Taking Stock of Explicit and Implicit Prejudice

Tobias H. Stark, Jon A. Krosnick, and Amanda L. Scott

Introduction

During the past century, racial attitudes in America have been radically transformed. One hundred years ago, this was a country of explicit racism, where separation of the races and discrimination against African Americans in particular were normative, formalized in laws, in the widespread practices of businesses and in the treatment of individuals by individuals every day. The civil rights movement of the 1960s brought about a landmark shift, eliciting widespread condemnation of racism, and setting the stage for the country's embracing of multiculturalism and implementing policies in many arenas of life to level the playing field and compensate for past discrimination. These changes in public practices were accompanied by a gradual transformation of public opinion in the United States: surveys documented a steady growth of endorsement of racial equality and a decline in explicitly stated racial prejudice. More and more Americans endorsed principles of racial equality and expressed support for various policies preventing discrimination.

And amidst all this, the women's movement spotlighted discrimination against and disadvantaging of women as well. An array of policies has also been implemented to attenuate such bias, and expressed support for equal rights for men and women rose steadily across the decades. Indeed, such a shift in laws and in public opinion has occurred with regard to many disadvantaged social groups, including gays and lesbians, the elderly, and disabled people. At the same time, similar shifts have been observed in many other countries around the world.

Yet in recent decades, observable events continued to occur illustrating that racism, sexism, and other forms of discrimination had not been eliminated. In the United States, evidence continued to document discrimination against African Americans, Asian Americans, Latino Americans, LGBTQ Americans, Native Americans, and women in several domains of life – healthcare, housing, education, employment, the justice system, and more (e.g., National Public Radio, the Robert Wood Johnson Foundation, and the Harvard T. H. Chan School of Public Health, 2018).

This incongruence between the rising endorsement of egalitarianism in surveys and the persistence of discrimination led to the first revolution in survey measurement of racist attitudes. Led by David Sears and his colleagues, the notion of symbolic racism was introduced to the survey research world. According to Sears, Americans had come to recognize that racism is disdained, so people who continued to hold racist opinions became increasingly unwilling to express those opinions explicitly in surveys and in life. So Sears and colleagues developed alternative measures that allowed survey respondents to express anti-Black attitudes shrouded in disguises that allowed those respondents to feel that their prejudice was hidden. And measures of symbolic racism were thought to document

considerably more prejudice than was documented by traditional, explicit measures. Similar measures of sexist attitudes were developed and performed similarly.

In the face of continued evidence of discrimination based on race decades later, social psychologists proposed another innovation: the notion that racism might be non-conscious. Specifically, these researchers proposed that implicit bias might be the result of widespread socialization throughout American society and might create potent associations in the long-term memories of huge majorities of people. Drawing on ideas from cognitive psychology, these social psychologists developed a bouquet of techniques for measuring non-conscious attitudes. And the designers of these techniques noted a useful side benefit that these measures afforded: because they do not rely on research participants' explicit self-reports, the new measures can document racism even when the holder of an opinion is aware of it but unwilling to reveal it. And indeed, researchers reported evidence apparently documenting nearly unanimous implicit bias in the adult population of the United States.

In this context, understanding the prevalence of bias in segments of society has substantial public benefit. In police departments, for example, knowing who holds implicit bias can help to focus training to reduce the number of instances in which innocent people receive unfair treatment and increase law enforcement's ability to protect populations as intended. In national defense environments, knowing who holds implicit bias can help to identify members of the military whose performance may be compromised. In the private sector, accurately understanding people's feelings about racial and ethnic groups can help businesses identify and correct biases within their organizations. In all of these ways and more, having accurate data on how people perceive and judge others can help farms, offices, factories, and professionals in numerous other settings to serve others effectively.

For these reasons, the identification of implicit bias has led many companies and government agencies to spend considerable resources training workers to minimize its impact on the work they do: caring for medical patients, enforcing laws, and much more.[1] These efforts are based on the understanding that racism and other forms of bias are widely prevalent in contemporary society and powerfully shape people's behavior in important arenas. And the notion of implicit bias has entered mainstream awareness; people in many circles outside of academia seem to view it as a well-established concept. Inherent in the zeitgeist of this period was the notion that explicit racism, which is easy to measure, was not the problem anymore.

And then came 2020. Among the many momentous events of that year was the surge of racism on the front pages of newspapers and the lead stories of broadcast news, not to mention in blog posts, tweets, and everywhere else on the Internet. Horrific violence directed at African Americans by police and others, unabashed protests by White supremacist groups, videos of White people accosting Black people in stores and parking lots, and more, have all painted portraits of explicit racism, leaving powerful, lasting impressions on the nation. The rise of the Black Lives Matter movement, nationwide counter-protests highlighting public outrage, and non-stop discussion of these issues by commentators, journalists, and others in the media have documented passionate condemnation of racist behavior. And a reaction to much of that, the claim that "all lives matter," is taken to be yet another manifestation of explicit racism.

[1] www.nytimes.com/2019/11/20/style/diversity-consultants.html

Thus, the discussion of race in America has changed again. We have entered what might be a new phase in the study of racism in America – a phase that puts implicit racism side by side with explicit racism, rather than discarding explicit racism as no longer worthy of study. We hope that this book is a helpful step into this new phase of research. To that end, this book takes stock of the existing literatures on racial bias and other forms of bias to gauge what we know and what we do not yet know about these issues. The chapters are written by many of the world's leading experts on racism and prejudice and review the literature, discuss the strengths and weaknesses of existing evidence, and identify fruitful directions for future work. We hope to help scientists and the general public to have a clear and complete understanding of the state of scientific evidence on the nature of implicit and explicit bias, their strengths, and their limitations.

Below, we offer a more detailed review of the history of the measurement of racism and set the stage for the rest of the book.

Traditional Explicit Prejudice Measures

In the early twentieth century, as quantitative social science was being born, American scholars developed explicit prejudice measures that gauged people's disliking of Black people (affect) or the stereotypes (cognition) they held about Black people (Allport, 1954; Bogardus, 1933; Hewstone et al., 2002). This was the time when Jim Crow racism in the United States ensured that many White people's contempt for and fear of African Americans was maintained through publicly endorsed associations of the Black community with negative traits such as low intelligence or laziness. As these perceptions were widely shared and even institutionalized, particularly in the United States South, many White Americans readily reported that they held such negative perceptions of African Americans. Explicit measures of prejudice thus asked people how they felt about Black people and how they would characterize members of the African American community (Axt, 2018; Ditonto et al., 2013; Dovidio et al., 1996; Mackie & Smith, 1998). Such questions are still used today to measure racial prejudice in social science research, as well as to measure prejudice toward gay people and lesbians, transgender people, disabled people, and other social groups (Burke et al., 2017; Crowson et al., 2013; Jackson et al., 2014; Norton & Herek, 2013).

Explicit affect measures focus on people's feelings toward members of a social group, for instance, how they feel toward African Americans and how comfortable they feel or would feel being around Black people. A variety of survey measures have been developed to tap such affect. These include feeling thermometers that ask how warm or cold a person feels toward members of a certain group (on a scale from 0 to 100), questions that ask how much people like or dislike Black people, and social distance scales on which people indicate how socially proximate they would be comfortable being with Black people (e.g., living near a Black neighbor, having a Black family member; Bogardus, 1933; Dovidio et al., 1996; Mackie & Smith, 1998). Such traditional explicit measures of affect are still widely used. For instance, the long-running American National Election Studies surveys continue to include feeling thermometers and questions about people's feelings toward various racial groups (www.electionstudies.org).

Explicit cognitive measures of prejudice assess the positive and negative attributes that people believe members of a social group possess, including physical characteristics, behavioral tendencies, values, personality traits, and preferences (Mackie & Smith, 1998). These can be measured via checklists, rating scale

questions asking for the degree to which certain traits describe a group, and semantic differential scales on which people place members of a group between two poles representing opposing attributes. These measures were originally developed to tap stereotypical beliefs about African Americans, such as that Black people are unclean, or that there are inborn differences between White and Black people that make African Americans less intelligent and less willing to work hard. Some of these beliefs, such as those about cleanliness, were easily debunked by increasing intergroup contact (Allport, 1954). Others were more persistent. For example, Huddy and Feldman (2009) asked in a 2006 national survey why people thought African American students get lower scores on standardized tests than White students, and 24 percent of White Americans said that this is to "some" extent or to "a great deal" due to racial difference in intelligence, and 20 percent said it is to "some" extent or to "a great deal" due to fundamental genetic differences between races.

Some implementations of traditional explicit measures do not assess affect toward or stereotypes of a specific group and instead compare feelings toward or perceptions of multiple groups (e.g., Goldman, 2012; Levin et al., 2003). This type of differential measurement approach builds on one common definition of prejudice as the tendency to evaluate one's own group *more* favorably than some other group (Brown & Zagefka, 2005; Hewstone et al., 2002). Comparing people's perceptions of their group to their perceptions of another group prevents mistaking generalized negativity toward all groups for prejudice toward a particular group.

Even though such differential prejudice measures are appealing, they have downsides. First, a differential measure may mischaracterize in-group favoritism as out-group bias (Brewer, 1979, 1999). Second, some people may consider outgroup members as "less good" but may not harbor hostility against them (Allport, 1954). Third, asking for people's perception of an outgroup and their ingroup requires twice as many questions, which increases study time and costs. The literature has not reached a conclusion about whether differential explicit measures are worth the cost.

Reduction of Prejudice?

Although many scholars have observed a general trend of traditional explicit measures of racial prejudice toward African Americans declining up through the turn of the century (Bobo, 2001; Krysan, 2011; Schuman et al., 1997), a more recent account paints a more complex picture: some explicit measures have shown a continuing decline, while others have shown stagnation or even slight increases (Moberg et al., 2019). This pattern of evidence is certainly consistent with the conclusion that some forms of explicit prejudice have declined, even if not all have. And numerous events that have occurred in the United States since 2016 have highlighted continued and vigorous prejudice.

Skeptics have raised the possibility that rather than attitudes and beliefs truly changing, what may have changed are societal norms for what one can comfortably acknowledge. As Americans have increasingly endorsed egalitarian values, many may now be unwilling to openly admit harboring inegalitarian attitudes. As a consequence, people who have negative feelings toward African Americans or who ascribe negative traits to African Americans may no longer report them honestly (Judd et al., 1995; Tarman & Sears, 2005).

This possibility is consistent with a huge literature in psychology on self-presentation. In general, people strive to present themselves favorably when interacting with others, even at the expense of honesty (Goffman, 1959). That

is why some people sometimes avoid reporting potentially embarrassing attitudes and behaviors, a tendency referred to as impression management social desirability bias (Paulhus, 1984, 1986, 2002; Paulhus & Reid, 1991). This social desirability bias is thought to be particularly likely in a survey interview where respondents have to openly admit potentially embarrassing attitudes to an interviewer who might pass judgment based on a reported response (Paulhus, 2002). In line with these concerns, some research suggests that reports of racial prejudice are influenced by the race of the interviewer, such that respondents report more favorable evaluations of Black people when speaking with Black interviewers than when speaking with White interviewers (e.g., Anderson et al., 1988a, 1988b; Finkel et al., 1991; Hatchett & Schuman, 1975). The reactivity of explicit racial prejudice questions in survey interviews has led scholars to ask "whether or not it is worth all the trouble to administer such surveys in the first place" (Corstange, 2009, p. 46).

Measures of "New" Racism

Sears and colleagues' proposed solution to this alleged problem is symbolic racism (Kinder & Sears, 1981; Sears, 1988; Sears & McConahay, 1973). This notion has been elaborated and transformed in the literature into related notions called modern racism (McConahay, 1986) and racial resentment (Kinder & Sanders, 1996). According to these theories that are generally described as positing "new" racism, racists nowadays no longer believe in the inferiority of African Americans. Instead, these "new" racists believe that there is no discrimination in America anymore and that any racial disparities are due to Black people's unwillingness to endorse fundamental American values, such as individualism and egalitarianism (Henry & Sears, 2002; Kinder & Sanders, 1996; Sears & Henry, 2005). Theories of "new" racism thus argue that racism has not necessarily declined but instead has changed its face. Accordingly, the decline detected with traditional explicit measures of prejudice (Krysan, 2011; Schuman et al., 1997) may have occurred because those measures no longer accurately tap people's racial attitudes (Sears et al., 1997).

Theories of "new" racism argue that the opinions assessed by traditional indicators of prejudice – negative affect and negative stereotypes about Black people's violation of traditional American values – are indeed involved in contemporary racists' convictions and in fact instigate those convictions. Anti-Black affect and negative stereotypes are said to be learned during socialization in childhood and adolescence (Kinder, 1986a, 1986b; McConahay, 1986) and "blend" together to form "new" racism (Kinder & Sears, 1981; Sears & Henry, 2003; Sears & McConahay, 1973). This blend has been measured with questions that tap four dimensions: the belief that Black people's disadvantages stem from their unwillingness to work hard, the belief that Black people demand too much, the belief that Black people no longer face racial discrimination, and the belief that Black people have already gotten more than they deserve (Henry & Sears, 2002).

Measures of "new" racism have been popular across the social and behavioral sciences, and these measures have proven to be effective predictors of discriminatory opinions about government policy and discriminatory behaviors (Sears et al., 1997; Sears & Henry, 2005). For instance, racial resentment was a stronger predictor of voting for Donald Trump in 2016 than other traditional predictors of vote choice (Enders & Scott, 2019; McElwee & McDaniel, 2017). Racial resentment was also a strong predictor of identification with the Republican party among White Americans

living in Southern states (Knuckey, 2017), and racial resentment predicted White people's unwillingness to vote for a Democratic candidate who associated with African Americans (Stephens-Dougan, 2016). Moreover, racial resentment predicted support of police violence (Carter & Corra, 2016) and opposition to gun control laws (Filindra & Kaplan, 2016).

Despite their wide use, measures of "new" racism have been the subject of criticism. One concern is about whether these questions tap only racial attitudes or are problematically confounded with attitudes toward a large and interventionist government or with fiscal or social conservatism (Rabinowitz et al., 2009; Simmons & Bobo, 2018; Sniderman et al., 2000; Wallsten et al., 2017; Zigerell, 2015). Another debate centers around the question of whether the construct tapped by these measures is sufficiently distinct from attitudes toward race-related policies, which the former are intended to predict (Carmines et al., 2011; Schuman, 2000; Sears & Henry, 2003, 2005). As these debates remain unresolved today, the continued widespread application of measures of "new" racism and their continued predictive success raises concerns among some scholars of prejudice.

Furthermore, the claim that "new" racism items measure prejudice without social desirability response bias has been disputed. Questions to measure "new" racism were initially designed to disguise expressions of negative attitudes toward a social group by providing seemingly race-neutral response options that nonetheless justify continued disadvantaging of disadvantaged groups (McConahay, 1986; Tarman & Sears, 2005). However, researchers have since argued and empirically shown that the intent of these measures is transparent to most research participants and that responses are easily manipulated if participants wish to engage in impression management (Brauer et al., 2000; Fazio et al., 1995; Holmes, 2009; Krysan, 1998). Thus, these researchers claim, measures of "new" racism are no solution for potential bias caused by socially desirable response behavior.

Moreover, measures of "new" racism, just like traditional explicit measures, may suffer from a different form of social desirability response bias: unconscious bias. Social norms against expressing negative attitudes toward Black Americans may have become sufficiently strong that people may not even be aware of their own bias (Devine, 1989; Greenwald & Banaji, 1995). That is, some well-meaning people who harbor prejudicial attitudes may not want to admit those attitudes to themselves, so their unconscious minds hide such attitudes from their conscious minds. These people might not be aware of their bias against members of certain groups, leading to what might be called self-deception social desirability response bias causing self-reports of attitudes to be inaccurate. In the face of these concerns, a new set of tools seemed to be needed to assess these unconscious attitudes.

Implicit Measures of Prejudice

To address this need, social psychologists proposed the notion of *implicit bias*, the idea that the unconscious mind might hold and use negative evaluations of social groups that are not and cannot be reported via explicit measures, no matter how disguised (Devine, 1989; Greenwald & Banaji, 1995, 2013). Measures developed to tap implicit bias include the Implicit Association Test (Greenwald et al., 1998), evaluative priming (Fazio et al., 1995), lexical decision tasks (Wittenbrink et al., 1997), the Affect Misattribution Procedure (Payne et al., 2005), and others (for a review, see Gawronski & De Houwer, 2014). These measures tap the

unconscious mind by measuring the speed of making judgments or the results of making seemingly neutral judgments. These measures seek to tap directly into long-term memory to assess bias, unfiltered by conscious awareness or self-presentational motives.

Measurements of implicit bias seek to detect associative processes that occur when people are confronted with members of other groups. Associative processes are simple, spontaneous reactions that occur in response to a stimulus based on the match between the stimulus and the individual's internal pre-existing network of stimulus-attribute associations. These reactions require little cognitive capacity, intention, or even awareness. When people are asked directly about their feelings toward African Americans, they might engage in a conscious inferential process to consider all of the propositions or statements that come to mind and that are considered relevant for the judgment at hand. This may include the societal norm favoring racial equality, which might inhibit some people from expressing their true attitudes toward African Americans. In contrast, implicit attitude measures are thought to capture spontaneous associations without involving respondents' deliberation or introspection. The focus of these measures is thus not on the content of the response but on the underlying associations that can be revealed by response speed or accuracy or other distinct indicators.

For example, an implementation of the Implicit Association Test might ask participants to make one of two judgments about each of a series of stimulus words and images. The words might vary in their valence (e.g., positive/negative), and the images might vary in their group-based content (e.g., showing faces of Black people and White people). If a participant's mind links a social group with a negative or positive evaluation, that linkage is presumed to affect the speed with which the person makes judgments. Thus, reaction time is thought to reveal prejudices of which a person might be unaware or that a person might be unwilling to express explicitly.

Evaluative priming is another implicit attitude measurement technique that relies on the same principles. Specifically, reaction time is used to measure the strength of a person's association of positive or negative affect with members of a target group (e.g., Fazio et al., 1995). For example, during an evaluative priming task, participants might be shown representations of target groups (e.g., photographs of faces of Black people and White people), each followed by the display of a word. Participants quickly categorize the word as positive or negative by pushing a button. Implicit attitudes are indicated if people are faster at recognizing negative words after seeing members of one target group (e.g., faster after seeing faces of Black people) and quicker at recognizing positive adjectives after seeing members of the other group. That is, if a photograph (i.e., a prime) activates positive or negative affect, that affect facilitates or interferes with performance of the participant's categorization task.

Lexical decision tasks likewise combine priming and reaction time measurement to assess the extent to which people have associations between groups of people and positive or negative stereotypical attributes (Wittenbrink et al., 1997). In this procedure, the priming is semantic and not evaluative, as participants are shown a prime that represents the group of interest and a non-valenced word (e.g., the words BLACK or WHITE). This prime is displayed so quickly that it remains outside of the participant's conscious awareness. After each prime, a measurement is made of the time it takes participants to correctly identify a string of letters as a word or non-word. Negative affect is indicated if a person is quicker at correctly identifying negative stereotypical traits of one group as words and is also quicker at

correctly identifying positive stereotypical traits of another group as words.

The Affect Misattribution Procedure does not rely on reaction time and instead measures negative affective responses to target groups by assessing the misattribution people make about the origin of their affect (Payne et al., 2005). A typical approach is to show participants Chinese ideographs with which they are unfamiliar and precede each ideograph with a very fast flash of either an African-American or a White face, which the participants are told to ignore. People's affective reaction to a face is thought to spill over onto their assessments of the following ideograph, which respondents rate as either pleasant or unpleasant. More pleasant ratings of ideographs that appear after White faces than after African American faces indicate automatic negative affect toward African Americans.

Some of these measures of implicit attitudes (e.g., lexical decision tasks) are so unobtrusive that respondents are not aware of what is being measured. Other measures, such as the Implicit Association Test, rely on spontaneous reactions that are difficult to manipulate (but see Cvencek et al., 2010; Fiedler & Bluemke, 2005) and are thus thought to detect even attitudes that a person might be unwilling to report explicitly or might be unaware of (Dasgupta & Stout, 2012). Hence, if these measures succeed in accurately measuring people's true attitudes toward social groups, they have a clear edge over explicit measures of prejudice. This notion has led some scholars to dismiss traditional explicit measures of prejudice and to favor measures of implicit bias that can assess mental content without social desirability response bias (Kim, 2003; Olson & Fazio, 2003, 2004; Orey et al., 2013). For instance, Ito et al. (2015) wrote, "The prospect that this mental content could be assessed without reliance on self-report was met with great enthusiasm, particularly as it was becoming clear at the time that self-reported intergroup attitudes were artificially positive, masking an underlying, stubborn basis of prejudice" (p. 187). Banks and Hicks (2016) wrote, "On explicit measures of racism, respondents can consciously avoid appearing racist by intentionally self-monitoring their responses – providing the most politically correct answer. Implicit measures reduce the likelihood that respondents can hide undesirable responses" (p. 642).

If this conviction is true, we should observe much more prejudice with implicit prejudice measures than with traditional explicit measures.

Inference about a Population

To answer questions about the prevalence of prejudice in a population, we need data that allow accurate inferences about that population. And scientists studying racism fall into two groups in terms of the types of data they have collected and analyzed. For decades, survey researchers (especially from sociology and political science, but from psychology as well) made use of data collected from truly random samples of the American adult population (e.g., Krysan, 2011; Schuman et al., 1997). This methodology provides a strong foundation for reaching conclusions about the population of interest, based both on sampling theory and on empirical evidence that such samples have provided and still provide highly accurate measurements (e.g., MacInnis et al., 2018), even in the face of dropping response rates (e.g., Holbrook et al., 2007).

Unfortunately, the cost of collecting data from random samples of American adults has increased in recent decades, and those costs have always vastly exceeded what most psychologists believed they could afford to pay to collect data with new measures of

prejudice. As a result, psychologists have studied prejudice almost exclusively using data from non-probability, convenience samples of participants.

In fact, this approach is very much in keeping with the traditional view of participant sampling in the field of psychology for over a century. As a discipline, psychology has not been especially concerned with collecting data from representative samples of well-defined populations. Instead, the presumption usually made has been that psychologists are studying fundamental processes that appear relatively uniformly across people, so it is possible to "generalize until proven otherwise." Yet rarely if ever have findings generated with convenience samples been checked against those obtained by representative samples. Instead, psychology has evolved confidence in findings by repeatedly observing them in similar convenience samples. Thus, it should come as no surprise that the recent literature on implicit bias in psychology has relied almost exclusively on convenience samples, thereby limiting the ability to justify strong conclusions about the United States population as a whole (or any other population). These convenience samples include students on college campuses enrolled in psychology courses, workers on Amazon's Mechanical Turk (MTurk), and members of opt-in online panels who volunteered to complete surveys for money.

This use of non-probability samples is problematic in many regards. First, there are notable differences between students and non-students in terms of their answers to measures of prejudice (Henry, 2008). Thus, student samples do not provide a solid basis for inference about prejudice in the general United States adult population. Samples from MTurk resemble the general population more than student samples do (Buhrmester et al., 2011). However, recent studies have shown that the data quality of MTurk samples is suboptimal (Cheung et al., 2017; Chmielewski & Kucker, 2020), perhaps because an increasing number of people on MTurk misrepresent who they actually are (MacInnis et al., 2020) or due to an increasing presence of bots (automatic computer programs that behave like respondents) and "farmers" (people who bypass MTurk's location restrictions by using server farms) (Stokel-Walker, 2018).

Likewise, results based on non-probability samples of people who actively volunteered to complete a survey are problematically inaccurate for describing populations (Chang & Krosnick, 2009). Experimental studies have shown that different opt-in panels produce very different results (Smith et al., 2016) and that these results often diverge considerably from benchmarks of truth, while representative samples do a much better job at representing such benchmarks (Macinnis et al., 2018; Malhotra & Krosnick, 2007; Yeager et al., 2011). Accordingly, survey methodologists advise against the use of non-probability samples for inference about a population (Cornesse et al., 2020).

Non-probability samples have been the focus of the data examined by the most prominent team studying implicit bias: Project Implicit (www.implicit.harvard.edu). Their website has allowed interested participants to complete an Implicit Association Test on their computer, and millions of these tests have been completed (Nosek, Smyth et al., 2007). Dozens of research articles have been written with data collected from people who made their way to the website to take the test. Many of the resulting publications have made claims about the prevalence of prejudice and implicit bias and their correlates (e.g. Nosek, Smyth et al., 2007). Although many of these publications contain a disclaimer about the non-probability nature of the sample in their limitations section, analyses have nonetheless often been conducted, and conclusions have nonetheless

often been drawn as if the sample allows confident generalization to the population. But the accumulated evidence of the inaccuracy of non-probability samples applies to these data.

Interestingly, collecting data from random samples has become more do-able for psychologists in recent years. Organizations in multiple countries have set up representative national online panels by first drawing probability samples from a population and then inviting selected people to complete online surveys regularly (Silber et al., 2018). Some of these panels allow researchers to collect data for free (e.g., the GESIS panel in Germany). In the United States, the National Science Foundation has funded TESS (Time-sharing Experiments in the Social Sciences) for decades, which pays for the collection of probability sample data from representative samples via the Internet and telephone. And in 2008, the American National Election Studies included implicit measures of racial prejudice in its national surveys of representative samples, and the data are available to all interested scholars at no cost (www.electionstudies.org). Therefore, it is not only desirable but also easy to rely on representative national samples if researchers wish to draw accurate inferences about characteristics and relationships between variables within the American adult population.

Prevalence of Prejudice

To address the claims and presumptions outlined above, we next offer assessments of the prevalence of prejudice as gauged by traditional explicit measures, measures of "new" racism, and implicit measures, using data from four studies of national random samples of White non-Hispanic Americans. Samples 1 through 3 (*N*s: 1,215; 1,037; 791) were drawn from the KnowledgePanel®, an online survey panel recruited via probability sampling. Data for Sample 4 came from the 2008–2009 American National Election Study (ANES) Panel Study ($N = 1{,}441$) that was also conducted online with a random sample of American adults (DeBell et al., 2010). The Appendix of this chapter provides a detailed sampling and sample description as well as the question wordings and coding scheme for each variable.

Each survey included two traditional explicit measures of anti-Black affect: people's *disliking* of Black people and a measure of *differential disliking* that compared participants liking of Black people to their liking of White people. All surveys also included traditional explicit measures of *stereotypes* of Black people (e.g., friendly, determined to succeed, violent). In Surveys 2, 3, and 4, the same questions were asked about White people, enabling the calculation of *differential stereotypes*. The first three surveys also contained items measuring *symbolic racism* (Henry & Sears, 2002) and *racial resentment* (Kinder & Sanders, 1996), two manifestations of "new" racism. Survey 4 included a four-item short version of the racial resentment scale. Last, in all four surveys, respondents completed the Affect Misattribution Procedure (AMP) to assess implicit prejudice toward African Americans (Payne et al., 2005). In addition, Survey 4 contained the brief version of the Black-White Implicit Association Test (IAT) (Sriram & Greenwald, 2009). All measures were coded to range from 0 to 1, with 1 indicating more prejudice, to ease interpretation (see the Appendix at the end of the chapter for details).

Across all surveys, implicit measures of prejudice and measures of "new" racism identified the largest group of participants with anti-Black attitudes (see Table I.1). For this analysis, the samples were split into those with "positive" attitudes toward African Americans (values below 0.5 on each measure), those with "neutral" attitudes (0.5), and those with "negative" attitudes (values above 0.5). According to the

Table I.1 *Distribution of prejudice measures in four representative national samples of White non-Hispanic Americans (average scores)*

	Attitudes toward African-Americans				
Measure	Positive (%)	Neutral (%)	Negative (%)	Total (%)	*Valid N*
Disliking of Black people					
Sample 1	48.09	43.22	8.68	100	1,207
Sample 2	48.20	45.71	6.08	100	797
Sample 3	42.60	51.27	6.14	100	767
Sample 4	25.68	69.77	4.55	100	1,425
Differential disliking of Black people					
Sample 1	1.49	74.42	24.10	100	1,206
Sample 2	0.97	80.37	18.66	100	796
Sample 3	1.25	79.23	19.52	100	766
Sample 4	2.10	76.02	21.87	100	1,425
Stereotypes of Black people[a]					
Sample 1	53.39	9.83	36.78	100	1,168
Sample 2	49.62	12.81	37.57	100	775
Sample 3	50.31	11.34	38.35	100	734
Sample 4	59.61	13.85	26.54	100	1,355
Differential stereotypes					
Sample 2	31.85	21.62	46.53	100	773
Sample 3	30.08	18.40	51.51	100	723
Sample 4	28.68	25.63	45.69	100	1,353
Symbolic Racism[a]					
Sample 1	37.81	3.78[a]	58.41	100	1,210
Sample 2	34.89	6.33	58.79	100	803
Sample 3	30.96	11.28	57.76	100	766
Racial Resentment[a]					
Sample 1	21.49	14.34	64.16	100	1,187
Sample 2	18.09	18.27	63.64	100	787
Sample 3	18.05	19.33	62.62	100	750
Sample (short version) 4	21.85	13.93	64.22	100	1,365
Implicit prejudice					
Sample 1 (AMP)	37.58	12.75	49.66	100	1,215
Sample 2 (AMP)	34.55	11.49	53.96	100	806
Sample 3 (AMP)	30.19	11.07	58.74	100	791
Sample 4 (AMP)	32.70	15.52	51.78	100	1,441
Sample 4 (IAT)	32.41	0.00	67.59	100	1,441

Note: All average scores were coded to range from 0 to 1. The absence of prejudice toward African-Americans is indicated by a score of .5 on all prejudice scales. Values above 0.5 indicate negative attitudes toward African Americans or more positive attitudes toward White people than toward Black people (differential scales).

[a] The non-differential measures of anti-Black affect, stereotypes, symbolic racism, and racial resentment do not possess a clearly labeled neutral midpoint. A value of .5 was chosen following Pasek et al. (2009), but it is not clear that respondents equated the middle answer choices with a neutral attitude.

AMP, 49.7 percent (survey 1), 54.0 percent (survey 2), 58.7 percent (survey 3), and 51.8 percent (survey 4) of the respondents scored above 0.5, which indicates negative attitudes toward Black people. The IAT in survey 4 yielded a larger figure: 67.6 percent of White Americans harbored implicit bias against African Americans.[2] Since both the AMP and the IAT have mid-points that are thought to be neutral (though see Blanton et al., 2015), these figures can be interpreted as indicating that a majority of the White American public holds prejudice against African Americans, though certainly not close to 100 percent.[3]

The measures of "new" racism identified similarly large proportions of biased participants. Between 57.8 percent (Sample 3) and 64.2 percent (Sample 4) of White Americans harbored negative attitudes toward Black people according to the symbolic racism and racial resentment scales. Neither of these measures includes an explicitly neutral midpoint. For instance, one question in both batteries asks respondents whether they agree or disagree (on a five-point rating scale) with the statement, "It's really a matter of some people just not trying hard enough; if Blacks would only try harder, they could be just as well off as Whites." The middle response option, "neither agree nor disagree," does not necessarily indicate a neutral attitude toward African Americans on its surface. However, previous work suggests that 0.5 is a reasonable value to use to identify neutrality on these batteries (Pasek et al., 2009), so we have used it here.

Differential measures of stereotypes identified slightly smaller percentages than those identified by the other measures: 45.7 percent to 51.5 percent reported characterizations of Black people that were less favorable than those of White people. Because the implicit attitude measures are all differential comparisons of evaluations of White people and Black people, these differential stereotype measures seem most comparable to them, and they yielded similar percentages of anti-Black White Americans: about 50 percent, plus or minus. However, the differential feeling thermometer scores indicated notably fewer White Americans manifesting anti-Black prejudice: 18.7 percent to 24.1 percent.

When examining the explicit affect and cognition measures that were *not* differential (i.e., ratings only of Black people), the proportions manifesting anti-Black prejudice were even smaller: 26.5 percent to 38.4 percent of White Americans characterized Black people as having unfavorable attributes as gauged by the stereotypes questions, and only 4.6 percent and 8.7 percent White Americans indicated disliking African Americans by rating them below the neutral point of 50 on the feeling thermometer.

Thus, the results in Table I.1 can be viewed as supporting a wide array of conclusions. First, one can argue that traditional measures, new racism measures, and implicit measures all document about half (or a bit more) of White Americans being anti-Black. Or one could claim (based on the IAT only) that about two-thirds of White Americans are anti-Black.

[2] Interestingly, this number is close to the 75 percent documented using data from the convenience sample of White people who voluntarily visited the Project Implicit website to diagnose their own level of implicit bias against Black people (Greenwald & Banaji, 2013, p. 47).

[3] It is important to note that these figures were generated using the midpoint as the precise location of neutrality, whereas measurement error no doubt caused some neutral respondents to score slightly above or below the midpoint. We have no way to estimate the numbers of such respondents in these surveys, but if they were present, the proportions of anti-Black respondents reported here are overestimates.

And one could argue that according to some explicit measures, only about one-third or even fewer White Americans manifest anti-Black attitudes. In other words, the conclusion one would reach about the prevalence of prejudice depends upon which measure one chooses to rely upon. And one cannot say unequivocally that measures of "new" racism and implicit measures document notably more anti-Black prejudice than do traditional, explicit measures, though one can certainly cherry-pick results by noting that the IAT yielded a much, much larger proportion of apparently prejudiced White people (about two-thirds) than did the non-differential affect measure (fewer than one-tenth).

Thus, in the face of this evidence, it is difficult to sustain the claim that explicit measures yield lower estimates of the prevalence of anti-Black prejudice because of social desirability bias. But it is certainly possible that social desirability bias attenuates the proportion of prejudiced White people documented by differential explicit measures and that once such bias is corrected for, the proportion of White people manifesting prejudice might be even higher. We turn to this possibility next.

Social Desirability Bias

Developers and users of implicit prejudice measures have often asserted that many research participants are likely unwilling to openly admit their prejudices toward social groups, especially Black people (Ito et al., 2015; Kim, 2003; Olson & Fazio, 2003, 2004; Orey et al., 2013). And this is of course a completely reasonable hypothesis. Traditional, explicit measures of prejudice and measures of "new" racism rely on people's willingness to accurately know and report their thoughts and feelings about social groups. Yet with changing social norms, perhaps people who are prejudiced feel that they will be punished if they express prejudice, so they are reluctant to be honest if asked to do so.

The debate about so-called cancel culture highlights these concerns; disdain has been voiced about the increasingly prevalent tendency to shame, attack, and boycott someone who has voiced an opposing or controversial opinion. More than 150 academics, writers, intellectuals, and other public figures signed an open letter on July 7, 2020, arguing that constructive, open social debate is at risk if people are not allowed to voice opinions that deviate from the norm – particularly if that norm reflects only the opinions of a small but loud "woke" group (Harper's, 2020). Given how aggressive some reactions have been to opinions that diverged from the norm, it is easy to imagine that many people may be unwilling to honestly report prejudicial attitudes in a survey.

Although the idea of social desirability response bias is well-established in the academic literature, the evidence cited in support for such a bias in the measurement of racial prejudice has several shortcomings. We will discuss here different types of evidence that are often cited as documenting the impact of social desirability bias: race-of-interviewer effects, experiments involving public and private response modes, randomized response techniques, the item count technique, and the bogus pipeline technique.

Race-of-Interviewer Effects

One type of evidence that has been used to discourage the use of traditional explicit prejudice questions and "new" racism measures is the effects of the race of the survey interviewer on the level of prejudice reported. More than 100 studies have been conducted on the effect of the race of a survey interviewer, and these race-of-interviewer effects are particularly

strong on questions that explicitly mention race (Anderson et al., 1988b; Davis, 1997; Schuman & Converse, 1971). The most prominent example of such race-of-interviewer effects comes from the 1971 Detroit Area Study. In that study, White respondents were randomly assigned to be interviewed by either a Black or a White interviewer. Only 26 percent of respondents interviewed by a White interviewer said that they would not "mind if a relative married a Negro." In contrast, a whopping 72 percent of respondents who were interviewed by a Black interviewer said that they would not mind (Hatchett & Schuman, 1975). This can be viewed as evidence of powerful social desirability bias, a difference of forty-six percentage points.

However, that same study also examined responses to other measures of racial attitudes and documented notably weaker effects: thirty-five percentage points on whether "Negro" and White children should attend the same school, thirty-one percentage points on the question of whether respondents would be disturbed if a "Negro with the same income and education as you moved into your block," and nine percentage points on the question of whether "Negro" and White children should be allowed to play together. One cannot deny that these numbers are quite different from one another, and if only one such result had been discovered, 9 percent is much smaller than 46 percent.

Nonetheless, these sorts of findings have often been put forward as evidence of social desirability bias in prejudice questions: respondents were not willing to honestly reveal their racial attitudes because they thought a Black interviewer might react negatively to unfavorable opinions about Black people (e.g., Anderson et al., 1988a, 1988b; Finkel et al., 1991). And it is interesting that researchers have uncritically accepted this evidence as proving that explicit measures underestimate the level of prejudice in the population due to social desirability bias. In fact, there is no evidence from Hatchett and Schuman (1975) or any other of the studies in this literature about which reports are more valid: the reports to White interviewers or the reports to Black interviewers.

Perhaps we are so inclined to accept the social desirability hypothesis that we assume the reports to White interviewers are the more valid ones, despite no evidence of that. But in fact, it is possible that reports made to White interviewers are the dishonest ones, intended to avoid appearing unprejudiced to people who might harbor anti-Black attitudes. Whether or not this seems plausible, studies of interviewer effects have not routinely ruled it out. So it is possible that reports to Black interviewers might have been more honest, and as a result, White interviewers caused the survey to overestimate true levels of prejudice.

Another important problem with many of the race-of-interviewer studies involving face-to-face interviews is that they did not randomly assign interviewers to respondents (e.g., Anderson et al., 1988b; Dohrenwend et al., 1968; Schaeffer, 1980). Random assignment rarely happens in face-to-face interviews, particularly with representative samples, because of the cost of sending interviewers across the country. Accordingly, interviewers typically interview people who live near them.

Therefore, interviewers who interview respondents of a different race are more likely to live in racially diverse areas than are interviewers who interview only respondents of their own racial group. Research has established that racial diversity facilitates intergroup contact, which in turn can reduce prejudice (Allport, 1954; Pettigrew & Tropp, 2006). Hence, an alternative explanation for the race-of-interviewer effects seemingly observed in these studies is that people who

are interviewed by an interviewer of another race actually have more favorable attitudes toward other racial groups. Thus, what looks like a race-of-interviewer effect is actually due to the higher likelihood of Black interviewers to interview White people with positive attitudes toward African Americans. Again, lack of direct evidence of which reports are more accurate leaves this ambiguity open.

Studies of race-of-interviewer effects in telephone surveys were often able to prevent this problem by randomly assigning interviewers to telephone numbers (Cotter et al., 1982; Davis, 1997; Finkel et al., 1991). These studies therefore seem to provide stronger evidence of social desirability bias. However, Kim and colleagues (2019) recently put forward an alternative – and largely overlooked – explanation for the seeming race-of-interviewer effect in telephone surveys. Kim et al. (2019) analyzed telephone surveys conducted before the 2008 United States presidential election. In line with previous research, what appeared to be a race-of-interviewer effect was observed: Black interviewers recorded more respondents saying they would vote for Barack Obama than did White interviewers. This suggests that the more interviews that are conducted by Black interviewers in a study, the larger would be the apparent proportion of Americans voting for Mr. Obama.

However, what looked like a race-of-interviewer effect turned out to be a consequence of differential respondent recruitment. Most people can identify the race of a person on the phone (Baugh, 2000), and the similarity-attraction hypothesis from social psychology suggests that an interviewer will be especially successful at recruiting participation from a potential respondent of the same race. That is, Black interviewers may be more likely than White interviewers to successfully recruit Black respondents. And indeed, this pattern was observed by Kim et al. (2019). Furthermore, White respondents who harbor anti-Black prejudice might be more reluctant to be interviewed by a Black person than a White person.

Taken together, this reasoning suggests that the proportions of respondents who were either Black or who were nonprejudiced Whites might have been greater among people interviewed by Black interviewers than among people interviewed by White interviewers. And not surprisingly, the proportion of Black respondents who said they would vote for Mr. Obama was greater than the proportion of White respondents who said they would vote for him. After controlling for the race of the respondent, there was no longer an effect of the race of the interviewer on reported voting intentions. Thus, although differences in results obtained by Black and White interviewers might have been the result of racial prejudice, they were not necessarily the result of intentional misreporting during the interviews.

To be sure, we are not claiming that there are no race-of-interviewer effects. We are merely pointing out that the basis to cite these effects as evidence for social desirability bias is less strong than has sometimes been assumed and that alternative explanations may account for observed evidence.

Public and Private Response Modes

Another widely cited body of evidence for distortion of reports due to social desirability bias and against the use of explicit measures of prejudice comes from studies that compared responses in survey modes with an interviewer present to responses in modes without an interviewer (e.g., Krysan, 1998; Tourangeau & Smith, 1996). Because respondents can reasonably expect interviewers to pass judgment on them based on their answers to survey questions, the social desirability argument suggests that people should be more willing to honestly

report embarrassing or controversial opinions and behaviors when no interviewer is present.

In line with this notion, research has found more reports of controversial behaviors (e.g., the number of sexual partners or drug use) being provided when respondents completed a questionnaire anonymously on a computer than when respondents provided responses to questions orally to interviewers (Tourangeau et al., 1995; Tourangeau & Smith, 1996). An early meta-analysis comparing computer-administered questionnaires to face-to-face interviews documented more reports of sensitive information in the former (Richman et al., 1999). Later studies have also documented higher frequencies of controversial or undesirable behaviors and opinions being reported in online surveys than in interviewer-administered modes, such as telephone surveys (Chang & Krosnick, 2009, 2010; Holbrook & Krosnick, 2010; Kreuter et al., 2008). Again, it is tempting to conclude that the presence of an interviewer causes reluctance to admit embarrassing things about oneself, so the reports made on computers are presumed to be more valid than the reports made to interviewers.

But there is an alternative view of this finding as well (see Holbrook & Krosnick, 2010; Kreuter et al., 2008). Specifically, reports of more undesirable attitudes and behavior given under complete privacy do not necessarily reveal more accuracy. This is possible because past studies did not typically assess accuracy and instead assumed that more reports are indicators of more honesty. Instead, interviewers may create a greater sense of accountability when respondents answer questions than when questionnaires are completed privately. And this accountability may inspire respondents to devote more cognitive effort to answering questions and to answer more accurately.

With regard to prejudice measurement, offering privacy to respondents may result in sloppy answering behavior that reduces the validity of prejudice measures. In line with this, Lelkes et al. (2012) found in three experiments that guaranteeing participants complete anonymity when completing a questionnaire sometimes increased socially undesirable answers but consistently decreased the accuracy of the responses. We should therefore not be too quick to equate findings of more prejudice in private answering modes with better measurement due to less social desirability bias, as some scholars have done (e.g., Gnambs & Kaspar, 2014; Piston, 2010; Richman et al., 1999; Tourangeau et al., 1995).

Stark et al. (2022) explored this alternative explanation using data from the 2008 American National Election Study, which contained an experiment that explored the effect of answering prejudice questions with and without potential social desirability concerns. Before and after the 2008 presidential election, the same representative national sample of 1,009 White non-Hispanic Americans participated in two computer-assisted interviews conducted by an interviewer in the respondents' homes. During both interviews, respondents answered two explicit prejudice questions about how (1) -hard-working and (2) intelligent they thought White people and Black people are. During the second interview, these questions were asked and answered orally. But during the first interview, these questions were asked with Audio Computer-Assisted Self-Interviewing (ACASI): the interviewer handed over a laptop to the respondent, who heard the questions being read aloud on headphones and saw the questions and answer choices on the laptop screen. Answers were typed confidentially on the keyboard without the interviewer seeing or hearing the questions or the answers. ACASI is thought to eliminate impression management concerns.

In line with social desirability bias, the mean ratings of Black people as hard-working

and intelligent were more pro-Black in the face-to-face interviews than in the ACASI interviews. On scales that were coded to range from 0 to 1 (1 meaning most hard-working), the mean rating of Black people decreased from 0.527 in the oral mode to 0.490 in the ACASI mode ($p < .001$). Also, significantly more participants said in the private mode that White people are more hard-working than Black people (oral: 44.97 percent, ACASI: 49.34 percent, $p = .018$) and more intelligent than Black people (oral: 39.74 percent, ACASI: 43.96 percent, $p = .018$). This fits the narrative of many proponents of implicit prejudice measures, as these results suggest that openly asking explicit prejudice questions in a survey interview leads to an underestimation of the real level of anti-Black prejudice (Piston, 2010).

However, not in line with this reasoning, significantly more people said that Whites are *less* hard-working and *less* intelligent than Black people in the private answer mode than in the oral administration mode (hard-working oral: 3.21 percent, ACASI: 8.46 percent, $p < .001$; intelligent oral: 1.75 percent, ACASI: 3.38 percent, $p = .037$). As a consequence, there was hardly any difference between modes in the differential measures comparing White and Black people's hard-working nature ($\Delta = 0.007$, $p = .119$) and intelligence ($\Delta = 0.012$, $p = .001$) on scales ranging from 0 to 1. The social desirability bias discussion typically centers around social norms that prevent the open expression of negative attitudes toward other social groups. Evidence of a substantive number of respondents seemingly hiding their negative attitudes toward their own group (Whites) does not fit with this narrative.

There was also no evidence that the privately measured attitudes were more valid than the orally reported attitudes when predictive validity was assessed by using the stereotypes measure to predict other explicitly measured attitudes and implicit measures. If oral reports of stereotypes suffer from distortion due to social desirability response bias, then the privately assessed attitudes should be more strongly related to known correlates of prejudice. This particular study examined the correlation with attitudes toward homosexuals that were measured in ACASI mode (and presumably undistorted by social desirability response bias), and this correlation did not vary by administration mode. Moreover, orally reported stereotypes of Black people were more strongly associated with implicit prejudice toward Black people measured with the Affect Misattribution Procedure (AMP) than were the stereotypes of Black people reported privately in ACASI mode. If implicit measures of prejudice are not affected by social desirability response bias, then one would expect a stronger association of the implicit measure with privately expressed prejudices than with openly expressed prejudices, but the opposite was observed.

We are not claiming that social desirability bias does not exist nor that it does not distort responses to explicit prejudice questions. Instead, we are claiming that the literature cited to support this conclusion using studies comparing administration modes is not as informative as it might seem. There is convincing evidence that offering privacy to respondents so that they do not have to fear judgment for their opinions leads to more reports of prejudicial attitudes. However, few studies were able to compare the reports in the private answer mode to a benchmark that allowed conclusions about the validity of these reports (for exceptions, though not with regard to prejudice, see Holbrook & Krosnick, 2010; Kreuter et al., 2008; Tourangeau & Smith, 1996). Most existing research relies on the "more-is-better" assumption (Krumpal, 2013) and equates reports of more controversial attitudes and behavior to more honest answers.

Stark et al.'s (2022) study suggests that this conclusion might be premature, as reports of "more" prejudice do not necessarily imply more accurate measurement.

Randomized Response Technique

Another prominent form of evidence for distortion of explicit prejudice measures by social desirability bias comes from studies implementing techniques that guarantee absolute privacy to respondents. One such approach is the randomized response technique (RRT). This procedure makes it impossible for the interviewer to know whether a respondent gave a prejudiced answer or not.

Various versions of RRTs have been developed that can be classified into forced-response variants and forced-question variants (Krumpal, 2013). All variants involve some form of randomizing (e.g., a coin toss or dice throw) that the respondent conducts without the interviewer being able to see the outcome. In forced-response RRTs, the outcome of the randomizing determines whether the respondent will answer a sensitive question truthfully (e.g., answer "yes" or "no" when a coin toss shows heads) or to simply respond with "yes" (when the coin shows tails). In forced-question RRTs, the outcome of the randomizing determines which of two questions the respondent will answer, only one of which is a sensitive question, and the other of which has a known expected distribution of responses. The interviewer conducting the RRT cannot know which question the respondent answered, and it is possible to implement calculations to derive the proportion of respondents who reported an embarrassing attribute.

Many studies have found more socially undesirable responses with the RRT than when the same question is asked directly of respondents. Because RRTs are typically limited to yes/no responses, the majority of studies concerned socially undesirable behaviors such as stealing (Wimbush & Dalton, 1997) or cheating (Scheers & Dayton, 1987) that people have either committed or not (for overviews, see Holbrook & Krosnick, 2010; Lensvelt-Mulders et al., 2005). But there are also studies that found more reports of socially undesirable attitudes such as prejudice with the RRT than in direct questions (e.g., anti-Semitism, Krumpal, 2012). For instance, a study in Germany found that about 25 percent more people said that they would mind if their hypothetical daughter had a relationship with a dark-skinned Nigerian immigrant with an RRT than when they were asked directly (Ostapczuk et al., 2009). This result is in line with the notion that social desirability response bias distorts results of explicit prejudice questions.

However, a shortcoming of most RRT studies is their reliance on the "more-is-better" assumption that also limits research comparing public and private survey modes. That is, without being able to compare the results to an objective benchmark, researchers can only assume that more undesirable responses in the private condition reflect more accurate answers (Krumpal, 2012).[4]

Moreover, plenty of studies have not found more undesirable reports using RRTs than using direct questions (e.g., Akers et al., 1983; Danermark & Swensson, 1987), and some even reported fewer such reports in RRTs than in

[4] This is particularly problematic for attitude questions for which benchmarks are typically not available. A few studies on undesirable behavior found reports in the RRT condition that matched medical or administrative records of the population more closely than open responses (Lensvelt-Mulders et al., 2005). However, scholars have pointed out that comparisons with population benchmarks does not guarantee that the response of an individual participant is indeed more accurate (Holbrook & Krosnick, 2010).

direct questions (e.g., Brewer, 1981; Holbrook & Krosnick, 2010; Williams & Suen, 1994). This suggests that the technique does not always work as intended. In fact, research has found that many respondents disobeyed the RRT instruction (Boeije & Lensvelt-Mulders, 2002; Ostapczuk et al., 2009). This may be caused by the complex and unexpected instructions of many RRTs (why do you need to toss a coin during an interview?) and the difficulty of understanding how the randomization guarantees respondents' privacy. As a consequence, some respondents may not trust that their privacy is protected. For instance, Holbrook and Krosnick (2010) found that the coins of respondents should have come up "heads" in 147 percent of the tosses to produce the results they found. Although new versions of RRTs have been developed that can account for such response biases to some extent (Coutts & Jann, 2011; Cruyff et al., 2007), the evidence in favor of social desirability bias produced by existing RRT studies is less unambiguous than many think.

Item Count Technique

Another method to collect information on sensitive attitudes and behavior is the item count technique (ICT, Miller, 1984), which is also sometimes called the list technique (Kuklinski, Sniderman, et al., 1997) or the unmatched count technique (Coutts & Jann, 2011). This approach has a clear advantage over the RRT as it can be implemented without requiring respondents to conduct a randomization themselves. In the ICT, respondents might be shown a list of statements and asked how many of these statements they find upsetting. Kuklinski, Sniderman and colleagues (1997) used this list:

- The federal government increasing the tax on gasoline
- Professional athletes getting million-dollar-plus salaries
- Large corporations polluting the environment

To ensure privacy, respondents are not asked which statements are upsetting, only how many. Respondents are randomly assigned to see the list of these three statements or to see a list with one additional and sensitive statement, such as "A Black family moving next door to you." The proportion of respondents who are upset by a Black family moving in next door can be inferred by comparing the mean number of statements reported in the condition with the short list to the mean number reported in the condition with the added statement.

Just like the randomized response technique, the ICT has been criticized for only reporting population averages that do not allow measurements of individual opinions or behaviors. However, statistical techniques have been developed that enable researchers to gauge correlations between variables (Blair & Imai, 2012; Holbrook & Krosnick, 2010; Imai, 2011). And the double list variant, the LISITT approach, and the item sum technique overcome issues of inefficient estimators and low statistical power (Corstange, 2009; Glynn, 2013; Trappmann et al., 2014).

Many of the existing ICT studies on prejudice are made ambiguous by the possibility of floor and ceiling effects (Kuklinski, Cobb, & Gilens, 1997; Martinez & Craig, 2010; Redlawsk et al., 2010). Such effects occur when respondents believe that all or none of the non-sensitive statements are upsetting, for example. That is, a respondent might be upset by increasing the tax on gasoline, million-dollar-plus salaries of athletes, and pollution by large corporations. This respondent cannot admit being upset by a Black family moving in next door, because the answer "four" would explicitly reveal this bias. Similarly, reporting

that none of the four concepts are upsetting would reveal a respondents' opinion about a Black family moving in next door. Such floor and ceiling effects have often been mentioned (Kuklinski, Cobb, & Gilens, 1997; Martinez & Craig, 2010; Redlawsk et al., 2010), but statistical techniques to detect and adjust for them have only been proposed fairly recently (Blair & Imai, 2012) and implemented rarely. To minimize this risk, researchers should select foil statements that are negatively correlated with one another (e.g., making it legal for two men to marry and teaching intelligent design in school) to decrease the likelihood that respondents will be upset by all or none of the statements (Glynn, 2013; Janus, 2010). But this has not always been done.

Another potential problem with the ICT is that respondents' evaluations of the foil statements might be changed by offering the sensitive statement, rendering responses to the short and long lists incomparable (Blair & Imai, 2012). Much research in psychology has shown that perceptions of objects change when perceived in relation to specific other objects (e.g. Higgins & Lurie, 1983; Stevens, 1957). Such perceptual contrast effects can take two forms in an ICT. First, by adding a statement that people find not at all upsetting, the upsetting nature of the other statements may change. For instance, compared to a Black family moving in next door, pollution by a large corporation may seem even more upsetting. This would lead to an overestimation of prejudice in the ICT.

Second, the evaluation of the additional statement may change when it is presented in a list as compared to when it stands on its own. Knoll (2013), for instance, found that 20 percent fewer Americans reported feeling that the American culture and way of life are threatened by foreign influence when asked in an ICT compared to an open question. Knoll (2013) argued that this is because some groups feel the social pressure to overreport feelings of threat, but it could just as well have been that the foreign influence felt much less threatening in comparison to the pollution of large corporations.

Furthermore, for an ICT to effectively measure racial attitudes, the sensitive item must be an unambiguous measure of prejudice that cannot be rejected due to principled, non-race-related stances (Sniderman & Hagendoorn, 2007). For instance, finding a government-implemented affirmative action policy upsetting could be due to prejudice against African Americans or due to a principled opposition to governmental intervention (Kuklinski, Sniderman, et al., 1997). Research on Islamophobia showed that negative attitudes toward Muslims are closely related to the rejection of Muslim practices such as wearing headscarves or Islamic education (Van der Noll, 2014). However, adding these practices to a list experiment could lead to an overestimation of negative attitudes toward Muslims because some groups of people reject such practices based on principles such as gender equality or separation of church and state without harboring prejudice toward Muslims (Adelman & Verkuyten, 2020; Dangubić et al., 2020; Sleijpen et al., 2020).

In sum, the ICT is a promising approach to avoiding social desirability response bias if the technique is implemented well. However, the existing evidence for such social desirability bias with regard to prejudice should be taken with a pinch of salt due to floor and ceiling effects (Kuklinski, Cobb, & Gilens, 1997; Martinez & Craig, 2010; Redlawsk et al., 2010) and perceptual contrast effects (Blair & Imai, 2012).

Bogus Pipeline

Another method used to eliminate social desirability bias is the bogus pipeline technique, wherein the researcher pretends to measure

people's "true" attitudes via their physical behavior (Jones & Sigall, 1971; Sigall & Page, 1971). For example, in one study, some undergraduates were randomly assigned to answer questions orally, and other undergraduates answered the same questions under "bogus pipeline" conditions, meaning that they were attached to what they believed to be a lie detector (Jones & Sigall, 1971). The idea is that embarrassment of being caught lying weighs more strongly than the potential embarrassment of holding an undesirable attitude. In line with this reasoning, White participants said that various derogatory attributes were truer of "Negroes" under the latter condition than under the former (Jones & Sigall, 1971). A meta-analysis of thirty-one studies confirmed that the bogus pipeline technique indeed elicits more reports of undesirable attitudes and behavior (Roese & Jamieson, 1993). This also applies to racial prejudice (e.g., Carver et al., 1978; Plant et al., 2003). However, because of the necessity of a lab setup wherein people must be attached to a bogus lie detector, traditional bogus pipeline approaches are not well suited to measuring prejudice in representative samples of the general public.

More recent applications of the bogus pipeline technique circumvent the necessity of a lab setup. Instead of pretending to measure people's true attitudes through their physical behavior, participants have simply been told that the study contains "sophisticated methods developed by psychologists" to detect deception, so lying is futile (Cohen et al., 2009, p. 293). For example, participants in one study were told the following information about a question regarding helping behavior ("How often do you stop for stranded motorists? (never, rarely, sometimes, usually, always)"): "This question might appear innocent enough, but, in fact, it is one of many tools psychologists use to detect people who lie to create a positive impression of themselves. With the possible exception of policemen on patrol, NO ONE 'usually' or 'always' stops for stranded motorists. People who say they do are most likely lying." This simple instruction led to a significant increase in antisemitic responses (Cohen et al., 2009). Although not widely used and ethically problematic because of its deception, this approach could be used in national surveys. In fact, some researchers have already implemented it to reduce socially desirable responses in online surveys (e.g. Beattie, 2017).

Like the other techniques reviewed in this section, however, bogus pipeline research has uniformly relied on the "more-is-better" assumption in order to claim that it yields more accurate measurements, even though more reports of undesirable attitudes do not necessarily imply more accurate reports (Lelkes et al., 2012). In fact, bogus pipeline techniques have been criticized for eliciting more undesirable responses only because of situational demands put on the participants (Arkin, 1981; Ostrom, 1973). Thus, it remains unclear if reports of more prejudice in a bogus-pipeline setup reflects more accurate reports.

Conclusions about Social Desirability

We conclude this discussion by returning to the question with which we began: Is social desirability response bias the explanation for any discrepancy between traditional explicit measures of prejudice and implicit measures? The research we have reviewed provides some evidence that socially desirable response bias can lead people to misrepresent their attitudes when they answer an explicit prejudice question. However, we have also identified a series of problems in the evidence that make socially desirable bias in prejudice measures less convincing than is generally assumed.

We have discussed alternative explanations for race-of-interviewer effects in studies that

did not use random assignment of interviewers. Moreover, some of the race-of-interviewer effects in studies with random assignment can be explained by differential non-response, not social desirability bias. We have also shown evidence that higher reports of prejudice in survey modes that provided absolute privacy may be partially or largely caused by less accurate answering. Moreover, results of studies making use of the randomized response technique and the item count technique are not as diagnostic as one would hope. The fundamental problem here is that researchers have relied on the "more is better" rule rather than directly assessing the accuracy of reports collected by various different methods. Therefore, it seems prudent to conclude that the literature on social desirability response bias contaminating explicit reports of prejudice is suggestive but not conclusive. Clearly, more work is needed. But because the evidence reported here that explicit prejudice against African Americans does not seem notably less common than implicit prejudice suggests that social desirability response bias is not an obvious source of error in explicit reports.

Critique of Implicit Measures

Although scholars generally agree that implicit bias exists, there is much disagreement about the quality of the existing instruments to measure implicit bias. Questions have been raised about the validity and reliability of implicit measures, as well as the quality of the data used to support some of the existing claims about implicit bias. There are also debates about what the existing instruments measure and how these measurements relate to people's behavior. Also, questions have been raised about the possibility that implicit attitudes can be changed and thereby produce a change in behavior, a notion that the original conceptualization of implicit attitudes deems exceedingly difficult to accomplish. Next, we discuss these concerns, as they set the stage for chapters in this book.

Issues of Validity and Reliability

Some observers have expressed concern about the near zero correlations often observed between explicit and implicit measures of the same construct. Of course, such correlations are not expected to be 1.0, because that would imply that the measures are entirely redundant and that the implicit measures provide no information not already conveyed by the explicit measures. But most attitude theories suggest that explicit acknowledgments of prejudice are likely to have roots in implicit attitudes, so there should be non-negligible correspondence between the two.

In the four surveys we analyzed earlier, correlations between the implicit measures (the AMP and IAT) and the explicit measures (disliking, stereotypes, symbolic racism) varied between $r = .06$ and $r = .30$ (with most being about $r = .20$). These numbers are similar to correlations between explicit and implicit measures in meta-analyses, which varied between .20 and .24 (Cameron et al., 2012; Hofmann et al., 2005). Data from the Project Implicit website yielded a correlation of $r = .27$ with regard to race-related attitudes (Nosek, Smyth, et al., 2007). In contrast, correlations between explicit measures of prejudice tend to be notably stronger. For instance, the surveys we analyzed earlier yielded correlations mostly above $r = .55$.

The weak correlations between explicit and implicit measures have been explained by some observers as being due to social desirability bias contaminating the explicit measures (e.g., Nosek & Smyth, 2007). In line with this argument, Nier (2005) found that race IAT scores were more strongly associated with answers to the explicit Modern Racism scale

in a bogus pipeline condition than in a condition without the bogus pipeline. However, Phillips and Olson (2014) showed that the weak correlation is not necessarily caused by social desirability bias in response to explicit prejudice questions. Their research suggests that the weak correlation might occur because people tend to deliberate about their responses to explicit prejudice questions, whereas implicit measures pick up people's gut feelings.

Scholars have also argued that the dissimilarity of the tasks that participants perform when answering an explicit question and when completing implicit measures might explain the weak correlations between implicit and explicit measurements (Hofmann et al., 2005). For instance, implicit measures such as the IAT or the AMP are measures of a preference for one group over another (e.g., a preference for White faces over Black faces). However, studies often compared such implicit measures to explicit measures that did not assess a preference and instead tapped attitudes toward individual objects (e.g., Cunningham et al., 2004; Nier, 2005). Another popular explanation considers implicit and explicit attitudes to be "related but distinct" constructs that explain different portions of variance in people's decisions and behaviors (Bar-Anan & Vianello, 2018; Blair et al., 2015; Nosek & Smyth, 2007). From this perspective, the low correlation is to be expected and is not a reason to question validity.

A more problematic finding that raises serious questions about the validity of implicit measures of prejudice is the fact that different measures of implicit bias correlate extremely weakly with one another (Blanton et al., 2015; Bosson et al., 2000; Brauer et al., 2000; Olson & Fazio, 2003; Rudolph et al., 2008). In the 2008 ANES data, the correlation between the AMP and IAT was only $r = .18$. If different implicit measures tap the same underlying evaluation of social groups, why are the measures essentially unrelated to one another?

One common explanation for such weak correlations between implicit measures is random measurement error (Nosek, Greenwald, & Banaji, 2007). Such error seems inevitable, and, if random, will weaken correlations between measures (Cunningham et al., 2001). One of the most common solutions to this problem is the application of multiple different measures to the same person. Scores on these multiple measures can be averaged so that random measurement errors cancel out, or the multiple measures can be used in latent variable structural equation models, which do an even better job of eliminating attenuation of relations due to random error. But one of the great strengths of the popular implicit attitude measures is that they each combine large sets of measurements to yield total scores. Thus, random measurement error in individual assessments seems unlikely to explain the strikingly weak correlations between different implicit measures.

The weak correlation between implicit measures is particularly concerning given that implicit bias has long been presented as an automatic and relatively uncontrollable characteristic that is resistant to change (Greenwald & Banaji, 1995). Studies in the early 2000s challenged this conceptualization of implicit bias, showing that scores on implicit measures such as the IAT are reactive to small external changes (e.g. Barden et al., 2004; Blair et al., 2001; Boysen et al., 2006; Dasgupta & Greenwald, 2001; Park et al., 2007; Wittenbrink et al., 2001). For instance, White college students manifested lower implicit racial bias on a race IAT in the presence of a Black experimenter than in the presence of a White experimenter (Lowery et al., 2001). This context-dependency has led to a reconceptualization of implicit bias as a person's attitude that is, just like attitudes in general, influenced

by the context the person is in (Greenwald & Banaji, 2013; Jost, 2019).

Although contextual influence may, at least partially, explain the low correlation between implicit measures, this raises a different question about the validity of such measures. The fact that the experimenter's race leads to lower implicit bias scores (Lowery et al., 2001) seems surprisingly similar to the race-of-interviewer effects discussed above (Anderson et al., 1988b; Davis, 1997; Schuman & Converse, 1971). Does this imply that implicit measures of prejudice are susceptible to social desirability response bias? If so, this would undermine one of the motivations to dismiss explicit measures of prejudice in favor of implicit measures.

Concerns have also been voiced about the unsatisfactory reliability of some measures of implicit bias. Ideally, a measurement instrument should lead to very similar conclusions if it is applied to the same people. However, research found low reliability of evaluative or affective priming, particularly when pictures of outgroup members were used (Banse, 2001). Reliability was much better when the method used unambiguous group labels instead of pictures (De Houwer, 2009). The reason for this low reliability might lie in the fact that priming methods do not make salient the category that is being primed (e.g., African Americans) and tap attitudes toward the used stimuli instead of the category (De Houwer, 2009; Olson & Fazio, 2003).

Research has also shown that the most widely used implicit measure, the IAT, has only weak to modest test-retest reliability (Bar-Anan & Nosek, 2014; Blanton et al., 2016; Devine et al., 2012). Over time, IAT scores have been found to be considerably less stable than scores on corresponding explicit measures (Gawronski et al., 2017). Similarly, the AMP also shows only modest test-retest reliability (Cooley & Payne, 2017). Whereas some consider this low reliability a central concern (Tetlock & Mitchell, 2009), others have pointed out that this concern is based on the no longer widely endorsed perception of implicit bias as a stable individual trait (Jost, 2019).

The view that implicit attitudes are stable individual characteristics changed in the early 2000s (Banaji, 2001; Cunningham et al., 2001). However, most scholars continued to focus on differences between participants in terms of implicit attitude measurements and other variables – an approach that assumes that the implicit measures tap stable individual differences. Newer accounts suggest that implicit bias does not reflect chronically stable individual differences and is instead responsive to situational triggers (Payne et al., 2017) or that there is a mix of stable and context-dependent components (Dentale et al., 2019). From these perspectives, low reliability scores are expected.

Low reliability may thus not be a reason for serious concern if implicit bias is considered context-dependent, but it does raise questions about the informativeness of a person's score on an implicit prejudice measure. For instance, after completing a race IAT on the Project Implicit website, a person may be told that he or she has a preference for White people over Black people. But, importantly, this result may be attributable to the context in which the measurement was made and not generalizable to other contexts. In other words, the person may be misled into believing they have a general disposition when no such thing exists. Some scholars have suggested treating implicit measurements like the results of blood pressure tests, in that multiple measurements made on the same person may produce different but equally valid results describing that person differently in different situations (Greenwald et al., 2019; Jost, 2019). Unfortunately, most research does not follow this logic and continues to use single applications of implicit measures to assess implicit bias of participants.

A final concern about implicit measures comes from the fact that people are quite able to guess their scores on these measures (Hahn et al., 2014). This suggests that people are not in fact unaware of their implicit attitudes. That again challenges a fundamental assumption of the implicit bias perspective.

Issues with the Metric

Questions have been also raised about the scoring of responses in the IAT that may lead to an overestimation of implicit bias (Blanton & Jaccard, 2006). Because the race IAT is based on a comparison of reaction times to White and Black faces in tasks involving categorizing words as good or bad, it seems natural to equate equally fast reaction times with no preference for Black people over White people or vice versa (Greenwald et al., 2019). Interestingly, the zero point on the IAT matches the neutral point on explicit measures (Pasek et al., 2009). But neutral behavior toward Black and White people corresponds to a point above zero on the IAT, a point normally interpreted as meaning pro-White preference (Blanton et al., 2015). This suggests that the proportion of people manifesting anti-Black prejudice on the IAT is smaller than the proportion of people with positive scores.

This finding might lead one to prefer to focus on associations between implicit bias scores and other variables (Jost, 2019). But this will not satisfy the interest in knowing how much implicit bias exists and how that quantity compares with the amount of explicit bias that is present. Furthermore, individuals want to know if they are prejudiced, and learning that requires interpreting absolute levels of implicit association test scores. Implicit bias scholars have routinely offered assessments such as that 68 percent of visitors of the Project Implicit website (Nosek, Smyth, et al., 2007) or 75 percent of White Americans (Greenwald & Banaji, 2013, p. 47) are implicitly biased against Black people. And the Project Implicit website invites people to take the test to learn if they are prejudiced. Thus, the desire for absolute scores seems powerful yet is now called into question. A related problem is the use of labels such as slight, moderate, or strong to characterize levels of implicit bias when those are arbitrary distinctions (Blanton & Jaccard, 2008). This raises the fundamental question of whether it is appropriate to inform study participants about their IAT scores (Blanton & Jaccard, 2006; Mitchell & Tetlock, 2017).

Sample Quality and Generalizability

As we have highlighted above, the vast majority of publications on implicit bias are based on data from non-probability samples that do not allow one to draw solid conclusions about the American adult population (or any other population). For example, many studies have explored correlates of implicit bias or interventions targeting implicit bias among college students. Since there are notable differences between students and non-students in terms of their scores on prejudice measures (Henry, 2008), it remains unclear how well these findings generalize. Moreover, hundreds of other studies make use of voluntarily completed IATs from people who visited the Project Implicit website to learn about themselves. The collection of millions of measurements does not legitimize generalizations to any populations. Only a very few studies have administered implicit prejudice measures with representative national samples (Ditonto et al., 2013; Kalmoe & Piston, 2013; Pasek et al., 2009, 2014; Payne et al., 2010), so most of the findings in this literature remain to be tested with representative samples and should be, since there is profound interest in applying prejudice measures to understand the nation.

Links to Behavior

Another line of research has raised the critique that few studies have documented sizable correlations between measures of implicit bias and discriminatory behavior (Oswald et al., 2013; Tetlock & Mitchell, 2009). Many of the existing studies on this matter have relied on data from university students and have found only weak correlations that appeared only under specific conditions (Blommaert et al., 2012). The number of studies that link implicit bias to real-world discrimination is increasing (Rooth, 2010; Van den Bergh et al., 2010), but the difficulty of implementing the measurement of implicit bias outside of the laboratory puts some limitation on their generalizability.

Meta-analyses of the existing literature (focusing on lab studies) have documented only weak to modest correlations ($r = .13$ to $r = .24$) between implicit racial bias measured by the IAT and discriminatory behaviors (Greenwald et al., 2009, 2015; Oswald et al., 2015) and a modest correlation of $r = .28$ between sequential priming measures and behavior (Cameron et al., 2012). In addition, a meta-analysis concluded that interventions aimed at reducing implicit bias in various domains rarely lead to changes in behavior (Forscher et al., 2019). A systematic review found a preference for White over Black people among most physicians but little evidence of a relationship with the physicians' decision making (Dehon et al., 2017). These kinds of findings have led to criticism (Mitchell & Tetlock, 2017) of earlier claims such as "the automatic White preference expressed in the Race IAT is now established as signaling discriminatory behavior" (Greenwald & Banaji, 2013, p. 47).

However, another recent meta-analysis on the link between the IAT and intergroup behavior revealed that the low correlations can partly be explained by methodological shortcomings in the existing literature (Kurdi et al., 2019). Many studies were underpowered, did not correct for measurement error, used variants of the IAT instead of the standard IAT, and had weak correspondence between the IAT and the criterion behavior. Among studies without these methodological shortcomings, the correlation between implicit attitudes and behavior was $r = .37$, which is certainly larger than that documented in earlier meta-analysis. However, these correlations were similar in size to correlations between explicit measures of attitudes and behavior (Kurdi et al., 2019). Accordingly, the conclusion that implicit measures of prejudice manifest more predictive validity than do explicit measures seems unsustainable at present, as is the claim that implicit attitudes are a primary driver of discriminatory behavior generally. However, this particular meta-analysis has been criticized for applying selection criteria and coding schemes that are insufficiently rigorous (Blanton & Jaccard, Chapter 12, this volume), so more work is needed on this issue.

Based on dual-process theories of cognition (Olson & Fazio, 2008; Petty et al., 2007), some scholars have argued that implicit prejudice measures are particularly well suited for predicting attitudes and behaviors that are automatic and less so for attitudes and behaviors that are more deliberate (see Blair et al., 2015; Dovidio et al., 2002). In line with these arguments, a meta-analysis found stronger correlations between implicit priming measures and behavior in studies in which participants had low motivation or low opportunity to deliberate their behavior (Cameron et al., 2012). Thus, the value of implicit measures for understanding prejudice may be limited to conditions of automatic thinking and action.

Can Implicit Bias Be Changed?

Decades of research in political science, sociology, and psychology have shown how

difficult it is to change opinions, and the slow decline of explicit anti-Black attitudes during the past sixty years illustrates how slowly prejudices typically change. If bias toward a social group is unconscious, changing this attitude may be even harder. However, there are a lot of well-meaning efforts around the world devoted to teaching people how to reduce their implicit and explicit bias, even though there is little evidence of how this can be done. Yet recent developments give hope that difficult attitude change may sometimes be attainable. For example, attitudes toward gay marriage have changed from a majority of Americans opposing it to a vast majority being in favor of it. And a recent study showed that attitudes toward transgender people can be lastingly changed through a relatively brief intervention (Broockman & Kalla, 2016).

An increasing number of studies have explored how implicit prejudice can be reduced. Research from lab studies showed that exposing people to counter-stereotypical members of a negatively associated social group can weaken both explicit and implicit racial bias (Dasgupta & Greenwald, 2001). For instance, exposure to Barack Obama during the 2008 election cycle was associated with lower levels of implicit bias against Black people (Plant et al., 2009). The problem in the current political climate – at least in the United States – is that people who harbor bias against minorities are more likely to be exposed to reinforcing negative information about members of these groups due to selective use of social media. This may fortify implicit bias, because negative behaviors by members of a social group influence implicit evaluations of a person from that group more strongly than do positive behaviors (Ratliff & Nosek, 2011).

More recent research has examined intervention programs attempting to reduce implicit bias. One study found that a combination of raising awareness of implicit bias and teaching bias reduction strategies for daily life reduced implicit racial bias (Devine et al., 2012). An imagined intergroup contact intervention has also been found to reduce implicit bias toward immigrants among school children (Vezzali et al., 2012). Moreover, a program that trained participants to use meaningful negation ("that's wrong") instead of simple negation ("no") reduced implicit racial bias significantly (Johnson et al., 2018). And a diversity training program reduced implicit bias against women in STEM among male (but not female) university faculty (Jackson et al., 2014).

Despite these promising results, many scholars are cautious about whether implicit attitudes can be changed easily. Many studies found that attempts to reduce implicit bias can be ineffective or even backfire and increase the bias (Dasgupta & Greenwald, 2001). Processes that have reduced explicit prejudices, such as increasing propositional knowledge, correction of prior information, and raising awareness of the likely consequences of information, have no effect on implicit bias (e.g. Petty et al., 2006; Teachman et al., 2003). Another strategy that has been deemed ineffective and sometimes even increases implicit bias is giving individuals global nonspecific instructions (e.g., "don't be biased" Frantz et al., 2004; Kim, 2003). Nonetheless, a highly publicized meta-analysis concluded that interventions can change implicit bias, though the effects are generally weak (Forscher et al., 2019). Importantly, changes in implicit bias did not necessarily translate into changes in behavior.

If implicit bias is a reflection of the cultural institutions in a society at a given moment in time (Payne et al., 2017), this bias may change as the cultural institutions change (Hardin & Banaji, 2013). Similar to changes in explicit prejudice (Krysan, 2011), research found that implicit biases with regard to race, skin-tone, and sexuality have become more egalitarian

during the last decade (Charlesworth & Banaji, 2019). Similar changes have not been observed with regard to obesity, age, or disability – perhaps because societal debate has centered less on discrimination with regard to these characteristics. Thus, implicit bias may be subject to change not as the result of training programs but rather because of changes in cultural institutions, the same institutions that are thought to give rise to implicit bias in the first place.

Conclusion

The issues we have discussed concerning claims about explicit and implicit measures of prejudice are reasons to take stock of the current state of affairs. Evidence suggests that explicit prejudice is currently alive and well and perhaps even widespread in contemporary America. And explicit measures of prejudice, which have long been considered ineffective due to social desirability concerns, might prove to be valuable after all. At the same time, implicit measures of prejudice and efforts to reduce that prejudice may merit reconsideration as well.

Jost (2019, p. 15) recently argued that advocates of research on implicit bias have moved on from the idea that implicit measures can and should replace explicit measures altogether: "The IAT as a 'magic bullet'—a panacea for solving the world's ongoing problems with racism and sexism and classism—is dead." This argument supports our notion that research on prejudice and bias has entered a fourth period, one in which explicit and implicit measures will help us understand societal problems by looking at the world from both angles. A meta-analysis revealed that implicit attitudes often explained additional variance after explicit attitudes are taken into account (Kurdi et al., 2019). This supports the notion that both types of measures can contribute to our understanding of social phenomena and that they can complement each other.

In sum, various factions of scholars are on various sides of these debates, and some efforts have been made to bring the factions together, though so far without complete success. There is no widely agreed-upon endorsement among relevant scholars of what we know about the measures and operation of prejudice and future research in this area is needed to confront and resolve deep concerns about whether explicit and implicit measures are indeed valid and of scholarly or practical value. We hope this volume moves scholars in that direction by fully considering the range of available evidence and the challenging questions raised by it.

References

Adelman, L., & Verkuyten, M. (2020). Prejudice and the acceptance of Muslim minority practices: A person-centered approach. *Social Psychology*, *51*(1), 1–16. https://doi.org/10.1027/1864-9335/a000380

Akers, R. L., Massey, J., Clarke, W., et al. (1983). Are self-reports of adolescent deviance valid? Biochemical measures, randomized response, and the bogus pipeline in smoking behavior. *Social Forces*, *62*(1), 234–251. https://doi.org/10.1093/sf/62.1.234

Allport, G. W. (1954). *The Nature of Prejudice*. Boston, MA: Addison-Wesley.

Anderson, B. A., Silver, B. D., & Abramson, P. R. (1988a). The effects of race of the interviewer on measures of electoral participation by Blacks in SRC National Election Studies. *Public Opinion Quarterly*, *52*(1), 53–83. https://doi.org/10.1086/269082

Anderson, B. A., Silver, B. D., & Abramson, P. R. (1988b). The effects of the race of the interviewer on race-related attitudes of Black respondents in SRC/CPS National Election Studies. *Public Opinion Quarterly*, *52*(3), 289–324. https://doi.org/10.1086/269108

Arkin, R. M. (1981). Self-presentational styles. In J. T. Tedeschi (Ed.), *Impression Management*

Theory and Social Psychological Research (pp. 311–333). New York, NY: Academic Press.

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science*, *9*(8), 896–906. https://doi.org/10.1177/1948550617728995

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger III, J. S. Nairne, I. E. Neath, & A. M. Surprenant (Eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder* (pp. 117–150). Washington, DC: APA.

Banks, A. J., & Hicks, H. M. (2016). Fear and implicit racism: Whites' support for voter ID laws. *Political Psychology*, *37*(5), 641–658. https://doi.org/10.1111/pops.12292

Banse, R. (2001). Affective priming with liked and disliked persons: Prime visibility determines congruency and incongruency effects. *Cognition and Emotion*, *15*(4), 501–520. https://doi.org/10.1080/0269993004200213

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, *46*(3), 668–688. https://doi.org/10.3758/s13428–013-0410-6

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

Barden, J., Maddux, W. W., Petty, R. E., et al. (2004). Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes. *Journal of Personality and Social Psychology*, *87*(1), 5–22. https://doi.org/10.1037/0022-3514.87.1.5

Baugh, J. (2000). *Beyond Ebonics: Linguistic Pride and Racial Prejudice*. Oxford: Oxford University Press.

Beattie, P. (2017). Anti-semitism and opposition to Israeli government policies: The roles of prejudice and information. *Ethnic and Racial Studies*, *40*(15), 2749–2767. https://doi.org/10.1080/01419870.2016.1260751

Blair, I. V., Dasgupta, N., & Glaser, J. (2015). Implicit Attitudes. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition* (pp. 665–691). Washington, DC: APA.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, *20*(1), 47–77. https://doi.org/10.1093/pan/mpr048

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, *81*(5), 828–841. https://doi.org/10.1037/0022-3514.81.5.828

Blanton, H., Burrows, C. N., & Jaccard, J. (2016). To accurately estimate implicit influences on health behavior, accurately estimate explicit influences. *Health Psychology*, *35*(8), 856–860. https://doi.org/10.1037/hea0000348

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, *34*, 277–297. https://doi.org/10.1146/annurev.soc.33.040406.131632

Blanton, H., Jaccard, J., Strauts, E., et al. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology*, *100*(5), 1468–1481. https://doi.org/10.1037/a0038379

Blommaert, L., van Tubergen, F., & Coenders, M. (2012). Implicit and explicit interethnic attitudes and ethnic discrimination in hiring. *Social Science Research*, *41*(1), 61–73. https://doi.org/10.1016/j.ssresearch.2011.09.007

Bobo, L. (2001). Racial attitudes and relations at the close of the twentieth century. In N. J. Smelser, W. J. Wilson, & F. Mitchell (Eds.), *America Becoming: Racial Trends and Their Consequences*, vol. 1 (pp. 264–301). Washington, DC: National Academy Press.

Boeije, H., & Lensvelt-Mulders, G. (2002). Honest by chance: A qualitative interview study to clarify respondents' (non-)compliance with computer-assisted randomized response. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*, *75*(1), 24–39. https://doi.org/10.1177/075910630207500104

Bogardus, E. S. (1933). A social-distance scale. *Sociology and Social Research*, *17*, 265–271.

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. https://doi.org/10.1037/0022-3514.79.4.631

Boysen, G. A., Vogel, D. L., & Madon, S. (2006). A public versus private administration of the implicit association test. *European Journal of Social Psychology*, *36*(6), 845–856. https://doi.org/10.1002/ejsp.318

Brauer, M., Wasel, W., & Niedenthal, P. (2000). Implicit and explicit components of prejudice. *Review of General Psychology*, *4*(1), 79–101. https://doi.org/10.1037/1089-2680.4.1.79

Brewer, K. R. W. (1981). Estimating marihuana usage using randomized response – some paradoxical findings. *Australian Journal of Statistics*, *23*(2), 139–148. https://doi.org/10.1111/j.1467-842X.1981.tb00771.x

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307–324. https://doi.org/10.1037/0033-2909.86.2.307

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate? *Journal of Social Issues*, *55*(3), 429–444. https://doi.org/10.1111/0022-4537.00126

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, *352*(6282), 220–224. https://doi.org/10.1126/science.aad9713

Brown, R., & Zagefka, H. (2005). Ingroup affiliations and prejudice. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the Nature of Prejudice: Fifty Years after Allport*. Oxford: Blackwell.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Burke, S. E., Dovidio, J. F., LaFrance, M., et al. (2017). Beyond generalized sexual prejudice: Need for closure predicts negative attitudes toward bisexual people relative to gay/lesbian people. *Journal of Experimental Social Psychology*, *71*, 145–150. https://doi.org/10.1016/j.jesp.2017.02.003

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. https://doi.org/10.1177/1088868312440047

Carmines, E. G., Sniderman, P. M., & Easter, B. C. (2011). On the meaning, measurement, and implications of racial resentment. *Annals of the American Academy of Political and Social Science*, *634*, 98–116. https://doi.org/10.1177/0002716210387499

Carter, J. S., & Corra, M. (2016). Racial resentment and attitudes toward the use of force by police: An over-time trend analysis. *Sociological Inquiry*, *86*(4), 492–511. https://doi.org/10.1111/soin.12136

Carver, C. S., Glass, D. C., & Katz, I. (1978). Favorable evaluations of Blacks and the Handicapped: Positive prejudice, unconscious denial, or social desirability? *Journal of Applied Social Psychology*, *8*(2), 97–106. https://doi.org/10.1111/j.1559-1816.1978.tb00768.x

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678. https://doi.org/10.1093/poq/nfp075

Chang, L., & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires. *An experiment. Public Opinion Quarterly*, *74*(1), 154–167. https://doi.org/10.1093/poq/nfp090

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Cheung, J. H., Burns, D. K., Sinclair, R. R., et al. (2017). Amazon Mechanical Turk in

organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361. https://doi.org/10.1007/s10869–016-9458-5

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cohen, F., Jussim, L., Harber, K. D., et al. (2009). Modern anti-semitism and anti-Israeli attitudes. *Journal of Personality and Social Psychology*, *97*(2), 290–306. https://doi.org/10.1037/a0015338

Cooley, E., & Payne, B. K. (2017). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin*, *43*(1), 46–59. https://doi.org/10.1177/0146167216675331

Cornesse, C., Blom, A. G., Dutwin, D., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Corstange, D. (2009). Sensitive questions, truthful answers? Modeling the list experiment with LISTIT. *Political Analysis*, *17*(1), 45–63. https://doi.org/10.1093/pan/mpn013

Cotter, P. R., Cohen, J., & Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, *46*(2), 278–284. https://doi.org/10.1086/268719

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research*, *40*(1), 169–193. https://doi.org/10.1177/0049124110390768

Crowson, H. M., Brandes, J. A., & Hurst, R. J. (2013). Who opposes rights for persons with physical and intellectual disabilities? *Journal of Applied Social Psychology*, *43*(Suppl.2), 307–318. https://doi.org/10.1111/jasp.12046

Cruyff, M. J. L. F., Van Den Hout, A., Van Der Heijden, P. G. M., et al. (2007). Log-linear randomized-response models taking self-protective response behavior into account. *Sociological Methods and Research*, *36*(2). https://doi.org/10.1177/0049124107301944

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin*, *30*(10), 1332–1346. https://doi.org/10.1177/0146167204264654

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

Cvencek, D., Greenwald, A. G., Brown, A. S., et al. (2010). Faking of the implicit association test is statistically detectable and partly correctable. *Basic and Applied Social Psychology*, *32*(4), 302–314. https://doi.org/10.1080/01973533.2010.519236

Danermark, B., & Swensson, B. (1987). Measuring drug use among Swedish adolescents: Randomized response versus anonymous questionnaires. *Journal of Official Statistics*, *3*(4), 439–448.

Dangubić, M., Verkuyten, M., & Stark, T. H. (2020). Understanding (in)tolerance of Muslim minority practices: A latent profile analysis. *Journal of Ethnic and Migration Studies*, *47*(7). https://doi.org/10.1080/1369183X.2020.1808450

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. https://doi.org/10.1037/0022-3514.81.5.800

Dasgupta, N., & Stout, J. G. (2012). Contemporary discrimination in the lab and field: Benefits and obstacles of full-cycle social psychology. *Journal of Social Issues*, *68*(2), 399–412. https://doi.org/10.1111/j.1540-4560.2012.01754.x

Davis, D. W. (1997). The direction of race of interviewer effects among African-Americans: Donning the Black Mask. *American Journal of Political Science*, *41*(1), 309–322. https://doi.org/10.2307/2111718

De Houwer, J. (2009). Comparing measures of attitudes at the procedural and functional level.

In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the New Implicit Measures*. London: Psychology Press.

DeBell, M., Krosnick, J. A., & Lupia, A. (2010). *Methodology Report and User's Guide for the 2008–2009 ANES Panel Study*. Stanford University and the University of Michigan. https://electionstudies.org/wp-content/uploads/2009/03/anes_specialstudy_2008_2009panel_MethodologyRpt.pdf

Dehon, E., Weiss, N., Jones, J., et al. (2017). A systematic review of the impact of physician implicit racial bias on clinical decision making. *Academic Emergency Medicine*, *24*(8), 895–904. https://doi.org/10.1111/acem.13214

Dentale, F., Vecchione, M., Ghezzi, V., et al. (2019). Applying the latent state-trait analysis to decompose state, trait, and error components of the self-esteem implicit association test. *European Journal of Psychological Assessment*, *35*(1), 78–85. https://doi.org/10.1027/1015-5759/a000378

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

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

Ditonto, T. M., Lau, R. R., & Sears, D. O. (2013). AMPing racial attitudes: Comparing the power of explicit and implicit racism measures in 2008. *Political Psychology*, *34*(4), 487–510. https://doi.org/10.1111/pops.12013

Dohrenwend, B. S., Colombotos, J., & Dohrenwend, B. P. (1968). Social distance and interviewer effects. *Public Opinion Quarterly*, *32*(3), 410–422. https://doi.org/10.1086/267624

Dovidio, J. F., Brigham, J. C., Johnson, B. T., et al. (1996). Stereotyping, prejudice, and discrimination: Another look. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and Stereotyping* (pp. 276–322). London: Guilford Press.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68. https://doi.org/10.1037/0022-3514.82.1.62

Enders, A. M., & Scott, J. S. (2019). The increasing racialization of american electoral politics, 1988–2016. *American Politics Research*, *47*(2), 275–303. https://doi.org/10.1177/1532673X18755654

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

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. https://doi.org/10.1207/s15324834basp2704_3

Filindra, A., & Kaplan, N. J. (2016). Racial resentment and Whites' gun policy preferences in contemporary America. *Political Behavior*, *38*(2), 255–275. https://doi.org/10.1007/s11109–015-9326-4

Finkel, S. E., Guterbock, T. M., & Borg, M. J. (1991). Race-of-interviewer effects in a preelection poll: Virginia 1989. *Public Opinion Quarterly*, *55*(3), 313–330. https://doi.org/10.1086/269264

Forscher, P., Lai, C. K., Devine, P. G., et al. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522–559. https://doi.org/10.17605/OSF.IO/DV8TU

Frantz, C. M., Cuddy, A. J. C., Burnett, M., et al. (2004). A threat in the computer: The race implicit association test as a stereotype threat experience. *Personality and Social Psychology Bulletin*, *30*(12), 1611–1624. https://doi.org/10.1177/0146167204266650

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality*

Psychology. Cambridge: Cambridge University Press, pp. 283–310.

Gawronski, B., Morrison, M., Phills, C. E., et al. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin*, *43*(3), 300–312. https://doi.org/10.1177/0146167216684131

Glynn, A. N. (2013). What can we learn with statistical truth serum? *Public Opinion Quarterly*, *77*(S1), 159–172. https://doi.org/10.1093/poq/nfs070

Gnambs, T., & Kaspar, K. (2014). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*(4), 1237–1259. https://doi.org/10.3758/s13428–014-0533-4

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Toledo, OH: Doubleday Anchor Books.

Goldman, S. K. (2012). Effects of the 2008 Obama presidential campaign on White racial prejudice. *Public Opinion Quarterly*, *76*(4), 663–687. https://doi.org/10.1093/Poq/Nfs056

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

Greenwald, A. G., & Banaji, M. R. (2013). *Blindspot: Hidden Bias of Good People*. New York, NY: Delacorte.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. https://doi.org/10.1037/pspa0000016

Greenwald, A. G., Brendl, M., Cai, H., et al. (2019). The Implicit Association Test at age 20: What is known and what is not known about implicit bias. University of Washington. Retrieved from: https://psyarxiv.com/bf97c

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., et al. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Hahn, A., Judd, C. M., Hirsh, H. K., et al. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Hardin, C. D., & Banaji, M. R. (2013). The nature of implicit prejudice, implications for personal and public policy. In E. Shafir (Ed.), *The Behavioral Foundations of Public Policy*. Princeton, NJ: Princeton University Press, pp. 13–31.

Harper's. (2020, July 7). *A Letter on Justice and Open Debate*. Harper's Magazine. https://harpers.org/a-letter-on-justice-and-open-debate/

Hatchett, S., & Schuman, H. (1975). White respondents and race-of-interviewer effects. *Public Opinion Quarterly*, *39*(4), 523–528. https://doi.org/10.1086/268249

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*(2), 49–71. https://doi.org/10.1080/10478400802049936

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology*, *23*(2), 253–283. https://doi.org/10.1111/0162-895X.00281

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*, 575–604. https://doi.org/10.1146/annurev.psych.53.100901.135109

Higgins, E. T., & Lurie, L. (1983). Context, categorization, and recall: The "change-of-standard" effect. *Cognitive Psychology*, *15*(4), 525–547.

Hofmann, W., Gawronski, B., Gschwendner, T., et al. (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. https://doi.org/10.1177/0146167205275613

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37–67. https://doi.org/10.1093/poq/nfp065

Holbrook, A. L., Krosnick, J. A., & Pfent, A. M. (2007). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In J. Lepkowski, B. Harris-Kojetin, P. J. Lavrakas, C. Tucker, E. de Leeuw, M. Link, M. Brick, L. Japec, & R. Sangster (Eds.), *Advances in Telephone Survey Methodology*. New York, NY: Wiley.

Holmes, J. D. (2009). Transparency of self-report racial attitude scales. *Basic and Applied Social Psychology*, *31*(2), 95–101. https://doi.org/10.1080/01973530902876884

Huddy, L., & Feldman, S. (2009). On assessing the political effects of racial prejudice. *Annual Review of Political Science*, *12*, 423–447. https://doi.org/10.1146/Annurev.Polisci.11.062906.070752

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association*, *106*(494), 407–416. https://doi.org/10.1198/jasa.2011.ap10415

Ito, T. A., Friedman, N. P., Bartholow, B. D., et al. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187–218. https://doi.org/10.1037/a0038557

Jackson, S. M., Hillard, A. L., & Schneider, T. R. (2014). Using implicit bias training to improve attitudes toward women in STEM. *Social Psychology of Education*, *17*(3), 419–438. https://doi.org/10.1007/s11218–014-9259-5

Janus, A. L. (2010). The influence of social desirability pressures on expressed immigration attitudes. *Social Science Quarterly*, *91*(4), 928–946. https://doi.org/10.1111/j.1540-6237.2010.00742.x

Johnson, I. R., Kopp, B. M., & Petty, R. E. (2018). Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes. *Group Processes & Intergroup Relations*, *21*(1), 88–110. https://doi.org/10.1177/1368430216647189

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349–364. https://doi.org/10.1037/h0031617

Jost, J. T. (2019). The iat is dead, long live the IAT: Context-sensitive measures of implicit attitudes are indispensable to social and political psychology. *Current Directions in Psychological Science*, *28*(1), 10–19. https://doi.org/10.1177/0963721418797309

Judd, C. M., Park, B., Ryan, C. S., et al. (1995). Stereotypes and ethnocentrism: Diverging interethnic perceptions of African American and White American youth. *Journal of Personality and Social Psychology*, *69*(3), 460–481. https://doi.org/10.1037/0022-3514.69.3.460

Kalmoe, N. P., & Piston, S. (2013). Is implicit prejudice against blacks politically consequential? Evidence from the AMP. *Public Opinion Quarterly*, *77*(1), 305–322.

Kim, D. Y. (2003). Voluntary controllability of the implicit association test (IAT). *Social Psychology Quarterly*, *66*(1), 83–96. https://doi.org/10.2307/3090143

Kim, N., Krosnick, J. A., & Lelkes, Y. (2019). Race of interviewer effects in telephone surveys preceding the 2008 U.S. Presidential Election. *International Journal of Public Opinion Research*, *31*(2), 220–242. https://doi.org/10.1093/ijpor/edy005

Kinder, D. R. (1986a). Presidential character revisited. In R. R. Lau & D. O. Sears (Eds.), *Political Cognition*. New York, NY: Erlbaum, pp. 233–255.

Kinder, D. R. (1986b). The continuing American dilemma: White resistance to racial change 40 years after Myrdal. *Journal of Social Issues*, *42*(2), 151–171. https://doi.org/10.1111/j.1540-4560.1986.tb00230.x

Kinder, D. R., & Sanders, L. M. (1996). *Divided by Color: Racial Politics and Democratic Ideals*. Chicago, IL: University of Chicago Press.

Kinder, D. R., & Sears, D. O. (1981). Prejudice and politics – symbolic racism versus racial threats to

the good life. *Journal of Personality and Social Psychology*, *40*(3), 414–431.

Knoll, B. R. (2013). Assessing the effect of social desirability on nativism attitude responses. *Social Science Research*, *42*(6), 1587–1598. https://doi.org/10.1016/j.ssresearch.2013.07.012

Knuckey, J. (2017). The myth of the "Two Souths?" Racial resentment and White Party identification in the Deep South and Rim South. *Social Science Quarterly*, *98*(2), 728–749. https://doi.org/10.1111/ssqu.12322

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research*, *41*(6), 1387–1403. https://doi.org/10.1016/j.ssresearch.2012.05.015

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity*, *47*, 2025–2047. https://doi.org/10.1007/s11135–011-9640-9

Krysan, M. (1998). Privacy and the expression of White racial attitudes : A comparison across three contexts. *The Public Opinion Quarterly*, *62*(4), 506–544. https://doi.org/10.1086/297859

Krysan, M. (2011). Data Update to Racial Attitudes in America. An update and website to complement H. Schuman, C. Steeh, L. Bobo, and M. Krysan (Eds.), *Racial Attitudes in America: Trends and Interpretations*, revised edition, 1997. Harvard, MA: Harvard University Press. www.igpa.uillinois.edu/programs/racial-attitudes/

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial attitudes and the "New South." *The Journal of Politics*, *59*(2), 323–349. https://doi.org/10.2307/2998167

Kuklinski, J. H., Sniderman, P. M., Knight, K., et al. (1997). Racial prejudice and attitudes toward affirmative action. *American Journal of Political Science*, *41*(2), 402. https://doi.org/10.2307/2111770

Kurdi, B., Seitchik, A. E., Axt, J. R., et al. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *American Psychologist*, *74*(5), 569–586. https://doi.org/10.1037/amp0000364

Lelkes, Y., Krosnick, J. A., Marx, D. M., et al. (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, *48*(6), 1291–1299. https://doi.org/10.1016/j.jesp.2012.07.002

Lensvelt-Mulders, G. J. L. M., Hox, J. J., Van Der Heijden, P. G. M., et al. (2005). Meta-analysis of randomized response research thirty-five years of validation. *Sociological Methods and Research*, *33*(3), 319–348. https://doi.org/10.1177/0049124104268664

Levin, S., van Laar, C., & Sidanius, J. (2003). The effects of ingroup and outgroup friendships on ethnic attitudes in college: A longitudinal study. *Group Processes & Intergroup Relations*, *6*(1), 76–92. https://doi.org/10.1177/1368430203006001013

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, *81*(5), 842–855. https://doi.org/10.1037/0022-3514.81.5.842

MacInnis, B., Krosnick, J. A., Ho, A. S., et al. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, *82*(4), 707–744. https://doi.org/10.1093/poq/nfy038

MacInnis, C. C., Boss, H. C. D., & Bourdage, J. S. (2020). More evidence of participant misrepresentation on Mturk and investigating who misrepresents. *Personality and Individual Differences*, *152*, 109603. https://doi.org/10.1016/j.paid.2019.109603

Mackie, D. M., & Smith, E. R. (1998). Intergroup relations: Insights from a theoretically integrative approach. *Psychological Review*, *105*(3), 499–529. https://doi.org/10.1037/0033-295X.105.3.499

Malhotra, N., & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences

about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with nonprobability samples. *Political Analysis*, *15*(3), 286–323. https://doi.org/10.1093/pan/mpm003

Martinez, M. D., & Craig, S. C. (2010). Race and 2008 presidential politics in Florida: A list experiment. *Forum*, *8*(2). https://doi.org/10.2202/1540-8884.1316

McConahay, J. B. (1986). Modern racism, ambivalence, and the modern racism scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, Discrimination, and Racism*. New York, NY: Academic Press, pp. 91–125.

McElwee, S., & McDaniel, J. (2017, May 8). Economic Anxiety Didn't Make People Vote Trump, Racism Did. *The Nation*. www.thenation.com/article/archive/economic-anxiety-didnt-make-people-vote-trump-racism-did/

Miller, J. D. (1984). *A New Survey Technique For Studying Deviant Behavior*. George Washington University.

Mitchell, G., & Tetlock, P. E. (2017). Popularity as a poor proxy for utility. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*. New York, NY: Wiley-Blackwell, pp. 164–195). https://doi.org/10.1002/9781119095910.ch10

Moberg, S. P., Krysan, M., & Christianson, D. (2019). Racial attitudes in America. *Public Opinion Quarterly*, *83*(2), 450–471. https://doi.org/10.1093/poq/nfz014

National Public Radio, the Robert Wood Johnson Foundation, and the Harvard T. H. Chan School of Public Health. (2018). Discrimination in America: Final Summary. www.rwjf.org/content/dam/farm/reports/surveys_and_polls/2018/rwjf443620

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A Bogus Pipeline approach. *Group Processes and Intergroup Relations*, *8*(1), 39–52. https://doi.org/10.1177/1368430205048615

Norton, A. T., & Herek, G. M. (2013). Heterosexuals' attitudes toward transgender people: Findings from a national probability sample of U.S. *Adults. Sex Roles*, *68*(11–12), 738–753. https://doi.org/10.1007/s11199–011-0110-6

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic Processes in Social Thinking and Behavior*. New York, NY: Psychology Press, pp. 265–292.

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test. *Experimental Psychology*, *54*(1), 14–29. https://doi.org/10.1027/1618-3169.54.1.14

Nosek, B. A., Smyth, F. L., Hansen, J. J., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science*, *14*(6), 636–639. https://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. https://doi.org/10.1207/s15324834basp2601_1

Olson, M. A., & Fazio, R. H. (2008). 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*. New York, NY: Psychology Press, pp. 19–63.

Orey, B. D. A., Craemer, T., & Price, M. (2013). Implicit racial attitude measures in black samples: IAT, subliminal priming, and implicit black identification. *Political Science and Politics*, *46*(3), 550–552. https://doi.org/10.1017/S1049096513000644

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, *39*(6), 920–931. https://doi.org/10.1002/ejsp.588

Ostrom, T. M. (1973). The bogus pipeline: A new ignis fatuus? *Psychological Bulletin*, *79*(4), 252–259. https://doi.org/10.1037/h0033861

Oswald, F. L., Mitchell, G., Blanton, H., et al. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192. https://doi.org/10.1037/a0032734

Oswald, F. L., Mitchell, G., Blanton, H., et al. (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. https://doi.org/10.1037/pspa0000023

Park, J., Felix, K., & Lee, G. (2007). Implicit attitudes toward Arab-Muslims and the moderating effects of social information. *Basic and Applied Social Psychology*, *29*(1), 35–45. https://doi.org/10.1080/01973530701330942

Pasek, J., Stark, T. H., Krosnick, J. A., et al. (2014). Attitudes toward Blacks in the Obama era: Changing distributions and impacts on job approval and electoral choice, 2008–2012. *Public Opinion Quarterly*, *78*(S1), 276–302. https://doi.org/10.1093/poq/nfu012

Pasek, J., Tahk, A., Lelkes, Y., et al. (2009). Determinants of turnout and candidate choice in the 2008 US presidential election illuminating the impact of racial prejudice and other considerations. *Public Opinion Quarterly*, *73*(5), 943–994. https://doi.org/10.1093/Poq/Nfp079

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality Assessment via Questionnaires: Current Issues in Theory and Measurement*. London: Springer, pp. 143–165.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp. 49–69). New York, NY: Erlbaum.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317. https://doi.org/10.1037/0022-3514.60.2.307

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

Payne, B. K., Krosnick, J. A., Pasek, J., et al. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology*, *46*(2), 367–374. https://doi.org/10.1016/j.jesp.2009.11.001

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*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Petty, R. E., Briñol, P., & DeMarree, K. C. (2007). The Meta-Cognitive Model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, *25*(5), 657–686. https://doi.org/10.1521/soco.2007.25.5.657

Petty, R. E., Tormala, Z. L., Brinol, P., et al. (2006). Implicit ambivalence from attitude change: an exploration of the PAST model. *Journal of Personality and Social Psychology*, *90*(1), 21.

Phillips, J. E., & Olson, M. A. (2014). When implicitly and explicitly measured racial attitudes align: The roles of social desirability and thoughtful responding. *Basic and Applied Social Psychology*, *36*(2), 125–132. https://doi.org/10.1080/01973533.2014.881287

Piston, S. (2010). How explicit racial prejudice hurt Obama in the 2008 election. *Political Behavior*,

32(4), 431–451. https://doi.org/10.1007/slll09–010-9108-y

Plant, E. A., Devine, P. G., & Brazy, P. C. (2003). The bogus pipeline and motivations to respond without prejudice: Revisiting the fading and faking of racial prejudice. *Group Processes & Intergroup Relations*, *6*(2), 187–200. https://doi.org/10.1177/1368430203006002004

Plant, E. A., Devine, P. G., Cox, W. T. L., et al. (2009). The Obama effect: Decreasing implicit prejudice and stereotyping. *Journal of Experimental Social Psychology*, *45*, 961–964. https://doi.org/10.1016/j.jesp.2009.04.018

Rabinowitz, J. L., Sears, D. O., Sidanius, J., et al. (2009). Why do White Americans oppose race-targeted policies? Clarifying the impact of symbolic racism. *Political Psychology*, *30*(5), 805–828. https://doi.org/10.1111/J.1467-9221.2009.00726.X

Ratliff, K. A., & Nosek, B. A. (2011). Negativity and outgroup biases in attitude formation and transfer. *Personality and Social Psychology Bulletin*, *37*(12), 1692–1703. https://doi.org/10.1177/0146167211420168

Redlawsk, D. P., Tolbert, C. J., & Franko, W. (2010). Voters, emotions, and race in 2008: Obama as the first black president. *Political Research Quarterly*, *63*(4), 875–889. https://doi.org/10.1177/1065912910373554

Richman, W. L., Weisband, S., Kiesler, S., et al. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*(5), 754–775. https://doi.org/10.1037/0021-9010.84.5.754

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin*, *114*(2), 363–375. https://doi.org/10.1037/0033-2909.114.2.363

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523–534. https://doi.org/10.1016/j.labeco.2009.04.005

Rudolph, A., Schröder-Abé, M., Schütz, A., et al. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment*, *24*(4), 273–281. https://doi.org/10.1027/1015-5759.24.4.273

Schaeffer, N. C. (1980). Evaluating race-of-interviewer effects in a national survey. *Sociological Methods & Research*, *8*(4), 400–419. https://doi.org/10.1177/004912418000800403

Scheers, N. J., & Dayton, C. M. (1987). Improved estimation of academic cheating behavior using the randomized response technique. *Research in Higher Education*, *26*(1), 61–69. https://doi.org/10.1007/BF00991933

Schuman, H. (2000). The perils of correlation, the lure of labels, and the beauty of negative results. In D. O. Sears, J. Sidanius, & L. Bobo (Eds.), *Racialized Politics: The Debate about Racism in America*. Chicago, IL: University of Chicago Press.

Schuman, H., & Converse, J. M. (1971). The effects of Black and White interviewers on Black responses in 1968. *Public Opinion Quarterly*, *35*(1), 44–68. https://doi.org/10.1086/267866

Schuman, H., Steeh, C., Bobo, L., et al. (1997). *Racial Attitudes in America: Trends and Interpretations*. Harvard, MA: Harvard University Press.

Sears, D. O. (1988). Symbolic racism. In P. A. Katz & D. A. Taylor (Eds.), *Eliminating Racism: Profiles in Controversy*. New York, NY: Plenum Press, pp. 53–84.

Sears, D. O., & Henry, P. J. (2003). The origins of symbolic racism. *Journal of Personality and Social Psychology*, *85*(2), 259–275. https://doi.org/10.1037/0022-3514.85.2.259

Sears, D. O., & Henry, P. J. (2005). Over thirty years later: A contemporary look at symbolic racism. *Advances in Experimental Social Psychology*, *37*, 95–150. https://doi.org/10.1016/S0065–2601(05)37002-X

Sears, D. O., & McConahay, J. C. (1973). *The Politics of Violence: The New Urban Blacks and the Watts Riot*. New York, NY: Houghton Mifflin.

Sears, D. O., van Laar, C., Carrillo, M., et al. (1997). Is it really racism? The origins of white

Americans opposition to race-targeted policies. *Public Opinion Quarterly*, *61*(1), 16–53.

Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology*, *18*(2), 247.

Silber, H., Stark, T. H., Bloom, A. G., et al. (2018). Multi-national study of questionnaire design. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multicultural, Multinational and Multiregional (3MC) Contexts*. New York, NY: Wiley, pp. 161–179.

Simmons, A. D., & Bobo, L. D. (2018). Understanding no special favors: A quantitative and qualitative mapping of the meaning of responses to the racial resentment scale. *Du Bois Review: Social Science Research on Race*, *15*(2), 323–352. https://doi.org/10.1017/S1742058X18000310

Sleijpen, S., Verkuyten, M., & Adelman, L. (2020). Accepting Muslim minority practices: A case of discriminatory or normative intolerance? *Journal of Community and Applied Social Psychology*, *30*(4), 405–418. https://doi.org/10.1002/casp.2450

Smith, S. M., Roster, C. A., Golden, L. L., et al. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, *69*(8), 3139–3148. https://doi.org/10.1016/j.jbusres.2015.12.002

Sniderman, P. M., Crosby, G., & Howell, W. (2000). The politics of race. In D. O. Sears, J. Sidanius, & L. Bobo (Eds.), *Racialized Politics: The Debate about Racism in America*. Chicago, IL: University of Chicago Press, pp. 236–279.

Sniderman, P. M., & Hagendoorn, L. (2007). *When Ways of Life Collide: Multiculturalism and its Discontents in the Netherlands*. Princeton, NJ: Princeton University Press.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, *56*(4), 283–294. https://doi.org/10.1027/1618-3169.56.4.283

Stark, T. H., Van Maaren, F. M., Krosnick, J. A., et al. (2022). The impact of social desirability pressures on Whites' endorsement of racial stereotypes: A comparison between oral and ACASI reports in a national survey. *Sociological Methods & Research*, *51*(2), 605–631. https://doi.org/10.1177/0049124119875959

Stephens-Dougan, L. (2016). Priming racial resentment without stereotypic cues. *The Journal of Politics*, *78*(3), 687–704. https://doi.org/10.1086/685087

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181. https://doi.org/10.1037/h0046162

Stokel-Walker, C. (2018). Bots on Amazon's Mechanical Turk are ruining psychology studies. *New Scientist*. www.newscientist.com/article/2176436-bots-on-amazons-mechanical-turk-are-ruining-psychology-studies/

Tarman, C., & Sears, D. O. (2005). The conceptualization and measurement of symbolic racism. *Journal of Politics*, *67*(3), 731–761.

Teachman, B. A., Gapinski, K. D., Brownell, K. D., et al. (2003). Demonstrations of implicit anti-fat bias: The impact of providing causal information and evoking empathy. *Health Psychology*, *22*(1), 68–78. https://doi.org/10.1037/0278-6133.22.1.68

Tetlock, P. E., & Mitchell, G. (2009). Implicit bias and accountability systems: What must organizations do to prevent discrimination? *Research in Organizational Behavior*, *29*, 3–38. https://doi.org/10.1016/j.riob.2009.10.002

Tourangeau, R., Jobe, B. J., Smith, W. T., et al. (1995). Sources of error in a survey on sexual behavior. *Journal of Official Statistics*, *13*(4), 341–366.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275–304. https://doi.org/10.1086/297751

Trappmann, M., Krumpal, I., Kirchner, A., et al. (2014). Item sum: A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology*, *2*(1), 58–77. https://doi.org/10.1093/jssam/smt019

Van den Bergh, L., Denessen, E., Hornstra, L., et al. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American*

Educational Research Journal, *47*(2), 497–527. https://doi.org/10.3102/0002831209353594

Van der Noll, J. (2014). Religious toleration of Muslims in the German Public Sphere. *International Journal of Intercultural Relations*, *38*(1), 60–74. https://doi.org/10.1016/j.ijintrel.2013.01.001

Vezzali, L., Capozza, D., Giovannini, D., et al. (2012). Improving implicit and explicit intergroup attitudes using imagined contact: An experimental intervention with elementary school children. *Group Processes and Intergroup Relations*, *15*(2), 203–212. https://doi.org/10.1177/1368430211424920

Wallsten, K., Nteta, T. M., McCarthy, L. A., et al. (2017). Prejudice or principled conservatism? Racial resentment and White opinion toward paying college athletes. *Political Research Quarterly*, *70*(1), 209–222. https://doi.org/10.1177/1065912916685186

Williams, B. L., & Suen, H. (1994). A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviors. *Psychological Reports*, *75*(3), 1531–1537. https://doi.org/10.2466/pr0.1994.75.3f.1531

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology*, *82*(5), 756–763. https://doi.org/10.1037/0021-9010.82.5.756

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262–274. https://doi.org/10.1037/0022-3514.72.2.262

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology*, *81*(5), 815–827. https://doi.org/10.1037/0022-3514.81.5.815

Yeager, D. S., Krosnick, J. A., Chang, L., et al. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747. https://doi.org/10.1093/poq/nfr020

Zigerell, L. J. (2015). Distinguishing racism from ideology: A methodological inquiry. *Political Research Quarterly*, *68*(3), 521–536. https://doi.org/10.1177/1065912915586631

Appendix

Sample Description and Operationalizations

Data Source

Data for Samples 1 through 3 came from the KnowledgePanel®, an online survey panel recruited via probability sampling techniques. People were invited to join the panel by Knowledge Networks (now part of Ipsos) through random-digit-dialing telephone (RDD) calls and in later years also through postal mail sent to nationally representative lists of addresses. Sample 1 consisted of 1,215 White, non-Hispanic Americans who completed two surveys (one with explicit measures, one with the implicit measure Affect Misattribution Procedure, AMP) in late August and early September of 2008 (completion rate survey 1 = 72.4%; cumulative response rate CUMRR1 survey 1 = 10.4%; completion rate survey 2 = 62.6%; CUMRR1 survey 2 = 9.2%, see Callegaro & DiSogra, 2008). Sample 2 consisted of 1,037 White, non-Hispanic Americans who completed a survey and the AMP between late October of 2009 and early January of 2010 (completion rate = 27.1%; CUMRR1 = 3.2%). Sample 3 comprised of 791 White, non-Hispanic Americans who completed a survey and the AMP between early August and early September of 2012 (completion rate = 44.5%; CUMRR1 = 4.3%).

Data for Sample 4 came from the 2008–2009 American National Election Study (ANES) Panel Study. Participants were recruited via RDD to complete internet surveys (DeBell, Krosnick, & Lupia, 2010). The response rate was 42 percent (AAPOR RR3). Sample 4 consisted of 1,441 White, non-Hispanic Americans who completed four surveys as part of this panel study that included measures of prejudice, including the AMP and the Implicit Association Test.

Measures

Explicit Measures

Disliking (Affect)
Samples 1, 2, and 3 were asked "How much do you like or dislike each of the following groups?" Responses for the group "Blacks" on a seven-point scale were coded to range from 0 "like a great deal" to 1 "dislike a great deal." The value 0.5 represents the neutral midpoint "neither like nor dislike."

Differential Disliking (Affect)
Samples 1, 2, and 3 were also asked how much they liked or disliked Whites. Disliking of Whites was subtracted from disliking of Blacks and transformed to range from 0 to 1, with higher values indicating a stronger dislike of Black people. The value 0.5 represents a neutral midpoint.

Feeling Thermometer (Affect)
Only members of Sample 4 were asked, "Do you feel warm, cold, or neither warm nor cold toward Blacks?" Answers on a seven-point scale were transformed to range from 0 "extremely warm" to 1 "extremely cold." The value 0.5 represents the neutral midpoint "neither warm nor cold."

Differential Feeling Thermometer (Affect)
In Sample 4, feelings toward Whites were subtracted from feelings toward Blacks and transformed to range from 0 to 1. Higher values indicated more negative feelings toward Black people than toward White people. The value 0.5 represents a neutral midpoint.

Racial Stereotypes (Cognition)
In Sample 1, 2, and 3, participants were asked how well nine positive traits (e.g., friendly, determined to succeed,) and five negative traits (e.g., violent, boastful) described Black people. Sample 4 was asked about three positive and three negative traits. Answers on five-point scales were transformed to range from 0 to 1 and then averaged, with higher values indicating less positive stereotypical perceptions of Black people. The value 0.5 stands for "moderately well" and can thus not be considered neutral.

Differential Anti-Black Stereotypes
Ratings of Whites that were also made in Samples 2, 3, and 4 were subtracted from ratings of Blacks on the same item and then transformed to range from 0 to 1. The average of the ratings represents differential anti-Black stereotypes and the value of 0.5 represents the neutral midpoint of the scale.

New Racism

Symbolic Racism
Symbolic racism was measured in Samples 1, 2, and 3 with the eight-items validated SR2K scale (Henry & Sears, 2002). Answers were transformed to range from 0 to 1 and then averaged. Higher values indicated less positive attitudes toward Black people. The midpoint of the scale does not represent a neutral attitude.

Racial Resentment

Racial resentment was measured in Samples 1, 2, and 3 with six items taken from Kinder and Sanders (1996). Four of these items are also part of the SR2K scale. Answers were transformed to range from 0 to 1 and then averaged. Higher values indicated less positive attitudes toward Black people. The midpoint of the scale does not represent a neutral attitude. Study 4 contained only a four-items short version of this scale.

Implicit Measures

Affect Misattribution Procedure

All four samples completed the AMP to assess implicit anti-Black prejudice (Payne et al., 2005). Participants saw four series of twelve Chinese ideographs that were each preceded by a very fast flash of either an African-American or a White face. Negative implicit bias toward Black people was assessed by subtracting the proportion of pleasantly rated ideographs after White faces in each of the four series from the proportion of pleasantly rated ideographs after African American faces. The resulting four indicators were first recoded to range from 0 (meaning pro-Black affect) to 1 (meaning anti-Black affect) and then averaged. The value 0.5 represents equal implicit associations with Black and White faces.

Implicit Association Test

The Brief-IAT (Sriram & Greenwald, 2009) was administered only to Sample 4. Respondents saw pictures of Black faces, White faces, positive words ("love," "good," or "friend"), and negative words ("hate," "bad," or "enemy"). Three indicators, each based on twenty-eight trials (after four test trials), were generated and transformed to range from 0 (indicating extremely pro-Black attitudes) to 1 (indicating extremely anti-Black attitudes). Two hundred and thirty-five participants with poor data quality in the IAT (who answered more than 10 percent of the trials faster than 300 ms or more than 5 percent slower than 4,000 ms and had an error rate of more than 35 percent and an average response latency above 2,500 ms) were excluded from the analysis (Greenwald et al., 2009). The value 0.5 represents equal implicit associations with Black and White faces.

References

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, *72*(5), 1008–1032.

DeBell, M., Krosnick, J. A., & Lupia, A. (2010). *Methodology Report and User's Guide for the 2008–2009 ANES Panel Study*. Retrieved from https://electionstudies.org/wp-content/uploads/2009/03/anes_specialstudy_2008_2009panel_MethodologyRpt.pdf

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., et al. (2009). Understanding and using the implicit association test: iii. meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41.

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology*, *23*(2), 253–283.

Kinder, D. R., & Sanders, L. M. (1996). *Divided by Color: Racial Politics and Democratic Ideals*. Chicago, IL: University of Chicago Press.

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

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology*, *56*, 283–294.