than will be offered in the current chapter. The present aim is simply to highlight the value of a particular way of viewing implicit bias.

was first proposed. Indeed, it was a desire to test the model more fully that led the research team to develop the evaluative priming procedure, the very first unobtrusive measurement procedure eventually subsumed under the umbrella term "implicit measures." For some 10 to 15 years earlier, we had examined the attitude-behavior relation only indirectly. We assessed the accessibility of attitudes from memory, typically by latency of response to an attitudinal query. The faster the individual could respond to a direct query, the more likely it was that the attitude had been activated from memory automatically (Fazio et al., 1986). Or, alternatively, we manipulated attitude accessibility, by having people rehearse their attitudes and, hence, strengthen the object-evaluation associations. We then could compare self-reported attitudes that were highly accessible to ones that were less accessible. For attitudes that were highly accessible, (i.e., ones that were capable of automatic attitude activation), we regularly observed stronger relations between the attitudinal self-reports and the judgmental or behavioral outcome measure of interest than was true for less accessible attitudes (e.g., Fazio & Williams, 1986; Houston & Fazio, 1989). Moreover, those effects were especially evident in situations in which any motivation to counter the influence of the attitude was lacking or in which individuals were forced to make decisions under time pressure (e.g., Sanbonmatsu & Fazio, 1990; Schuette & Fazio, 1995).

What was unfortunate about this approach, however, is that there was a sense in which it had never really tested the theory, because the research did not measure automatically-activated attitudes. It was this theoretical challenge that actually prompted us to develop the evaluative priming procedure, because we wanted to index what happens in memory when an individual is presented with an attitude object. What is automatically activated? Could we get a snapshot of that? And, could we more directly test the MODE model by having such a snapshot?

The essence of the evaluative priming procedure is very simple. Participants' primary task concerns a set of evaluative adjectives. Their job is to simply tell us as quickly as possible whether the connotation of the adjective is positive or negative, i.e., whether it means good or bad. The adjectives themselves are unambiguously positive or negative, so participants make very few errors. We are interested in the latency with which the responses are made. Based on prior work that we had conducted using a priming paradigm to study automatic attitude activation, we knew that people would be faster to tell us that an adjective like "nasty" means bad if it was preceded by, for example, the word "cockroaches" as opposed to the word "chocolate" (Fazio et al., 1986). The former would automatically activate negativity and, hence, facilitate responding to the evaluatively-congruent target adjective, whereas the latter would activate positivity and potentially interfere with fast responding.

Our very first series of experiments attempting to measure automatically-activated attitudes used this paradigm to assess racial attitudes (Fazio, Jackson, Dunton, & Williams, 1995). After being told that the experiment concerned judging word meaning, the participants underwent the adjective connotation task. They then were introduced to a face-learning task in which they were exposed briefly to a large number of photos of individuals' faces with the instruction to study the faces so as to be able to pick them out later. Immediately thereafter, a face-detection task began in which participants were asked to indicate whether the face was one they had seen earlier. We then transitioned to the critical priming phase of the experiment in which the adjective connotation and face learning tasks were conjoined. Participants were told that if judging word meaning is indeed an automatic skill, then they should be able to perform the adjective connotation task just as well as they had at the beginning of the experiment even when given something else to do at the same time. That additional task was to learn faces. So, on

each trial, a face was presented, which participants needed to study for a later detection task. The photo was followed by the target adjective whose connotation the participants needed to indicate as quickly as possible by pressing either the "good" key or the "bad" key. Included were faces of Hispanics and Asians (intended to obscure our interests), but most critical for our purposes were sets of photos of White faces and photos of Black faces that had been followed by the very same positive and negative target adjectives.

As mentioned earlier, participants make very few errors during the connotation task. The response latencies are the data of interest. In particular, we are concerned with the extent to which the latency data show a pattern indicative of automatically-activated negativity in response to Black individuals. As shown in Table 1, a pattern indicative of negative racial attitudes involves relatively faster responding to negative adjectives when they are preceded by Black faces in comparison to White faces and/or relatively slower responding to positive adjectives when they are preceded by Black faces in comparison to White faces. The extent to which a participant exhibits this interactive pattern between the race of the photos and the valence of the target adjective serves as the estimate of individual's racial attitudes.

Race of Photo	Positive Adjectives	Negative Adjectives
White faces	***Fast***	Slow
Black faces	Slow	***Fast***

Table 1. A pattern of latencies indicative of prejudice

The findings from one of our very early studies using this evaluative priming procedure to assess automatically-activated racial attitudes will allow us to offer a number of observations relevant to the current perspective on implicit bias. In this particular study (Fazio et al., 1995,

Study 4), the scores derived from the priming procedure were used to predict verbal behavior – specifically, responses to the Modern Racism Scale, a commonly-employed questionnaire assessing racial prejudice (McConahay, Hardee, & Batts, 1981). The scale asks respondents to indicate the extent to which they endorse various statements like "Discrimination against Blacks is no longer a problem in the United States" and "Over the past few years, Blacks have gotten more economically than they deserve." We were interested in whether we could predict such endorsements from our estimates of participants' automatically-activated racial attitudes. According to the MODE model, we should be able to do so, but that will depend on the extent to which the individual experiences any motivation to control prejudiced reactions. We developed and employed a simple measure asking the college student participants about such motivation. To what extent are they concerned about acting prejudiced (e.g., "I get angry with myself when I have a thought or feeling that might be considered prejudiced")? To what extent are they willing to restrain themselves so as to avoid dispute with or about Black people (e.g., "If I were participating in a class discussion and a Black student expressed an opinion with which I disagreed, I would be hesitant to express my own viewpoint")? It is with this scale that we assessed the extent to which individuals were motivated to control prejudiced reactions.

Our interest was in how the two predictor variables – automatically-activated racial attitudes and motivation to control prejudice – might jointly relate to verbal expressions of prejudice on the Modern Racism Scale, for which higher numbers reflect greater negativity. The data revealed a significant interaction between the two predictor variables, which is presented in Figure 1. For people low in motivation to control prejudice reactions, the relation is just as we would expect. The more negativity exhibited in response to Black faces relative to White faces during the evaluative priming procedure, the more negativity expressed on the Modern Racism

Scale. As motivation to control prejudice reactions increased, that relationship attenuates and even reverses.

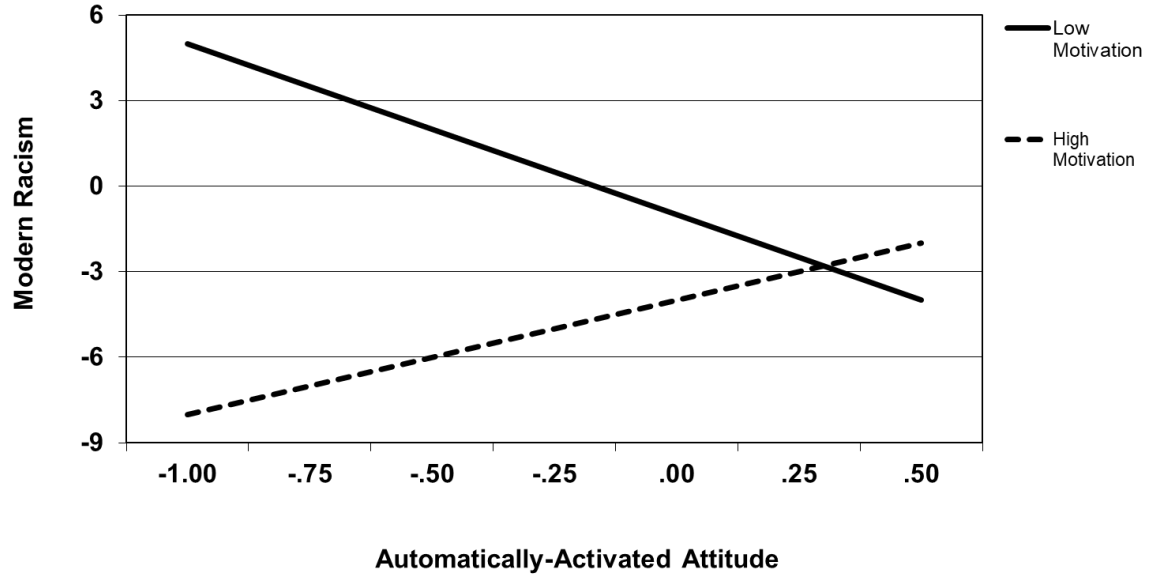

Figure 1. Scores on the Modern Racism Scale as a function of automatically-activated racial attitudes and Low versus High Motivation to Control Prejudice. Higher scores on the y-axis reflect greater prejudice; more positive scores on the x-axis represent more automatically-activated positivity toward Black faces relative to White faces. Adapted from Fazio et al. (1995).

The findings enable us to identify three different classes of people, roughly speaking (see Figure 2). One group, those depicted in the right side of the graph, might be referred to as "truly unprejudiced." They do not have negativity automatically activated in response to Black faces. At the top left, we have the "truly prejudiced." For such people, negativity is automatically activated, and they have absolutely no qualms about expressing that negativity. The bottom left represents a very interesting group. For want of a better term, we can refer to them as the

"motivated egalitarians." For such individuals, negativity is automatically activated, but they are bothered by the fact that it happens. They don't appreciate that they are experiencing a negative reaction, and they are motivated to control that prejudiced reaction. Moreover, that motivation seems to lead to an overcorrection. They actually present themselves as less prejudiced than people for whom no automatically-activated negativity occurs.

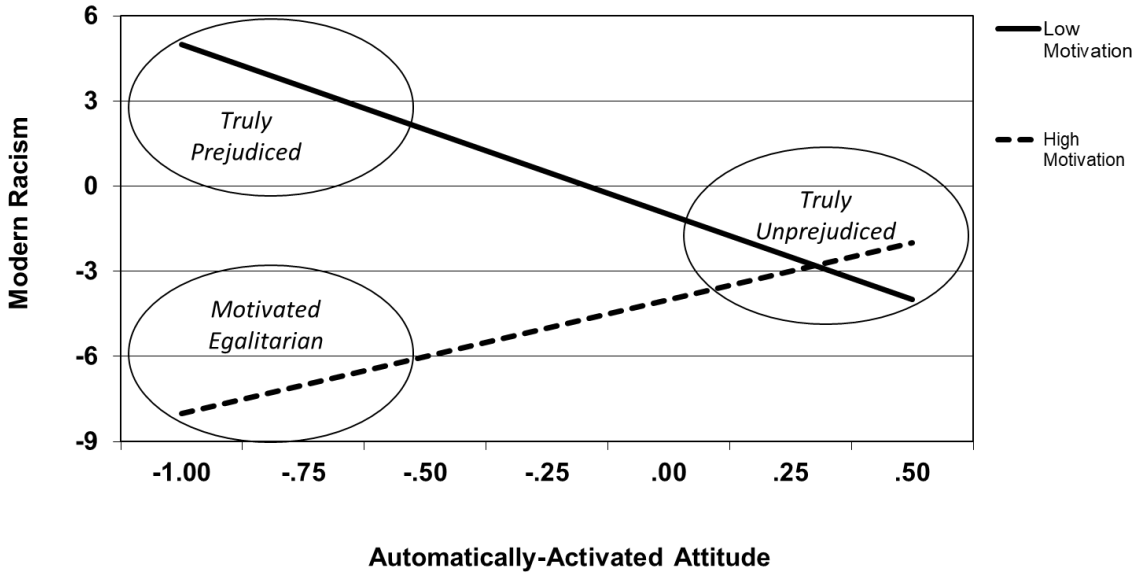

Figure 2. Three rough categories of people superimposed upon the regression lines depicted in Figure 1.

The laboratory observed similar patterns of data in many later studies (see Olson & Fazio, 2009; Fazio & Olson, 2014, for reviews). However, before briefly overviewing some additional research, let us offer a few observations based on the findings depicted in Figure 1. Let's first consider what the data reveal about individuals who are not generally motivated to control prejudiced reactions. What is noteworthy here is that these individuals displayed

considerable variability with respect to their automatically-activated racial attitudes, i.e., in the extent to which Black faces evoked negativity for them. A score of zero on the x-axis is indicative of no greater activation of positivity or negativity in response to Black faces compared to White faces. For some participants, negativity was automatically activated, for others less so, and for still others positivity was activated. The participants were White college students from a Midwestern state university. At least some of these students developed racial attitudes that were not consistent with the common cultural stereotype associating Black people with negative attributes. This speaks to the point offered earlier about holding individuals responsible for their racial attitudes. If some can overcome the influence of the culture in which they are immersed, why cannot others? The cultural transmission of negative racial attitudes, whether it be via parental socialization, friendship networks, or the media, is unquestionably a very influential factor, but the fact that some people display a very different pattern of automatic attitude activation indicates that it is by no means the only determining force. Some people, whether as a consequence of pleasant interracial interactions, egalitarian values, or whatever, do develop more positive racial attitudes. Given this variability, it seems evident that individuals themselves should be held accountable for prejudicial attitudes. Assigning blame solely to the culture is not justifiable.

A second observation concerns the moderating influence of motivational factors that was observed in the study. Analyses that focus only on simple correlations, ignoring the moderating influence of motivation and opportunity, will fail to capture the predictive utility of implicit measures of attitudes. The impact of implicitly-measured attitudes will be most evident when individuals lack the motivation and/or opportunity to counter their automatically-activated influence. Indeed, given that some people with more negative racial attitudes actually succeed in

correcting for their negativity when expressing their attitudes or engaging in some judgmental process, we can easily underestimate the influence of automatically-activated attitudes and the predictive value of implicit measures that capture this activation.

Finally, and possibly most importantly, consider the implications of the data revealed by the set of individuals referred to as motivated egalitarians. These participants seemed to *overcorrect* for the influence of their automatically-activated negativity. Their verbal behavior was characterized by even less prejudiced endorsements than was true of individuals whose performance during the priming procedure was indicative of more positive racial attitudes. Doesn't such overcorrection by individuals who are motivated to control prejudiced reactions indicate that these people are aware of their negative attitudes? What reason would they have to engage in corrective action if their attitudes were unconscious?

The Importance of Motivation

According to the MODE model then, experiencing some motivation to counter the influence of one's automatically-activated attitude and having the opportunity to do so is critical. If either of those conditions is lacking, then the subsequent judgments or behaviors should be a more direct downstream consequence of the automatically-activated attitude. In the study just described, the verbal behaviors that constituted the outcome measure involved expressing one's agreement or disagreement with statements from the Modern Racism Scale. The statements were obviously race-related. Indeed, they could not have been any more direct in that regard. Hence, the very nature of the situation should have evoked any motivation that individuals might have to control prejudiced reactions.

The same was true in other studies in which we observed moderating effects of motivation. For example, participants in one study were asked about their willingness to interact

with a variety of strangers, including a Black person, in various situations. Willingness to enter such situations was a function of their automatically-activated racial attitudes moderated by the extent to which they were concerned about acting in a prejudiced manner (Towles-Schwen & Fazio, 2003). Given little motivation, those with more negative attitudes anticipated less comfort in such interactions. Once again, as motivation increased, this relation was attenuated and eventually reversed, just as in Figure 1. Conceptually parallel findings were observed when participants were asked to list and then rate the feelings the came to mind when they thought of the "typical Black male undergraduate" (Dunton & Fazio, 1997). Yet another study focused on participants' first impressions upon being presented with photos of individuals varying in race, gender, and occupational cues (Olson & Fazio, 2004a). Trait ascriptions (e.g., likeability, intelligence) regarding Black (versus White) men and women depended jointly upon racial attitudes as measured by the priming procedure and the motivation variable. For individuals with lower motivation, those with more positive attitudes judged the Black targets more positively, but this relation was attenuated and, once again, even reversed, with increased motivation. Responding to direct questions about Black people is obviously race-related and, hence, likely to evoke any egalitarian concerns and/or any desire to avoid dispute (see Olson & Fazio, 2009, for a detailed consideration of these two different forms of motivation and their roles in the studies summarized above).

However, other studies from our research program have revealed that such motivational concerns are not always evoked. Situations do not always make it obvious that the judgment being rendered is race-related. Indeed, we have conducted studies in which we have observed an influence of automatically-activated attitudes *unmoderated* by motivation to control prejudiced reactions.

As one such example, consider a study in which participants were asked to offer judgments of the quality of a set of candidates to the Peace Corps (Olson & Fazio, 2007). For each of the four candidates, participants were provided with a four-page dossier which included a resume, a college transcript, a personal statement in which the applicant discussed reasons for wishing to join the Peace Corps, and the summary report of an official who had interviewed the applicant. The first page of the dossier, i.e., the resume, included demographics, high school and college extracurricular activities, grade point averages, academic awards, volunteer activities, employment history, and, most importantly, a race-revealing color photograph of the applicant. Two of the applicants were White females, one of whom had excellent credentials, whereas the other's qualifications and record were obviously very weak. The two remaining applicants were the focus of the study. Both were males and both were presented as having rather mixed and ambiguous credentials. One was White ("Jason Heinrich") and the other was Black ("Jamal Wills"). The difference in how a given participant evaluated these two candidates served as the major outcome variable.

What is noteworthy about this situation is the wealth of information that was provided about each candidate. In this context, an assessment of Jamal hardly seems to be a matter of his race. The judgment is presumably based on all the details offered in his dossier. Yet, those details are open to interpretation; strengths or weaknesses can appear especially serious and the mixed nature of the record requires some integration of those characteristics. Decades of research in social psychology, well before the onset of any research conducted under the rubric of implicit bias, suggest that attitudes can guide the processing of such ambiguous information (e.g., Fazio & Williams, 1986; Hastorf & Cantril, 1954; Houston & Fazio, 1989; Lord, Ross, & Lepper, 1979). Indeed, that is what was found in this study. Automatically-activated racial attitudes, as

estimated by the evaluative priming procedure earlier in the study, predicted the extent to which the Black and the White candidates were evaluated differentially. Those with more negative racial attitudes evaluated the Black applicant more negatively relative to the White applicant. Importantly, these judgments were *not* moderated by motivation to control prejudiced reactions. The observed relation with racial attitudes was unaffected by the motivational variable. Thus, it appears that participants did not view the setting and the judgment of Jamal as race-related. Nevertheless, their racial attitudes must have been activated upon their seeing the resume, and that activation colored their assessments of his qualifications for the position. Without the realization that the candidate assessment process was at all race-related, even individuals who described themselves as highly motivated to control prejudiced reactions were affected by their automatically-activated attitudes. Those attitudes served as the starting point to an attitudinally-biased processing of Jamal's dossier.

The Importance of Opportunity

According to the MODE model, it is not only motivation that matters in terms of the possibility of countering the influence of automatically-activated attitudes, but also opportunity. One must have the time and the resources to engage in any motivated correction process. Essentially, the model views opportunity as a gateway that, if open, allows for effective effortful control. Once again, a lengthy literature, some stemming from research conducted decades ago, speaks to the importance of such opportunity factors as time pressure (e.g., Jamieson & Zanna, 1989; Kruglanski & Freund, 1983; Sanbonmatsu & Fazio, 1990). Individuals' self-reported attitudes prove all the more predictive of their subsequent judgments when the situation forces them to reach those decisions rapidly.

More recent research has demonstrated conceptually parallel effects when attitudes are measured implicitly. One such experiment highlights both the joint influence of motivation and opportunity postulated by the MODE model, and also a different operationalization of the opportunity factor, one that centers on fatigue. Candy consumption (M&M's eaten) was examined as a function of attitudes toward M&M's (as assessed via an IAT), individuals' expressed motivation to control their body weight, and fatigue (Hofmann, Rauch, & Gawronski, 2007). The M&M's were made available to participants under the guise of a product-testing study after they had watched a 7-minute clip from an emotionally evocative movie. Participants assigned to an effortful condition were told to suppress their emotions and remain completely neutral while watching, an endeavor which prior research has demonstrated to be very taxing (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Gross & Levenson, 1997). Attitudes were more predictive of actual consumption in this condition than in a control condition in which participants were simply asked to watch the movie just as they would in a theater. In the control condition, candy consumption related to the motivation variable and not attitudes. Thus, when both motivation and opportunity were both present (in the control condition), participants' commitment to watching their weight overrode their attitudes. However, when participants did not have the resources to control their eating behavior, because they were fatigued by having had to keep their emotions suppressed, their attitudes toward M&M's were much more forceful in determining how much they ate.

Hoffman and Friese (2008) examined the effect of alcohol intake in a similar candy consumption procedure. Alcohol may diminish either individuals' motivation to control their desires or the resources that they have to do so, or both. As such, alcohol consumption should exacerbate the impact of automatically-activated attitudes toward candy in the taste-testing

situation. Precisely that was observed. Candy consumption was largely a function of participants' motivation to restrain their eating in the control condition. However, among participants who had drunk an orange juice and vodka mix in an earlier portion of the experiment (as opposed to orange juice alone), the amount of M&M's eaten was predicted largely by an implicit measure of attitudes toward M&M's.

In a conceptually similar manner, alcohol consumption has been found to interfere with the extent to which individuals control the verbal expression of their racial attitudes (Loersch, Bartholow, Manning, Calanchini, & Sherman, 2015). Implicitly measured racial attitudes were unrelated to explicit reports of prejudice and discrimination in either a control condition in which participants knowingly drank only tonic water or a placebo condition in which the participants drank a mixture that retained the taste and smell of alcohol although it contained only a diluted vodka substitute. The null relations suggest that these participants effectively controlled any automatically-activated negativity while responding to the explicit measures. As we would expect on the basis of the MODE model, however, the implicit measure of racial attitudes was significantly more predictive of the explicit measures among participants who consumed alcohol.

In some of our own research, the laboratory has pursued yet another approach to limiting the opportunity to engage in motivated effortful control of one's behavior. We turned to interracial dormitory roommate relationships as a context in which a White student could not possibly control prejudicial reactions on a continual, day-to-day basis, no matter how motivated the individual might be to uphold egalitarian values and avoid dispute with or about Black people. Surely, there will be days when the White student is rushed, overloaded with task demands, fatigued, or simply in a sour mood, all of which will make it difficult to overcome the influence of any automatically-activated negativity. Moreover, this dormitory roommate context

allows us to take advantage of an effect of aggregation over time. At any given moment in time, the opportunity to monitor and control one's behavior might be present, but across the many episodes of interaction that characterize a roommate relationship, there will be numerous instances in which such control proves difficult. In effect, then, studying roommate relationships over time benefits from aggregation, in much the same way that aggregating across a large sample of individual participants offers a means of assessing prejudicial attitudes within a given community (Hehman, Flake,& Calanchini, 2017; Payne, Vuletich, & Lundberg, 2017).

A pair of studies recruited as participants White college freshmen who had been randomly paired by the university to share a dormitory room with a Black student (Towles-Schwen & Fazio, 2006). The first study compared such participants to a sample of White students who had been randomly paired to room with a member of their own race. Two major findings emerged, one concerning self-reports and the other unmistakably behavioral. The White students in interracial roommate relationships reported engaging in far fewer joint activities with their roommates and reported less satisfaction with the relationship than did those in the same-race settings. Most importantly, however, substantially more of the interracial roommate relationships actually dissolved before the end of the semester; one of the two dyad members moved out of the room. The second study focused primarily on this latter variable. Early in the semester, a sample of White participants in interracial dyadic roommate relationships completed both the evaluative priming measure assessing their automatically-activated racial attitudes and the Motivation to Control Prejudiced Reactions scale. At the end of the year, the university housing office provided data regarding the last date that the students officially shared a room with their roommates. Only 43% of the interracial dyads remained intact on the last day of the academic year. The number of days the roommates lived together, which ranged from a mere 24

to the full 252 days of the year, offered a continuous measure of the success of the relationship. Racial attitudes, as measured at the beginning of the year, significantly predicted the longevity of the relationships. The motivational variable did not; nor did it moderate the relation between attitudes and the duration of the relationship. Thus, whether White students reported being motivated to control prejudiced reactions proved irrelevant. Presumably, even if motivated to do so, the opportunity to control prejudiced reactions was sometimes restricted. Hence, automatically-activated racial attitudes influenced appraisals of the interaction events, at least at times, without even the motivated students having the resources to counter the influence of any automatically-activated negativity. Accumulated over time, such negative appraisals apparently led to dissolution of the interracial roommate relationship.

Although somewhat of a digression, interestingly similar findings have been observed in research concerning romantic relationships. Implicit measures of attitudes toward the romantic partner predict the quality of the relationship prospectively over and above explicit measures. Naturally, newlyweds report being ecstatic about their partners. However, one of the most robust findings in the literature on close relationships is that marital satisfaction declines over time (e.g., Glenn, 1998). In one such study of newlyweds, the researchers administered an evaluative priming measure of attitudes toward the partner and then surveyed the couples periodically over the next four years (McNulty, Olson, Meltzer, & Shaffer, 2013). The implicit measure correlated with the decline in marital satisfaction over time. Those couples for whom the implicit measure revealed less positivity toward (and, hence, some hesitations about) the newly-wedded marital partner experienced a greater decline. Although those initial hesitations apparently were dismissed as inconsequential or otherwise obscured, in light of the romantic bliss the couples initially enjoyed, they prevailed over time. Yet another study focused on relationship dissolution

(Lee, Rogge, & Reis, 2010). Similar to the roommate relationship findings, an implicit measure of attitudes toward the romantic partner predicted the likelihood that the couple would break up over the following 12 months. Moreover, this relation was observed above and beyond the predictions of traditional self-report measures of relationship satisfaction and conflict.

Taken together, the summarized research highlights the importance of the MODE model's postulate regarding opportunity. When an individual's capacity to control the influence of an automatically-activated attitude is restricted by time pressure or resource depletion, the biasing impact of the attitude on judgments or behavior becomes evident even among those who are typically motivated to behave differently. Additionally, when considered over lengthy periods of time and, hence, likely to include the accumulation of numerous instances in which the opportunity to control was limited, the power of automatically-activated attitudes becomes all the more evident.

Implications for Implicit Bias Training

This may be a useful point at which to consider briefly the implications of the MODE model for implicit bias training. The model suggests that there are two very different modes by which prejudicial behavior might be mitigated. Obviously, one is to change the automatically-activated attitude, but that will be difficult to accomplish, especially for the subgroup of people previously referred to as the "truly prejudiced." Focusing on the development of motivations to control prejudiced reactions may prove more effective for such people. Individuals comprising the "motivated egalitarian" subgroup are different; they value the relevant motivational goals. Hence, for them, the major issue may be to increase the likelihood that the motivation is evoked in any and all relevant situations. What may be most helpful is to lead them to understand that

race or gender attitudes may be influencing their appraisals even in situations in which they have an overwhelming amount of information with which to render judgments.

A very relevant and interesting parallel exists with respect to anxiety disorders and their treatment. After all, a phobia is essentially automatically-activated negativity in response to some feared object or situation. Changing such disorders surely is not easy, which is exactly why therapies have been developed. Exposure treatment has been found to be remarkably effective in producing short-term change, but not necessarily long-term change (e.g., Craske, Kircanski, Zelikowsky, Mystikowski, Chowdhury, & Baker, 2008). In collaboration with a group of clinical scientist colleagues, our laboratory has considered the issue of treatment effectiveness through the lens of the MODE model, essentially asking what may change as a consequence of exposure therapy (Vasey, Harbaugh, Buffington, Jones, & Fazio, 2012). On the one hand, the actual attitudinal representation of the feared situation may change. On the other, the original attitude may not have been modified, but the individual may have developed both the motivation and the skills to control the automatically-activated fear. The latter is certainly laudable, but it does mean that the negative attitude remains capable of activation and that, if fatigued or otherwise lacking in opportunity, the individual will have difficulty controlling the fear effectively. A failure experience is likely to erode whatever confidence the individual may have developed. Hence, relapse is far more likely in this latter case than when the attitude toward the object or situation has changed.

This conceptual analysis received empirical support in a clinical trial involving individuals with social anxiety disorder and, in particular, clinically diagnosed phobia regarding public speaking (Vasey et al., 2012). By the end of exposure treatment, participants had improved markedly at delivering brief speeches while facing a video camera, as indicated by

their subjective reports of less distress, diminished heart rate, and judges' ratings of the speech quality. At a follow-up assessment one month later, however, some participants showed evidence of relapse, displaying what is referred to as "return of fear" in the clinical literature. Interestingly, an implicit measure of attitudes toward public speaking that was administered immediately after the exposure treatment was predictive of such relapse.[2] Individuals for whom the implicitly-measured attitudes toward public speaking had grown more positive as a result of the exposure session were more comfortable delivering speeches one month later.[3] Those participants for whom attitudes remained relatively negative, despite their success during the treatment sessions themselves, were more likely to experience return of fear. Presumably, the latter had learned to control their fears, but this did not mean that they would not experience negativity when again facing a public speaking situation.

Parallel processes may occur with respect to implicit bias training. Participants' recognition of the possibility that their appraisals in any given situation may be influenced by their attitudes may increase. Their motivation to control such bias might be enhanced, and those motivational forces might be evoked more readily across a wider variety of situations (see Forscher, Mitamura, Dix, Cox, & Devine, 2017; Stone, Moskowitz, Zestcott, & Wolsiefer, in

[2] The implicit measure employed in this study was a personalized IAT, a variant of the traditional IAT that focuses the measure on personal attitudes by examining the association between labels referring to things "I Like" and things "I Don't Like" and, in this case, public speaking. Such personalization has been found to diminish the sensitivity of the measure to extrapersonal factors, unrelated to individuals' attitudes, that can influence participants' interpretations of the ambiguous labels ("Good/Bad" or "Pleasant/Unpleasant") traditionally employed in IAT measures and the ease with which they can use the dual meanings of a given response key simultaneously. Discussion of features of the IAT that can make it differentially sensitive to extrapersonal considerations, including momentarily salient contextual factors, is well beyond the scope of the present chapter. Extensive empirical comparisons of the personalized and traditional variants of the IAT are available in other papers (Han, Czellar, Olson, & Fazio, 2010; Han, Olson, & Fazio, 2006; Olson & Fazio, 2004b). The research findings indicate that changes in traditional IAT scores need not imply an actual change in attitudes, whereas changes in personalized IAT score are more likely to do so.

[3] Changes in an explicit measure of attitudes toward public speaking were not similarly predictive of relapse one month later.

press; for research concerning interventions focused on enhancing motivation, awareness, and effort). Just as with anxiety disorders, such improvements with respect to controlled processing are to be lauded. Yet, they are not sufficient in terms of yielding consistent reduction in prejudicial judgments and behavior. It may take considerable time, effort, and practice, well beyond what is achieved in any short-term implicit bias training session, for the attitude itself to change. Nevertheless, change in the attitudinal representation should be the ultimate goal, shifting those with negative attitudes (the "truly prejudiced" or the "motivated egalitarian") toward the "truly unprejudiced" status. Attitudes that are more positive foster more benign appraisals, even in situations involving limited opportunity to reflect upon one's judgments.

Conceptualizing Implicit Bias

Our goal in this chapter has been to frame "implicit bias" as grounded in decades of social psychological theory and research, and to offer the MODE model as a theoretical perspective by which to view and understand the meaning of the concept. Essentially, consideration of the model and the associated literature has allowed the current discussion to address three questions, each of which will now be reviewed briefly.

When and How Does Implicit Bias Arise?

Implicit bias is no different than any other form of attitudinally-driven bias. Hence, it has the potential to arise whenever an attitude is automatically activated from memory. Such automatic activation occurs without any conscious reflection or deliberation on the part of the individual. It is a consequence of a strong association in memory between the attitude object and one's evaluation of that object. Merely encountering the object may be sufficient to activate the associated evaluation from memory, without the individual's intent, even if he or she is attempting to engage in some other task. Any such activation need not reach the level of

awareness; that is, the individual may not be aware that the attitude has been evoked. However, the activation does increase the likelihood that the evaluation will influence subsequent information processing (Bruner, 1957; Fazio, 2007; Higgins, 1996). In particular, the activated attitude serves to disambiguate available information and, hence, affects the individual's current appraisals.

Can One Effectively Overcome Implicit Bias?

Yes, one can effectively counter the influence of an automatically-activated attitude, but doing so requires (a) that relevant motivational factors be evoked *and* (b) that the opportunity to engage in effortful control is available. The latter involves properties of both the situation (e.g., time pressure) and the individual's current state (e.g., fatigue). The process by which motivated individuals correct for the influence of a biasing factor is obviously complex. Relevant social psychological theory postulates that upon suspicion that they are falling prey to an undesired judgmental bias, individuals may attempt to correct for the bias (Wegener & Petty, 1995). They adjust their judgments on the basis of naive theories that they hold regarding the direction and magnitude of the unwanted influence. The accuracy of these naïve beliefs, especially those concerning the magnitude of bias that one is experiencing at any given moment in time, is questionable. As a result, individuals' judgments sometimes show evidence of their actually having overcorrected for the biasing influence. We reviewed findings illustrating such overcorrection on the part of individuals who experienced automatically-activated negativity but were motivated to control their prejudiced reactions – the subgroup earlier referred to as the "motivated egalitarian" individuals. Interested readers are referred to a paper by Olson and Fazio (2009) for additional examples of such overcorrection in the context of racial prejudice, as well as an analysis of various forms that it might assume.

Critically, correcting for unwanted bias requires that individuals suspect that bias might be occurring and experience some motivation to counter the bias. The motivation may involve a desire to be accurate, to be fair, to adhere to one's egalitarian values, or simply to manage the impression one is creating. Whatever goal it might entail, the countervailing motivation must be evoked. In the context of prejudice, that will typically involve recognizing that the judgmental situation is race or gender related. Individuals must recognize the pervasive influence their attitudes can have even in situations in which it is not transparently obvious that the judgment involves matters of race or gender.

What is Implicit Bias?

Finally, and most importantly, we can return to the original question regarding the very nature of implicit bias. Instead of offering vague references to an empirically contradicted "unconscious attitude" or "hidden attitude," scientists, educators, and interested laypersons need to recognize that it is far more appropriate to characterize implicit bias is as an *effect* of attitude. Attitudes can bias our perceptions and judgments, without our awareness that the effect is happening. As argued earlier, both the science and the politics of implicit bias benefit from this theoretical perspective. We should be wary of conflating measures and constructs, and certainly avoid reifying measures as constructs. The very terms "implicit attitude" and "explicit attitude" imply a theoretical distinction at the level of the constructs. Hence, our very terminology can lead us astray in terms of interpretations of data and the questions we ask.

Moreover, the politics of implicit bias benefit from the advocated conceptualization. As an effect of attitude, instead of an "unconscious" attitude, implicit bias is much more readily viewed as the personal responsibility of the individual. Individuals should not be able to excuse any prejudicial behavior as a hidden by-product of the culture in which they have been

socialized. Given that some individuals are able to "rise above" any such cultural influences and show little or no evidence of automatically-activated negativity in response to Blacks, the responsibility for negativity clearly lies within the individual. In addition, anyone can become more sensitive to the possibility that their attitudes are influencing their appraisals in any given situation, once again highlighting individuals' personal responsibility for that influence.

References

Balcetis, E., & Dunning, D. (2006). See what you want to see: The impact of motivational states on visual perception. *Journal of Personality and Social Psychology*, *91*, 612–625.

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*, 147–152.

Baumeister, R.F., Bratslavsky, E., Muraven, M., & Tice, D.M. (1998). Ego-Depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*, 1252-1265.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64,* 123-152.

Bruner, J. S., & Goodman, C. C. (1947). Value and need as organizing factors in perception. *Journal of Abnormal Social Psychology*, *42*, 33–44.

Cameron, C. D., Payne, B. K., & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research, 23*, 272–289.

Craske, M. G., Kircanski, K., Zelikowsky, M., Mystikowski, J., Chowdhury, N., & Baker, A. (2008). Optimizing inhibitory learning during exposure therapy. *Behaviour Research and Therapy*, *46*, 5-27.

Czarnecki (2018, July 06). Timeline of a crisis: Starbucks' racial bias training. *PRWeek*. Retrieved from https://www.prweek.com/article/1486260/timeline-crisis-starbucks-racial-bias-training.

Daumeyer, N. M., Onyeador, I. N., Brown, X., & Richeson, J. (2019). Consequences of attributing discrimination to implicit vs. explicit bias. *Journal of Experimental Social Psychology*, *84*, 133–146

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316-326.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75-109). San Diego: Academic Press.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition, 25,* 664-703.

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

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Fazio, R. H., & Olson, M. A. (2014). The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 155-171). New York: Guilford Press.

Fazio, R. H., Roskos-Ewoldsen, D. R., & Powell, M. C. (1994). Attitudes, perception, and attention. In P. M. Niedenthal & S. Kitayama (Eds.), *The heart's eye*: *Emotional influences in perception and attention* (pp. 197-216). New York: Academic Press.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229-238.

Fazio, R. H., & Williams, C. J. (1986). Attitude accessibility as a moderator of the attitude-perception and attitude-behavior relations: An investigation of the 1984 presidential election. *Journal of Personality and Social Psychology*, *51*, 505-514.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W. T. L., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, time course, and longevity. *Journal of Experimental Social Psychology*, *72*, 133–146.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*.

Glenn, N. D. (1998). The course of marital success and failure in five American 10-year marriage cohorts. *Journal of Marriage and the Family, 60*, 569-576.

Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology, 20*, 419-445.

Gross, J. J., & Levenson, R. W. (1997). Emotional suppression: physiology, self-report, and expressive behavior. *Journal of Abnormal and Social Psychology*, *64*, 970–986.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*, 1369-1392.

Han, H. A., Czellar, S., Olson, M. A., & Fazio, R. H. (2010). Malleability of attitudes or malleability of the IAT? *Journal of Experimental Social Psychology, 46,* 286-298.

Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally-created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology*, *42*, 259-272.

Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, *49*, 129–134.

Hehman, E., Flake, J. K., & Calanchini, J. (2017). Disproportionate use of lethal force in policing associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*, 393-401.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133-168). New York: Guilford Press.

Higgins, E. T, Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, *13*, 141-154.

Hofmann, W., & Friese, M. (2008). Impulses got the better of me: Alcohol moderates the influence of implicit attitudes toward food cues on eating behavior. *Journal of Abnormal Psychology, 117*, 420-427.

Hofmann, W., Rauch, W., & Gawronski, B. (2007). And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology*, *43*, 497-504.

Houston, D. A., & Fazio, R. H. (1989). Biased processing as a function of attitude accessibility: Making objective judgments subjectively. *Social Cognition, 7*, 51-66.

Jamieson, D. W., & Zanna, M. P. (1989). Need for structure in attitude formation and expression. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 383-406). Hillsdale, NJ: Erlbaum.

Koole, S. L., Dijksterhuis, A., & van Knippenberg, A. (2001). What's in a name: implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology*, *80*, 669-685.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, *19,* 448-468.

Lee, S., Rogge, R. D., & Reis, H. T. (2010). Assessing the seeds of relationship decay: Using implicit evaluations to detect the early stages of disillusionment. *Psychological Science, 21*, 857-864.

Loersch, C., Bartholow, B. D., Manning, M., Calanchini, J., & Sherman, J. W. (2015). Intoxicated prejudice: The impact of alcohol consumption on implicitly and explicitly measured racial attitudes. *Group Processes & Intergroup Relations*, *18*, 256-268.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585-589.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality & Social Psychology*, *37*, 2098-2109.

McConahay, J. B., Hardee, B. B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution, 25,* 563-579.

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

McNulty, J. K., Olson, M. A., Meltzer, A. L., & Shaffer, M. J. (2013). Though they may be unaware, newlyweds implicitly know whether their marriage will be satisfying. *Science, 342,* 1119-1120.

Olson, M. A., & Fazio, R. H. (2004a). Trait inferences as a function of automatically-activated racial attitudes and motivation to control prejudiced reactions. *Basic and Applied Social Psychology*, *26*, 1-11.

Olson, M. A., & Fazio, R. H. (2004b). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, *86*, 653-667.

Olson, M. A., & Fazio, R. H. (2007). Discordant evaluations of Blacks affect nonverbal behavior. *Personality & Social Psychology Bulletin*, *33*, 1214-1224.

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.

Olson, M. A., Fazio, R. H., & Hermann, A. D. (2007). Reporting tendencies underlie discrepancies between implicit and explicit measures of self-esteem. *Psychological Science*, *18*, 287-291.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*, 233-248.

Plant, E.A., & Devine, P.G. (1998) Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811-832.

Ranganath, K. A., Smith, C. T., & Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*, *44*, 386-396.

Redford, L., & Ratliff, K. A. (2016). Perceived moral responsibility for attitude-based discrimination. *British Journal of Social Psychology, 55*, 279–296.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology*, *39*, 475-543.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*, 1043-1056.

Roskos-Ewoldsen, D. R., & Fazio, R. H. (1992). On the orienting value of attitudes: Attitude accessibility as a determinant of an object's attraction of visual attention. *Journal of Personality and Social Psychology, 63*, 198-211.

Sanbonmatsu, D. M., & Fazio, R. H. (1990). The role of attitudes in memory-based decision making. *Journal of Personality and Social Psychology, 59*, 614-622.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 501-518.

Schuette, R. A., & Fazio, R. H. (1995). Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model. *Personality and Social Psychology Bulletin, 21*, 704-710.

Stone, J., Moskowitz, G. B., Zestcott, C. A., & Wolsiefer, K. J. (in press). Testing active learning workshops for reducing implicit stereotyping of Hispanics by majority and minority group medical students. *Stigma and Health*.

Tang, T. (2018, May 28). Experts: Starbucks training a first step in confronting bias. *Associated Press*. Retrieved from https://www.apnews.com/a7afd1dc61ae4481a1ed9f5e67de1259.

Towles-Schwen, T., & Fazio, R. H. (2003). Choosing social situations: The relation between automatically-activated racial attitudes and anticipated comfort interacting with African Americans. *Personality and Social Psychology Bulletin*, *29*, 170-182.

Towles-Schwen, T., & Fazio, R.H. (2006). Automatically activated racial attitudes as predictors of the success of interracial roommate relationships. *Journal of Experimental Social Psychology*, *42*, 698-705.

Vasey, M. W., Harbaugh, C. N., Buffington, A. G., Jones, C. R., & Fazio, R. H. (2012). Predicting return of fear following exposure therapy with an implicit measure of attitudes. *Behaviour Research and Therapy*, *50*, 767-774.

Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, *68*, 36-51.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, *107*, 101-126.

Winawer, J., Witthoft, N., Frank, M.C., Wu, L.M., Wade, A.R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 7780-7785.

Young, A. I., & Fazio, R. H. (2013). Attitude accessibility as a determinant of object construal and evaluation. *Journal of Experimental Social Psychology*, *49*, 404-418.