Journal of Experimental Social Psychology 97 (2021) 104202

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Registered Report Stage 2: Full Article

Implicit attitudes matter for social judgments of others' preference, but do not make those judgments more or less accurate☆

Jeffrey Lees

John E. Walker Department of Economics, Clemson University, 309M Wilbur O. and Ann Powers Hall, Clemson, SC 29634, United States of America

ARTICLE INFO

Keywords:
Social judgment accuracy
Implicit attitudes
Registered report
Open data
Truth and bias model

ABSTRACT

Drawing from a large dataset of responses to implicit and explicit attitude measures and social judgments of others' preferences (N = 97,176) across 95 distinct attitude domains, this Registered Report utilized a componential analysis of judgment accuracy to examine whether implicit attitudes affected the accuracy of social judgment. I found evidence that judgments of the population's preferences were associated with the population's true implicit (but not explicit) attitudes, and that individuals projected their implicit attitudes in addition to the projection of explicit attitudes when judging the population's true preferences. However, I found no evidence that stronger or weaker implicit attitudes were uniquely associated with greater or less accuracy in judging the population's true preferences. These results provide generalizable evidence that implicit attitudes matter greatly for social judgment accuracy in distinct and nuanced ways.

1. Introduction

Successfully navigating the social world requires understanding the preferences and attitudes of others (Eyal, Steffel, & Epley, 2018; Hall & Goh, 2017; Kenny & Albright, 1987; Vazire & Mehl, 2008). While individuals are often accurate in their social judgments (Carlson, Vazire, & Furr, 2011; Jussim, 2017; Lewis, Hodges, Laurent, Srivastava, & Biancarosa, 2012), misperceiving the feelings and preferences of others has negative consequences for romantic relationships (Carlson, 2016; Human, Carlson, Geukes, Nestler, & Back, 2018), intergroup (Judd, Ryan, & Park, 1991; Li & Hong, 2001) and gender (Goh, Rad, & Hall, 2017; Rudman & Fetterolf, 2014) relations, politics (A. M. Enders & Armaly, 2018; Westfall, Van Boven, Chambers, & Judd, 2015), and one's reputation (Barranti, Carlson, & Furr, 2016; Solomon & Vazire, 2016). Yet despite a long scholarly tradition examining interpersonal and social judgment accuracy, along with the factors which affect accuracy (Biesanz, 2010; Funder, 1995; Kenny, 2004; West & Kenny, 2011), past work on the relationship between attitudes and social judgment has focused primarily on explicit attitudes in shaping judgment accuracy (Judd & Park, 1993; Krueger & Clement, 1994; Prentice & Miller, 1993; West, 2016). In doing so this work has overlooked a potential antecedent of social judgment accuracy: implicit attitudes.

Here I empirically tested whether implicit attitudes served a unique biasing role in social judgment accuracy, independent of the long documented effects of explicit attitudes in driving such judgments (Ames, 2004; Cronbach, 1955; Hoch, 1987). If implicit attitudes are related to but distinct from explicit attitudes (Lai & Banaji, 2020), it is likely they play a similarly biasing role in affecting social judgment accuracy. I measured social judgment accuracy as judgments of the average person's preferences within a specific attitude domain, for example "Does the average person prefer Traditional Values or Feminism?" This form of judgment accuracy is conceptually related to stereotypic (Cronbach, 1955) or normative accuracy (Biesanz, 2010; Furr, 2008) – the ability to accurately judge the true-mean of a given trait in the population (e.g., how extroverted is the average person) – and is distinct from interpersonal accuracy (e.g., how extroverted is Person X). Yet even at the interpersonal level, normative accuracy influences how we judge specific others by anchoring judgments on what people believe is typical (Biesanz & Human, 2010; Carlson, 2016; Krzyzaniak, Colman, Letzring, McDonald, & Biesanz, 2019; Wood & Furr, 2016). Social psychology has also long observed that perceptions of social norms and others' preferences strongly influence conformity in social behavior (Asch, 1956; Brewer & Miller, 1984; Cialdini & Goldstein, 2004; MacCoun, 2012), and can even shift individuals' own preferences (Berns, Capra, Moore, & Noussair, 2010; Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010; Zaki, Schirmer, & Mitchell, 2011).

Nonetheless, misperceiving the average preferences of others can lead to negative social outcomes. Such inaccurate social judgments

☆ This paper has been recommended for acceptance by Dr. Jarret Crawford.

E-mail address: jmlees@clemson.edu.

https://doi.org/10.1016/j.jesp.2021.104202
Received 23 January 2020; Received in revised form 10 July 2021; Accepted 16 July 2021

underly false-consensus (Krueger & Clement, 1994; Ross, Greene, & House, 1977) and pluralistic ignorance (Miller & Nelson, 2002; Prentice & Miller, 1993), where individuals over- and underestimate the prevalence of their preferences among others, respectively. Inaccurate social judgments contribute to false polarization (Monin & Norton, 2003; Westfall et al., 2015) and overly negative beliefs about what others believe, which can serve to exacerbate intergroup conflict (Lees & Cikara, 2020). Inaccurate social judgments can also contribute to the continuation of oppressive social policies, for example in Saudi Arabia where the vast majority of married men privately support female labor force participation, but inaccurately believe that many other men do not (Bursztyn, Gonzalez, & Yanagizawa-Drott, 2018). These findings point to a growing need to better understand the cognitive antecedents of inaccurate social judgments regarding the average preferences of others. If implicit attitudes partially contribute to such inaccuracies, as this Registered Report hypothesized, then understanding the nature of that relationship may provide new avenues for studying and correcting inaccurate social judgments.

Implicit attitudes–automatic and unconscious associations–are distinct from but related to the explicit attitudes which individuals consciously possess (Kurdi, Gershman, & Banaji, 2019; Kurdi, Mann, Charlesworth, & Banaji, 2019; Nosek & Hansen, 2008) and are measured through methods such as the Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998), sequential priming (Cameron, Brown-Iannuzzi, & Payne, 2012), and the Affect Misattribution Procedure (Payne, Cheng, Govorun, & Stewart, 2005). Implicit attitudes are often studied in the context of attitudes toward marginalized groups (e.g., African Americans), where implicit biases against such groups and in favor of majority-group (e.g., White Americans) are consistently observed (Charlesworth & Banaji, 2019; Jost et al., 2009; Lai et al., 2014). Yet implicit attitudes are no less pronounced or impactful in attitude domains outside of marginalized groups, including the domains of politics (Arcuri, Castelli, Galdi, Zogmaister, & Amadori, 2008), moral judgments (Reynolds, Leavitt, & DeCelles, 2010), sports (Brand, Heck, & Ziegler, 2014), and even food (Hofmann & Friese, 2008).

Scholars have begun to examine how implicit attitudes are influenced and can be changed. Many interventions designed to reduce implicit racial biases have failed to produce lasting change in implicit attitudes (Lai et al., 2014), yet recent work suggests that implicit attitudes do shift over time (Charlesworth & Banaji, 2019), are sensitive to immediate environmental and social cues (Mann, Kurdi, & Banaji, 2019; Vuletich & Payne, 2019), and can be durably changed when they relate to novel stimuli (Cone, Mann, & Ferguson, 2017). If implicit attitudes have a unique effect on social judgment accuracy, which this Registered Report sought to test, then interventions that can reliably shift implicit attitudes may also be effective at increasing social judgment accuracy and attenuating the negative outcomes associated with inaccurate social judgments regarding the preferences of others.

By expanding the scope of implicit attitudes to include dozens of varied attitude domains, these findings related to social judgment accuracy are generalizable beyond the study of intergroup relations and speak to phenomena that impact myriad domains of human cognition and behavior. To do so I drew from an existing dataset of ~97,000 IAT responses (Nosek & Hansen, 2008) administered over 95 distinct attitude domains. These domains varied substantially and included attitudes about social groups (e.g. "African Americans - European Americans" and "Muslims - Jews"), prominent individuals (e.g. "Princess Diana - Mother Teresa" and "Tom Cruise - Denzel Washington"), commercial institutions (e.g. "McDonald's - Burger King" and "Apple - Microsoft"), consumption and aesthetic preferences (e.g. "Tea - Coffee" and "Dramas - Comedies"), philosophical ideas (e.g. "Rebellious - Conforming" and "Innocence - Wisdom"), and political preferences (e.g. "Republicans - Democrats" and "Gun Rights - Gun Control"), among many others.

Critically, this dataset included measurements of explicit attitudes (self-reported attitudes toward the attitude domain), implicit attitudes (IAT D-scores toward the attitude domain), and individuals' judgments of the population's preference within the attitude domain. From this dataset I calculated the population's true implicit and explicit preferences using a weighted-averaging raking procedure. Despite its size, the survey sample was not perfectly representative of the general US population. As such, I calculated weights for all responses based on population distributions of the demographic variables age, race, education, sex, and political orientation. These weights ensured that when I averaged participants' implicit and explicit attitudes within the attitude domain as a measurement of the population's true preferences, the means were representative of the general population rather than the specific, unrepresentative sample.

With the following pieces of information across 95 discrete attitude domains: individuals' explicit and implicit attitudes, individuals' judgments of the population's preferences, and the population's true explicit and implicit preferences, I was able to robustly examine the extent to which judgments were attracted to the true values and the extent to which they were biased by individuals' own attitudes. To guide this analysis I utilized the Truth & Bias Model of Judgment (West & Kenny, 2011) ("T&B Model"), a conceptual and statistical framework for understanding and disentangling the components of human judgment, specifically the ways in which judgment is pulled by "truth forces", here the population's true implicit and explicit preferences, and pulled by "bias forces", here individuals' implicit and explicit attitudes (i.e., projection/assumed similarity). Because I drew from a diverse and representative set of 95 attitude domains of judgment, attitude domains were modeled as a random effect.

Using the T&B Model I tested seven hypotheses (four confirmatory, one exploratory, and two competing) regarding the relationships between judgment, accuracy, and implicit attitudes (see Table 1a for a summary of all hypotheses). Hypothesis 1a predicted that participants' judgments of the population's preferences would be positively associated with the population's true explicit preference, which would be evidence for normative accuracy (Biesanz, 2010; Furr, 2008) in judgments of mean population preferences. Hypothesis 1b was exploratory and predicted that there would be directional biases in participants' judgments within specific attitude domains. Directional biases in the T&B Model are equivalent to differential elevation accuracy (Cronbach, 1955), the correspondence between the validation measure (here the population's true explicit preference within a domain) and the main judgment within a domain. In other words, directional bias is mean over- or underestimation of a single value, relative to the true value. Hypothesis 1b was exploratory because the breadth of attitude domains examined made it impractical to provide a theoretical framework for predicting what attitude domains would exhibit underestimation, overestimation, and no bias. As such, the goal of Hypothesis 1b was to provide a descriptive account of directional biases in judgment.

Hypothesis 2 predicted that participants' judgments of the population's preferences would be positively associated with the population's true implicit preferences, which would both be evidence for accuracy in judgment and suggest judgments are discriminating between others' implicit and explicit attitudes. Growing evidence suggests that implicit attitude measures are at their most reliable when aggregated (Hehman, Calanchini, Flake, & Leitner, 2019; Schimmack, 2021), as they were here with the measure of the population's true implicit preferences. It was therefore reasonable to hypothesize that judgments of the population's preferences would track with the population's true implicit preferences, along with the population's true explicit preferences.

Hypotheses 3a predicted that participants' judgments of the population's preferences would be positively associated with their own explicit attitudes, and Hypothesis 3b predicted that participants' judgments of the population's preferences would be positively associated with their own implicit attitudes. Such findings would be evidence for the projection/assumed similarity (Ames, 2004; Cronbach, 1955; Holtz & Norman, 1985; Robbins & Krueger, 2005) in relation to both explicit

Table 1a
Summary of Hypotheses.

Hypothesis	Model	Prediction	Confirmatory or Exploratory	Level of Analysis
Hypothesis 1a	Model 1	*Judgment Accuracy Hypothesis*: Linear relationship between estimates of and the population's true explicit preferences	Confirmatory	Level 2, between domain association
Hypothesis 1b	Model 1	*Directional Bias Hypothesis*: Mean-level over and under estimation of the population's true explicit preferences within domain	Exploratory (no a priori prediction regarding domains exhibiting directional bias, or in what direction)	Level 1, within domain point estimate
Hypothesis 2	Model 2	*Implicit Detection Hypothesis*: Linear relationship between estimates of the population's preferences and the population's true implicit preferences	Confirmatory	Level 2, between domain association
Hypothesis 3a	Model 3	*Explicit Attitude Projection Hypothesis*: Linear relationship between participant's estimates of the population's preferences and their own explicit attitudes	Confirmatory	Level 1, between participant association
Hypothesis 3b	Model 3	*Implicit Attitude Projection Hypothesis*: Linear relationship between participant's estimates of population preferences and their own implicit attitudes	Confirmatory	Level 1, between participant association
Hypothesis 4	Model 4	*Implicit Attitudes Increase Accuracy Hypothesis:* Judgment accuracy will increase as the population's implicit attitudes trend toward the population's true explicit preference	Competing (no a priori prediction as to whether H4 will be supported over H5)	Level 1 x Level 2 interaction
Hypothesis 5	Model 5	*Implicit Attitude-Extremity Increases Accuracy Hypothesis:* Judgment accuracy will increase as the population's implicit attitudes trend away from the population's true explicit preferences (i.e., exhibit negative curvature)	Competing (no a priori prediction as to whether H5 will be supported over H4)	Level 1 x Level 2 interaction

and implicit attitudes. Projection/assumed similarity refers to the common tendency of individuals' judgments of others' traits and preferences to correlate with their own traits and preferences, and Hypotheses 3a and 3b predicted projection of individuals' own explicit and implicit attitudes in predicting the population's true preferences. In the T&B Model projection/assumed similarity were conceptualized "bias forces" on judgment.

Yet just because judgments are "biased" by projection/assumed similarity does not necessarily mean those judgments will be less accurate. If one's attitudes happen to be closely aligned with the population's true preferences then such projection/assumed similarity may reflect actual similarity and therefore increase accuracy. This possibility, along with other potential ways in which one's own attitudes affected judgment accuracy, were tested in the competing Hypotheses 4 and 5.

Hypotheses 4 and 5 were competing hypotheses related to the way in which individuals' implicit attitudes affect judgment accuracy, with both hypotheses making predictions that implicit attitudes would increase judgment accuracy under some circumstances. What differed was that Hypothesis 4 assumed that the direction of one's implicit attitudes relative to the population's true preference mattered, whereas Hypothesis 5 predicted that direction would not matter and that the extremity of one's implicit attitudes was what affected accuracy.

Hypothesis 4 predicted that as implicit attitudes trend above the population's true mean preferences, judgments of those preferences would become more accurate because assumed similarity would, on average, reflect actual similarity. Hypothesis 4 therefore also predicted that accuracy would decrease as implicit attitudes trended in the opposite direction of the population's true preferences. The potential for projection/assumed similarity to increase judgment accuracy has long been observed (Cronbach, 1955; Hoch, 1987; Murray, Holmes, Bellavia, Griffin, & Dolderman, 2002; Neyer, Banse, & Asendorpf, 1999), and as such Hypothesis 4 reflected this possibility in the domain of implicit attitudes.

Hypothesis 5 also predicted that implicit attitudes will increase judgment accuracy, but in contrast to Hypothesis 4 it predicted that the extremity of implicit attitudes would increase accuracy. Hypothesis 5 was based on the theory that implicit attitudes reflect, in part, knowledge rather than unconscious appraisals (Banaji, Nosek, & Greenwald, 2004; Dambrun & Guimond, 2004; Hahn, Judd, Hirsh, & Blair, 2014; Nosek & Hansen, 2008). This meant that Hypothesis 5 predicted that those with strong implicit attitudes in *either direction* would in fact be the most knowledgeable regarding the attitude domain at hand and therefore the most accurate in judging the population's true preference toward that domain. This possibility is supported by work on interpersonal perception which finds that those with more extreme political attitudes are more accurate in judging the political attitudes of others (Ivanov, Muller, Delmas, & Wänke, 2018). Hypothesis 5 was tested using polynomial regression to examine whether there was a negative curvilinear relationship between accuracy and implicit attitudes.

Hypotheses 4 and 5 were competing hypotheses reflecting different theories as to the mechanisms driving the judgment accuracy and implicit attitudes relationship, yet both hypothesized that implicit attitudes would positively moderate the accuracy of social judgment. Hypothesis 4 was based on an account of projection/assumed similarity between individuals and their judgments of the population's true preferences, whereas Hypothesis 5 was based on the theory that implicit attitudes reflect knowledge that would itself increase individuals' judgment accuracy. I made no a priori prediction as to which would be supported by the analyses herein.

Table 1a below summarizes all the tested hypotheses. The T&B Model allowed me to test and explore these hypotheses while disentangling the unique role implicit and explicit attitudes each played in driving judgment and affecting accuracy. Table 1a also highlights the multilevel structure of the models. As the measures of true explicit and implicit preferences were necessarily averaged within attitude-domain, they were Level 2 variables within the hierarchical models. Accuracy therefore was constituted as a Level 2 association in the models, and as such was an attribute of the population, *not* an attribute of individuals in these analyses. And individuals' own explicit and implicit attitudes were measured at the individual level (Level 1), so examining the interaction between them and the population's true explicit and implicit preferences (Level 2 variables) constituted cross-level analyses.

All analyses for this Registered Report are available on the Open Science Framework: https://osf.io/k5v8x/

2. Method

2.1. Anticipated sample

The full AIID dataset (N ≃ 200,000) was collected on the Project Implicit website (https://implicit.harvard.edu/implicit/) between 2004 and 2007 and served as the basis for Study 7 in Nosek and Hanen's 2008 paper (Nosek & Hansen, 2008). This dataset was being intentionally withheld as part of a call for Registered Reports to use the data. It was only released to the author upon acceptance of a Stage 1 Registered Report. As part of the call for Registered Reports using this dataset, a small portion of the overall dataset (15% stratified subset) was released to allow researchers to set up analysis scripts and conduct power

analyses. See the Pilot Data section of the Stage 1 manuscript for information regarding this subset of data (link: https://osf.io/zvucw/).

In the original collection of the data, participants registered on the Project Implicit website and were randomly assigned to one of 95 different attitude domains (e.g. "Democrats" vs. "Republicans"), and randomly assigned to assess the domain in terms of their implicit identification (self/other) or implicit attitudes (positive/negative). Afterwards participants completed self-report attitude items (including their explicit attitudes) and individual difference measures.

To be included in the analyses for this Registered Report responses must have met three criteria. First, participants must have been assigned to the implicit attitude measures, as I would not be analyzing the implicit identification data. Second, participants must have provided their judgments of others' preferences, as this was the dependent variable across all models. For all independent variables I utilized pairwise deletion for incomplete responses. Third, participants must not have failed one or more of the standard IAT exclusion criteria that were used to screen participants who were carelessly completing the IAT tasks. See the supplementary materials for specifics of these IAT exclusion criteria. I was provided with a masked version (i.e. all responses were masked as "1," but missing data were observable) of the original dataset so that I could examine how many participants would meet these exclusion criteria. I anticipated that of the ~200,000 completed responses in the original dataset approximately ~97,000 would meet these criteria.

Of the ~97,000 responses in the full dataset I planned to analyze, I anticipated those responses to come from approximately ~68,000 unique participants, meaning that a plurality of participants took the IAT on Project Implicit more than once. Of the ~68,000 unique participants, I anticipated ~52,000 would have taken it once, ~10,000 would have taken it twice, ~3000 would have taken it three times, and ~ 3000 would have taken it four or more times. Repeated responses raised the potential for within-participant effects, which I discuss in the analysis section.

2.2. Deriving population's true preferences

Of primary interest in the analyses was the relationships between participants' judgments of the population's preferences within each attitude domain, analyzed as a dependent variable, and the true mean of the population's implicit and explicit attitudes within each attitude domain, analyzed as independent variables. Examining the linear relationship between these values and participants' estimates of these values allowed for an examination of judgment accuracy, i.e. whether judgments tracked with both the true mean implicit and explicit preferences of the general population.

To derive the population's true implicit and explicit preferences I utilized weighted-means within domain, across the full dataset. This method has been used in previous IAT research (Charlesworth & Banaji, 2019). The population's true explicit attitudes were weighted-means within the attitude domain of participants' responses to the self-report explicit attitude items. The population's true implicit attitudes were weighted-means within the attitude domain of participants' D-scores on the implicit attitude measures. While the full dataset was large, it was not perfectly representative of the general US population, as the sample skewed younger, more educated and wealthy, and more politically liberal than the general US population. As such I utilized a raking function to generate response-level weights based on 2005 census data and General Social Survey (Smith, Davern, Freese, & Stephen, 2019) responses along the following characteristics: sex, age, race, education, and political orientation (see supplimentary materials for exact weights).

The raking function assigned each response a weight corresponding to its demographic representativeness relative to the demographic frequencies in the dataset. For example, if liberal white men were overrepresented in the full dataset, their data were down-weighted (i.e. given a weight less than one) for the purposes of deriving the mean of the population's explicit and implicit preferences within each attitude domain. Note that weights were generated for each response, not each individual. As a plurality of participants responded to the IAT multiple times, setting the raking function to generate weights by individuals rather than by their responses would have led to those who took the IAT multiple times being overweighted relative to the majority of participants who took it once. And since participants who took the survey more than once were assigned a different attitude domain to judge, there was no possibility that a single participant's multiple responses were included in any one averaged value for the population's true preferences.

Participants must have responded to all the relevant demographic questions in order to be included in the generation of weights and aggregating of the population's true preferences. Because all participants responded to the explicit attitude measure, but not all participants responded to the implicit attitude measure (some instead responded to an implicit identification measure), two weights were generated for each response in the dataset. One weight reflected the full dataset of responses to the explicit attitude measure and was used in deriving the mean of the population's explicit preferences, and the other reflected the subset of data for responses to the implicit attitude measure and was used in deriving the mean of the population's implicit preferences. These procedures ensured that the derived population's true preferences, both for implicit and explicit preferences, reflectived the general US population's preferences.

2.3. Anticipated data processing

To allow for clear comparisons of individuals' responses across attitude domains, I ensured that all domains were ordered such that a positive value reflected an attitude directionally aligned with the population's true explicit preferences. For example, take one attitude domain from the data: "Dogs - Cats." When participants rated their explicit attitudes toward this category they did so on an −3 to +3 scale, with −3 representing a strong preference for dogs over cats, and + 3 representing a strong preference for cats over dogs. Imagine that the population's true explicit attitude was −1.5 for dogs vs. cats, indicating a relative preference for dogs. In this case, because the population's true preference was negative I flipped the "Dogs - Cats" category, and all corresponding individual responses related to that attitude domain, so that the presentation of the domain became "Cats - Dogs". Therefore, the population's true explicit preference was transformed from −1.5 to +1.5 and all individual-level values related to the domain (e.g., explicit attitudes, D-scores, estimates of the population's preference) also had their signs flipped.

These transformations were necessary to allow for several of the analyses. Hypotheses 4 and 5 examined whether implicit and explicit attitudes affected the accuracy of judgments of the population's implicit and explicit preferences. By aligning all the attitude domains such that all positive values indicated attitudes that trended directionally toward the population's true explicit preferences, I was able to examine whether both the extremity and direction of one's own attitudes affected judgment accuracy. For example, if I had found that individuals with higher D-scores were more accurate, I could interpret that to mean that individuals whose implicit attitudes were above the population's true preferences were more accurate, and conversely that those whose implicit attitudes directionally diverged from the population's true preference were less accurate. Such an inference would have been impossible if the data were not transformed as such.

2.4. Measurements

There were seven total variables that were used in testing Hypotheses 1–5: the dependent variable of judgments of the population's true preferences, four fixed independent variables of individual's explicit and implicit attitudes, along with the true population mean explicit and

implicit attitudes, and two random variables of attitude domain and participant-effects. The measurement of an individual's implicit attitudes was their IAT D-score (Greenwald, Nosek, & Banaji, 2003), a standardized measure of participants' relative preference for one of the presented categories over the other, bound between −2 and + 2, and calculated by dividing the mean difference in reaction times across trial blocks by the overall standard deviation in reaction times. A positive D-score represented a relative evaluative preference for the right-sided attitude domain category, relative to the left, presented herein. For example, if a participant's D-score was positive 0.25 in the domain "Cats vs. Dogs," they had a relative implicit preference for dogs over cats. The measure of the population's true implicit preferences was simply the weighted-mean average of the D-scores for each of the 95 attitude domains. The attitude domain of judgment was analyzed as a categorical random effect.

Individuals' explicit attitudes were measured by a single 7-point Likert item, numbered −3 to +3 with labels "Strongly Prefer X over Y" to "Strongly Prefer Y over X," where "X" and "Y" were the contrasting categories within an attitude domain. The measure of the population's true explicit preferences was simply the weighted-mean average of this item for each of the 95 attitude domains.

Individuals' judgments of the population's preferences, the dependent variable, was measured by taking the mean of three 7-point Likert items which asked participants to estimate the preferences of other people. Those three items were "Does the average person prefer X or Y", "Does the culture you live in prefer X or Y", and "Do most people prefer X or Y". The reliability of these three items in the pilot data ($N = 31{,}262$) was $\alpha = 0.86$. The original paper that collected these items (Nosek & Hansen, 2008) collected a random subset of four items across these three and another three items. As such, all participants responded to at least one of these three items. In the pilot data ~20% received all three, ~60% received two of three, ~20% received one of three. These three items were also anchored on "Strongly Prefer X over Y" to "Strongly Prefer Y over X," meaning that the judgment items and validation items (the true values) utilized the same scales on the same anchors, allowing confidence in interpreting (in)congruence between them as (in) accuracy.

2.5. Planned analyses

All hypotheses were tested using linear mixed-effects modeling and were based on the Truth & Bias Model of Judgment (West & Kenny, 2011) ("T&B Model"), a conceptual framework for understanding and disentangling the components of human judgment, specifically the ways in which judgment is pulled by "truth forces" and "bias forces." For my purposes, the judgment of interest were participants' estimates of the population's true preferences within a given attitude domain, and were analyzed as the dependent variable. The two "truth values" were the population's true mean explicit and implicit attitudes within the given attitude domain. The "truth forces" were the extent to which judgments were attracted toward these two values. Conversely, the two "bias values" were individuals' explicit and implicit attitudes with the given attitude domain, and the "bias forces" were the extent to which judgments were attracted toward these two values rather than the true values (i.e., projection/assumed similarity). In addition to the main effects of the truth and bias values in driving judgment, I also examined the interaction between the truth and bias values. The 95 attitude domains were modeled as random intercepts with correlated random slopes for the individual-level bias values (individuals' implicit and explicit attitudes). As the truth values were static within domain (i.e., Level 2 variables), there was no slope variance to examine, hence no random slopes for the population's true preferences. Participant-effects were also modeled with random intercepts, as many participants took the IAT more than once in the dataset.

This framework allowed for a parsimonious test of the stated hypotheses through five linear mixed-models, each building on the last. All analyses were conducted using the statistical software R (v. 4.0.3) and the R package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017) to model the regressions and calculate *p*-values using Welch-Satterthwaite approximation.

Model 1: Model 1 tested Hypotheses 1a and 1b. In Model 1, I regressed the dependent variable of judgment onto the single independent variable of the population's true explicit preferences. Hypothesis 1a predicted that there would be a positive linear relationship between judgments of the population's preference, within a given attitude domain, and the population's true explicit preference for that domain. Such a linear relationship would constitute evidence of judgment accuracy in the population, as the true values were Level 2 variables. Hypothesis 1b explored whether there was directional bias in judgments of the population's preferences, and whether directional bias would vary across attitude domains. Directional bias was modeled as the intercept value in Model 1, and variance in directional bias was constituted as the random intercept estimates for each attitude domain. For Model 1, J_{di} was participant i's judgment of the population's true preference within domain d. Intercept b_0 provided the overall directional bias, and modeled directional bias within each domain as random intercept D_0 in domain d. Random intercept P_{0i} modeled within-participant effects for each participant i. TE_d was the truth value of the population's true explicit preference within domain d, and t_1 was the truth force of the population's true explicit preferences on judgment. e_{di} was random error across individuals and domains.

$$J_{di} = b_0 + D_{0d} + P_{0i} + t_1 TE_d + e_{di}$$

Model 2: Model 2 tested Hypothesis 2. In Model 2, I added a second independent variable of the population's true implicit preferences to Model 1. Hypothesis 2 predicted that there would be a positive linear relationship between judgments of the population's preferences, within a given domain, and the population's true implicit preference in that domain. Such a linear relationship would both constitute evidence of judgment accuracy in the population, and evidence that judgments were discriminating between implicit attitudes at the population level and the population's explicit attitudes. In the T&B framework this is conceptualized as having two truth forces, both of which are independently associated with judgment. Analytically Hypothesis 2 was examined by entering both the population's true explicit and implicit preferences as independent predictors of participants' judgments in the model. In Model 2, TI_d was the truth value of the population's true implicit preference within domain d, and t_2 was the truth force of TI_d on judgment.

$$J_{di} = b_0 + D_{0d} + P_{0i} + t_1 TE_d + t_2 TI_d + e_{di}$$

Model 3: Model 3 tested Hypotheses 3a and 3b. In Model 3, I added two new independent variables, individuals' own explicit and implicit attitudes, to Model 2. Hypothesis 3a predicted that there would be a significant positive linear relationship between judgments of the population's preference, within a given domain, and individuals' own explicit attitudes within that domain. Hypothesis 3b predicted that there would be a significant positive linear relationship between judgments of the population's preference, within a given domain, and individuals' own implicit attitudes within that domain. Plainly, I predicted that individuals' judgments of the population's preferences would be associated with their personal implicit and explicit attitudes, which would constitute evidence for projection/assumed similarity. In the T&B framework, individuals' own implicit and explicit attitudes were bias values, and Hypotheses 3a and 3b predicted that these bias values would have significant bias force. Analytically, individuals' implicit and explicit attitudes were entered as predictors into Model 2, along with adding correlated random slopes for implicit and explicit attitudes within attitude domains. From Model 3, I added the bias values of explicit attitudes *BE* and implicit attitudes *BI*, both for individual i and domain d. The bias force of individuals' explicit attitudes on judgment was b_1, whereas b_2 was the bias force of individuals' implicit attitudes on judgment. Lastly, D_{1d} and D_{2d} modeled random slopes for individuals'

explicit and implicit attitudes within domain *d*, respectively.

$$J_{di} = b_0 + D_{0d} + P_{0i} + t_1 TE_d + t_2 TI_d + (b_1 + D_{1d})BE_{di} + (b_2 + D_{2d})BI_{di} + e_{di}$$

Models 4: Models 4 examined Hypothesis 4. Model 4 added four two-way interactions to Model 3. I interacted the population's true explicit preferences with individuals' explicit and implicit attitudes separately, and I interacted the population's true implicit preferences with individuals' explicit and implicit attitudes separately. Hypotheses 4 suggested that implicit attitudes affect judgment accuracy. A significant positive interaction would indicate that higher implicit or explicit attitudes in the direction of the population's true preferences were associated with an increase in the linear relationship between judgments and the true values (i.e., greater accuracy). I included the interactions between participants' explicit attitudes and the population's true implicit and explicit preferences to ensure any interaction of implicit attitudes and the population's preferences was uniquely the result of individuals' implicit attitudes.

I modeled Hypotheses 4 as four two-way interactions between the bias values and true values. Of primary interest to the hypotheses were x_3 and x_4. x_3 modeled the moderating force of implicit attitudes BI_{di} on the truth force (i.e. accuracy) of the population's true explicit preferences TE_d. x_4 modeled the moderating force of implicit attitudes BI_{di} on the truth force of the population's true implicit preferences TI_d. I also modeled the moderating role of explicit attitudes on the population's true explicit and implicit attitudes as x_1 and x_2, respectively, as control variables to ensure that the observed estimates of x_3 and x_4 were unique to individuals' implicit attitudes.

$$J_{di} = b_0 + D_{0d} + P_{0i} + t_1 TE_d + t_2 TI_d + (b_1 + D_{1d})BE_{di} + (b_2 + D_{2d})BI_{di} + x_1 BE_{di} TE_d + x_2 BE_{di} TI_d + x_3 BI_{di} TE_d + x_4 BI_{di} TI_d + e_{di}$$

Model 5: Model 5 examined Hypothesis 5, which predicted that irrespective of direction the extremity of implicit attitudes would moderate the accuracy of judgment. In other words, Hypothesis 5 predicted a negative curvilinear relationship between accuracy and implicit attitudes, where the lowest level of accuracy would be observed when implicit attitudes were equal to the population mean, and accuracy would increase as implicit attitudes deviated in both directions from the mean.

Hypotheses 4 and 5 competed in their predictions of whether or not the direction of individuals' implicit attitudes mattered for affecting judgment accuracy. Hypothesis 4 predicted that directionality of implicit attitudes would matter, such that the primary difference in accuracy would be between individuals with strong implicit attitudes in line with the population's true preferences (who would be more accurate) and individuals with strong implicit attitudes which diverged from the population's true preferences (who would be less accurate). Hypothesis 5, conversely, explored whether only extremity of implicit attitudes matters, such that the primary difference in accuracy would be between those with typical implicit attitudes (i.e. a D-score equal to the population mean) and those with implicit attitudes that were strongly atypical (who would be more accurate).

Analytically, Hypothesis 5 was examined using polynomial regression estimates, a common method for examining judgment accuracy (Barranti, Carlson, & Côté, 2017; Edwards & Parry, 1993; Humberg, Nestler, & Back, 2019). Model 5 added two additional predictors to Model 4: the squared value of the population's true explicit attitudes TE_d^2, and the squared value of individuals' implicit attitudes BI_{di}^2. Drawing from techniques developed for response surface analysis (Barranti et al., 2017), Hypothesis 5 would be considered confirmed over Hypothesis 4 if (1) the interaction x_3 between the population's true explicit attitudes and individuals' implicit attitudes were positive and significant, (2) either of the quadratic values s_1 or s_2 were negative and significant, and (3) Model 5 significantly improved model fit over Model 4 (see section below on model comparisons).

$$J_{di} = b_0 + D_{0d} + P_{0i} + t_1 TE_d + t_2 TI_d + (b_1 + D_{1d})BE_{di} + (b_2 + D_{2d})BI_{di} + x_1 BE_{di} TE_d + x_2 BE_{di} TI_d + x_3 BI_{di} TE_d + x_4 BI_{di} TI_d + s_1 TE_d^2 + s_2 BI_{di}^2 + e_{di}$$

Centering: All variables in the multilevel models were either grand mean centered or group mean centered within attitude-domain (i.e., within-cluster; Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995). The population's true explicit preferences TE_d and true implicit preferences TI_d were grand mean centered. Judgments of the population's preferences J_{di} were centered on the grand mean of TE_d, in other words true grand-mean centered. This technique meant that the intercept estimate of Model 1 reflected the average difference of J_{di} and TE_d and provided a direct test of the directional mean biases proposed in Hypothesis 2. Individual's explicit attitudes BE_{di} and implicit attitudes BI_{di} were each centered within-domain on the population's true explicit preferences TE_d and true implicit preference TI_d respectively (i.e., on the raw value of TE_d and TI_d). This centering method addressed two potential issues. First, it orthogonalized the Level 1 predictors from the Level 2 predictors, which is ideal when modeling interactions across model levels as was done in Models 4 and 5 (C. K. Enders & Tofighi, 2007). Second, it addressed potential issues with interpreting centered variables which, in raw form, were bipolar. Take, for example, the D-scores used to measure an individual's implicit attitudes (TI_{di}). If for domain-A the mean D-score were − 1, and in domain-B it were + 1, then centering on the within-cluster sample mean would lead to a situation where for domain-B going from 0 to 1 in the regression meant *increasing* implicit bias, but in domain-A it meant *decreasing* implicit bias. This would present serious interpretive issues if such values were entered into statistical interactions. However, by centering on the true population means derived herein via the weighting measure, which were crucially transformed such that all explicit preference means are positive (see "Anticipated Data Processing" section), it guaranteed that the within-cluster centered mean reflected a positive raw value, meaning that an increase in the centered variable was interpreted as an increase in implicit/explicit attitude strength.

Model Comparisons and R^2: Models 1–5 were analyzed using maximum-likelihood estimation. While restricted maximum-likelihood (REML) estimation provides less biased estimates of the random effects, REML estimation precluded the usage of likelihood-ratio tests to compare model fits across Model 1–5, whereas maximum-likelihood estimation allowed for such comparative statistics. I used likelihood-ratio tests to examine if adding predictors to the models provided significantly better model fit, and these tests required that the models be embedded within one another. As such, I compared Model 1 to Model 2, Model 3 to Model 4, and Model 4 to Model 5. I also calculated R^2 for each model using "Nakagawa's R^2" (Nakagawa, Johnson, & Schielzeth, 2017), which calculated both the variance explained by just the fixed effects and the variance explained by both fixed and random effects. The relative R^2 values for each model, along with the likelihood-ratio significance tests, were used to make inferences about whether or not adding certain variables to the models explained meaningful variance. Given the anticipated sample size (~97,000), it was possible that I would find statistically significant main effects (or interactions) that were nonetheless so small in effect size they would not be ecologically meaningful. The relative R^2 values and model comparisons assisted in interpreting the meaningfulness of such results.

Power Analysis: To ensure that the anticipated sample of ~97,000 was sufficient to detect small effect sizes I focused on the effects across the hypotheses that would be the most difficult to detect: the interactions in Models 4 and 5. First, I used the pilot data (provided as part of this Registered Report) to estimate effect sizes in the final sample. Of primary interest was x_3 and x_4 in Models 4 and 5, i.e. whether implicit attitudes affected the accuracy of judgments. In the pilot data the absolute, unstandardized effect sizes for estimates of x_3 and x_4 ranged from 0.037 to 0.072 (none were statistically significant in the pilot data). Because the sample size of the full dataset was fixed, I performed

prospective power analyses to determine the smallest effect size detectable at 80% power with a final sample of ~97,000. I found that an unstandardized x_3 value of 0.09 in Model 4 could be detected with 80.80% power [95% CI = 77.07, 84.16] in the full dataset, based on 500 Monte Carlo simulations using the simr R package (Green & MacLeod, 2016). Note that $x_3 = 0.09$ is an unstandardized coefficient, and while it would be inappropriate to report standardized coefficients in the final results due to the fact that in analyses of accuracy the raw data have inherent meaning (Biesanz, 2010; West & Kenny, 2011), here I transformed this coefficient for the sole purpose of making the effect size interpretable in the context of this power analysis. Standardized, the coefficient is $\beta = 0.009$, meaning that these analyses are highly powered to detect the moderating effect of implicit attitudes on judgment accuracy at a magnitude of approximately one-twentieth of a standard deviation. I argue that any true effect of implicit attitudes on judgment accuracy that is smaller in magnitude than this is so small as to be scientifically meaningless. Therefore, the anticipated sample of ~97,000 was highly powered to detect the effects predicted by Hypotheses 1–5.

Multiple Responses from Participants: Across the anticipated sample of ~97,000 responses ~68,000 were anticipated to be unique participants, with ~52,000 having one response, ~10,000 having two responses, and the remaining ~6000 having responded three or more times. Given that relatively few participants responded more than once and that the vast majority of those who did only responded twice or three times, there was potentially little to no variance to be explained by within-participant effects. Participant random intercepts were modeled across all models, and only excluded from any given model if model convergence could not be obtained due to lack of variance in said random intercept estimates. The procedure for this determination is below.

Handling Model Non-Convergence: I did not anticipate serious issues with model convergence, but analyses of the pilot data suggested that modeling of participant random intercepts may lead to singular models. If this issue arose I utilized the alternative optimization options available through the lme4 (Bates, Maechler, Bolker, & Walker, 2015) R package and outlier removal procedures prior to removing variables from the models to reduce complexity, as removing random slopes for the sake of model parsimony can potentially bias model outcomes (Barr, Levy, Scheepers, & Tily, 2013). If any model failed to converge or was singular when using the base optimizer for linear mixed-effects models, I used the allFit() function to examine whether any of the alternative optimizers led to models that neither failed to converge nor were singular. If one of the optimizers converged and was not singular then I used that optimizer. If more than one optimizer converged then I used the optimizer with the higher log-likelihood, as this indicated better fit. If the log-likelihoods were the same, I privileged linear over non-linear optimization formulas. If I was still left with multiple optimizers, then I simply privileged the one ordered first in the function.

If no models converged/lacked singularity I examined whether this was due to the inclusion of participant random intercepts. If (1) every optimizer produced singular models, and (2) the variance explained by participant random intercepts was zero, then I removed the participant random intercepts from the model. If these criteria were not met, for example if some optimizers were not singular but still would not converge, then I utilized a Bonferroni outlier test to identify up to ten potential outliers. The test identified mean-shift outliers by examining the distribution of studentized residuals. Bonferroni-adjusted p-values would be calculated for all studentized residuals, and a p-value <0.05 would indicate the response could be considered an outlier. I would remove up to ten such outliers, and then begin again the process of examining all potential model optimizers. If this process eventually led to a point where no models converged/lacked singularity and no outliers were identifiable, I would then begin removing random slopes or intercepts to simplify the random structure. I would first remove the random variable which had a lower standard deviation (in the first model that failed to converge), then repeat all the steps above before removing another random variable.

3. Results

3.1. Data availability

All data and analyses conducted as part of the Stage 2 Registered Report will be publicly available on the Open Science Framework at the following link, once the full dataset is released publicly by the AIID team: https://osf.io/95qcy/. The preregistered analyses, Stage 1 manuscript, and pilot data can also be found on the OSF: https://osf.io/r7j4h/. The preregistration can be found here: https://osf.io/qv8cy If you have any questions regarding the analyses, please do not hesitate to contact the author.

An overview of results for the tests of Hypotheses 1–5 can be found in Table 1b. In summary, I found evidence that implicit attitudes played an important role in social judgment accuracy. In the T&B framework, implicit attitudes exhibited significant truth force and bias force on judgments of the population's true preferences, above the truth force and bias force of explicit attitudes. However, I found no evidence that an increase in implicit attitudes was associated with an increase in social judgment accuracy (or vice versa).

Preregistered Analyses & Sample Size: All planned analyses adhered to the preregistered Stage 1 analysis plan, and any *ex post* analyses conducted which were not specified in the preregistration are explicitly stated as such below. I experienced minimal model convergence and singularity issues, and as such all models included random intercepts for participants (P_{0i}), along with all other variables in the models specified in the Method section. Furthermore, the final sample size of $N_{observations} = 97{,}176$ for Models 3–5 was closely aligned with the anticipated sample size, preserving the integrity of the Stage 1 power analysis. Note as well that, as anticipated, the sample size for Models 1–2 was larger, $N_{observations} = 150{,}643$, than the sample on which Models 3–5 was tested. This was due to the fact that many participants in the original data were administered measures of implicit identification, not implicit attitudes. Such participants could not be included in the dataset testing Models 3–5, but could be included in the dataset testing Models 1–2 as those Models did not include participant's own implicit attitudes. Testing Models 1–2 on the smaller dataset used to test Models 3–5 found practically identical results to those presented below on the larger dataset (note: this specific test of Models 1–2 on the smaller dataset was not preregistered).

Models 1 & 2: Table 2 presents the results of Model 1, used to test the Judgment Accuracy Hypothesis (Hypotheses 1a) and the Directional Bias Hypothesis (H1b), and Model 2, used to test the Implicit Detection Hypothesis (H2). H1a and H2 predicted significant and positive linear relationships, true forces, between judgments of the population's

Table 1b
Summary of Results.

Hypothesis	Model	Prediction	Confirmatory or Exploratory	Results
Hypothesis 1a	Model 1	Judgment Accuracy Hypothesis	Confirmatory	*Partially Supported*
Hypothesis 1b	Model 1	Directional Bias Hypothesis	Exploratory	*Observed Bias*
Hypothesis 2	Model 2	Implicit Detection Hypothesis	Confirmatory	*Supported*
Hypothesis 3a	Model 3	Explicit Attitude Projection Hypothesis	Confirmatory	*Supported*
Hypothesis 3b	Model 3	Implicit Attitude Projection Hypothesis	Confirmatory	*Supported*
Hypothesis 4	Model 4	Implicit Attitudes Increase Accuracy Hypothesis	Competing	*Not Supported*
Hypothesis 5	Model 5	Implicit Attitude-Extremity Increases Accuracy Hypothesis	Competing	*Not Supported*

preferences and the population's true explicit and implicit preferences, respectively. H1b predicted mean level over- and under-estimation of the population's true explicit preferences within domain.

Model 1 found a significant positive linear relationship, a truth force, between the population's true explicit preferences and participants' judgments of the population's preferences, which initially supported H1a. However Model 2, where the population's true implicit preferences were added as a predictor alongside the population's true explicit preferences, no longer observed a significant truth force between the population's true explicit preferences and participants' judgments. Rather, Model 2 found only a significant linear truth force between the population's true implicit preferences and participants' judgments of the population's preferences, in support of H2. As such, support for H1a was only partial. To compare whether Model 2 explained meaningfully more variance than Model 1, I conducted a likelihood-ratio test. The LR test found that Model 2 explained significantly more variance than Model 1, $\chi^2(1) = 4.61$, $p = 0.032$.

To test exploratory Hypothesis 1b regarding directional biases in judgment, I examined the random intercept estimates across attitude domains in Model 1. Fig. 1 plots the observed random intercept estimates with 95% confidence intervals. Note that the intercept estimate for Model 1 is −0.27 ($p = 0.001$), suggesting that on average participants were underestimating the strength of the population's true explicit preferences. The dotted vertical red line in Fig. 1 was therefore situated at +0.27, representing the value where mean judgments of the population's true preferences were accurate and displayed no directional bias. Random intercept estimates for which the dotted red line passes through the confidence interval were interpreted as attitude domains which exhibited no directional bias.

Note that in Fig. 1 the labels are situated such that the right side category is where the population's true explicit preferences laid. For example, the topmost estimate is for the attitude domain "Fat People - Thin People." This estimate is to the far right (and does not cross the dotted red line), which indicated that participants significantly overestimated the population's explicit preference for "thin" over "fat" people. In other words, while the population's true explicit preference was in fact for "thin" over "fat" people, on average participants overestimated the magnitiate of that preference while getting the direction right. For another example take the bottommost attitude domain, "Television - Books." The fact that "Books" is on the right side of the label indicated the population's true explicit preference was for books over television. Yet the far left plotted intercept estimate indicated that participants significantly underestimated the extent to which the population values books over television, suggesting that in this domain participants' directional bias was in the wrong direction.

Model 3: Table 3 presents the results of Model 3. Model 3 tested the Explicit Attitude Projection Hypothesis (Hypotheses 3a) and the Implicit Attitude Projection Hypothesis (H3b) which predicted positive linear relationships, bias forces, between participants' own explicit and implicit attitudes and their judgments of the population's preferences, respectively. To test these hypotheses participants' own explicit and implicit attitudes were added as predictors to Model 2.

First, Model 3 found further support for H2. As with Model 2, Model 3 found a significant positive linear truth force between the population's true implicit (but not explicit) preferences and participants' judgments. Model 3 also found support for H3a and H3b, as participants' own implicit and explicit attitudes were associated with participants' judgments of the population's true preferences, above the population's actual true preferences. This constituted evidence for implicit and explicit attitude projection, the bias force of attitudes on judgment.

Model 3 could not be compared to Model 2 using a likelihood-ratio test, as the samples used for each were distinct. However, the Conditional R^2 (Nakagawa et al., 2017), which captured the estimated percentage of variance explained by both the fixed and random effects, for Models 2 and 3 were qualitatively compared. Model 2 explained 28.5% of the variance in participant's judgments whereas Model 3 explained 31.6% of the variance, a difference of 3.1%.

Models 4 & 5: The results of Models 4 and 5 can be found in Table 4. Models 4 and 5 tested the competing Implicit Attitudes Increase Accuracy Hypothesis (H4) and Implicit Attitude-Extremity Increases Accuracy Hypothesis (H5). Both hypotheses predicted that implicit attitudes would affect the accuracy of judgment. But whereas H4 suggested the direction of implicit attitudes would matter in affecting accuracy, H5 suggested the extremity of attitudes irrespective of the direction would matter.

I found no evidence to support either H4 or H5. H4 lacked support as I observed no significant interactions between implicit attitudes and the population's true explicit or implicit preferences in Model 4. H5 too lacked support as I observed no significant polynomial effects or interactions of implicit attitudes or the population's true explicit attitudes in Model 5. Note however that both Models 4 and 5 replicated the pattern of results found in Model 3, which provided further evidence for H2, H3a and H3b. Using likelihood-ratio tests I observed no model improvement above Model 3 for either Model 4, $\chi^2(4) = 2.39$, $p = 0.664$, or Model 5, $\chi^2(6) = 4.08$, $p = 0.665$, nor did Model 5 improve upon Model 4, $\chi^2(2) = 1.69$, $p = 0.430$. As such, Model 3 was the best fitting and most parsimonious model observed across the analyses.

Non-Preregistered Analyses: All analyses presented below were not preregistered as part of the Stage 1 manuscript. I ran a single non-preregistered model, henceforth Model 6. The argument for Model 6 is that Hypothesis 5, as preregistered, ought to be amended *ex post* in light of the observed partial support for Hypothesis 1. The observed preregistered findings in Models 1–5, in total, suggested that judgments of the population's preferences were uniquely associated with the population's true implicit but not explicit preferences, contrary to the prediction of Hypothesis 1. As such, it is arguable that Hypothesis 5, which predicted

Table 2
Linear mixed-effects models predicting judgments of the population's preferences.

	Model 1			Model 2		
Predictors	*Estimates*	*95% CI*	*p*	*Estimates*	*95% CI*	*p*
(Intercept)	−0.27	−0.44 to −0.11	**0.001**	−0.27	−0.43 to −0.11	**0.001**
True Explicit Preference	0.39	0.06–0.72	**0.024**	0.18	−0.19–0.55	0.347
True Implicit Preference				0.60	0.06–1.14	**0.032**
Random Effects						
σ^2	1.79			1.79		
τ_{00}	0.02 $_{user\_id}$			0.02 $_{user\_id}$		
	0.66 $_{domain}$			0.63 $_{domain}$		
ICC	0.27			0.26		
N	95 $_{domain}$			95 $_{domain}$		
	96,950 $_{user\_id}$			96,950 $_{user\_id}$		
Observations	150,643			150,643		
Marginal R^2 / Conditional R^2	0.015 / 0.285			0.027 / 0.285		

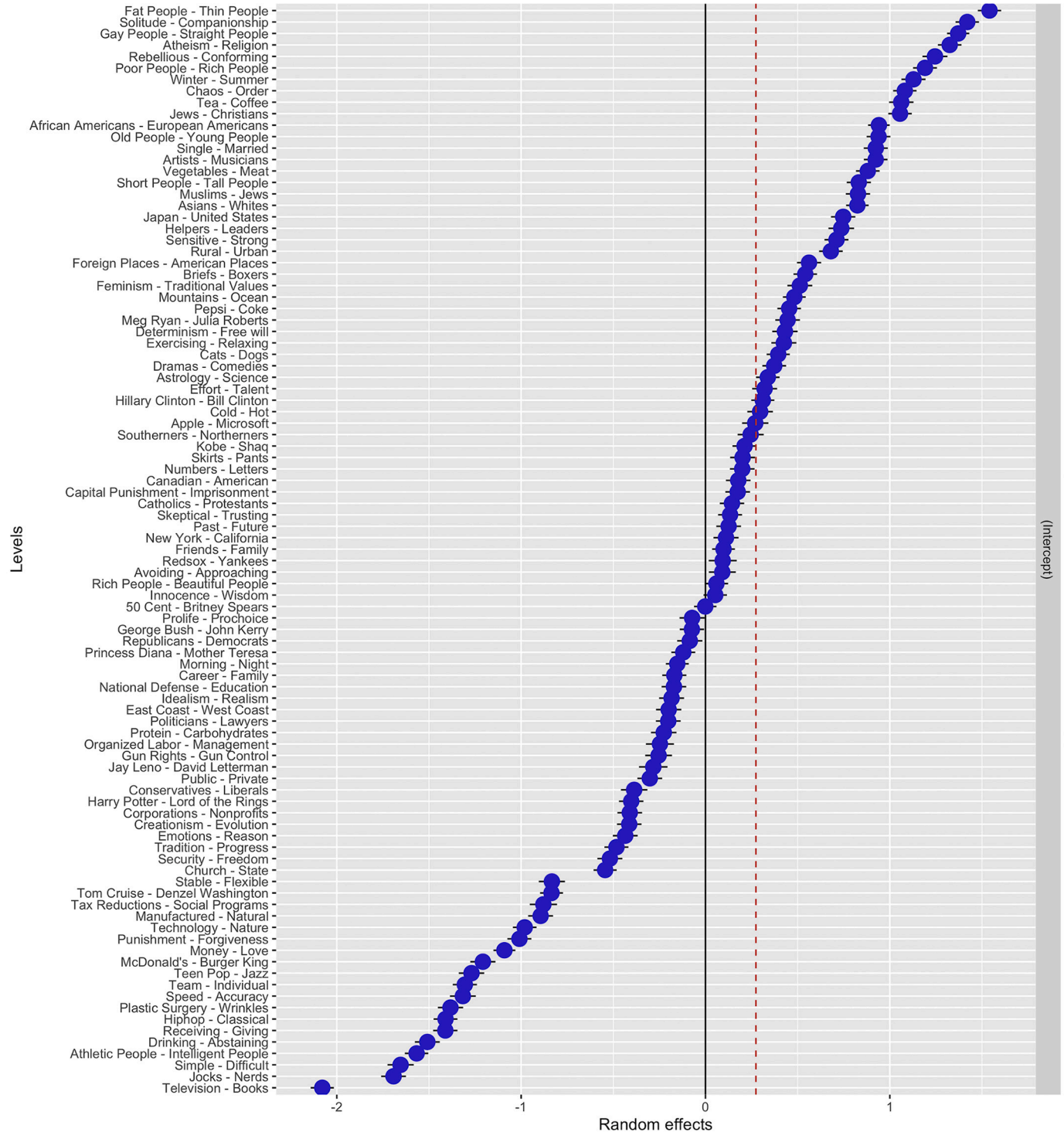

Fig. 1. Random Intercept Plot with 95% Cls. Note: Dotted vertical line is situated at +0.27, representing the value where mean judgments of the population's true preferences were accurate

a negative curvilinear relationship between judgments of the population's preferences and the population's true preferences, would be better operationalized as the curvilinear effect of the population's true *implicit,* rather than the true explicit, attitudes in predicting judgment. Below is the specification of Model 6. Relative to Model 5, in Model 6 I removed the squared value of the population's true explicit preferences TE_d^2 and added the squared value of the population's true implicit preferences TI_d^2 for the estimate s_1.

$$J_{di} = b_0 + D_{0d} + P_{0i} + t_1 TE_d + t_2 TI_d + (b_1 + D_{1d})BE_{di} + (b_2 + D_{2d})BI_{di} + x_1 BE_{di} TE_d + x_2 BE_{di} TI_d + x_3 BI_{di} TE_d + x_4 BI_{di} TI_d + s_1 TI_d^2 + s_2 BI_{di}^2 + e_{di}$$

The results for Model 6 can be found in Table 5. Model 6 found no support for H5, as there was no significant polynomial relationship between either participants' implicit attitudes or the population's true implicit attitudes in predicting participants' judgments. Like Model 5, Model 6 conformed to the pattern observed in Model 3, which provided

Table 3
Linear mixed-effects model predicting judgments of the population's preferences.

	Model 3		
Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	−0.26	−0.43 to −0.10	**0.002**
True Explicit Preference	0.16	−0.21–0.53	0.396
True Implicit Preference	0.68	0.14–1.22	**0.015**
Explicit Attitude	0.08	0.06–0.10	**<0.001**
Implicit Attitude (D)	0.04	0.01–0.07	**0.016**
Random Effects			
σ^2	1.70		
$\tau_{00\ user\_id}$	0.01		
$\tau_{00\ domain}$	0.64		
$\tau_{11\ domain.imp\_att}$	0.02		
$\tau_{11\ domain.exp\_att}$	0.01		
$\rho_{01\ domain.imp\_att}$	0.06		
$\rho_{01\ domain.exp\_att}$	0.20		
ICC	0.29		
N_{domain}	95		
N_{user_id}	68,418		
Observations	97,176		
Marginal R^2 / Conditional R^2	0.038 / 0.316		

p-values less than 0.05 are bolded

further support for H2, H3a, and H3b. Using a likelihood-ratio test I found no evidence that Model 6 explained more variance than Model 3, $\chi^2(6) = 2.93$, $p = 0.818$.

4. Discussion

Across six models this Registered Report examined several hypotheses related to how implicit attitudes affected social judgment accuracy, beyond the effects of explicit attitudes on social judgment. I found that judgments of the population's true preferences across 95 attitude domains were associated with the population's true implicit preferences, but not explicit preferences, and that individuals project their own implicit and explicit attitudes when judging the population's true preferences. Contrary to the preregistered hypotheses, I found no evidence that an increase in implicit attitudes, either in pure magnitude or directional strength, was associated with increased accuracy in judging the population's true preferences.

4.1. Model interpretations

A critical consideration that should preface all interpretations of the results are the levels of analyses. The level of analysis for each hypothesis is explicated in Table 1a. Because the true values examined in juxtaposition to participants' judgments were static within domain, many of the hypotheses were necessarily tested at the level of the *population* (Level 2 in the models), not at the level of the individual (Level 1 in the models). The only hypotheses tested at the level of the participants were Hypothesis 1b, which examined mean-level directional biases in judgment, and Hypotheses 3a and 3b, which examined projection of explicit and implicit attitudes, respectively. As such, the results supported the inference that *individuals* project their explicit and implicit attitudes when judging the population's true preferences, and that individuals exhibit directional biases in their point estimates of the population's true preferences. However, Hypotheses 1 and 2 were tested at the level of the population, as they examined population-level linear associations, and Hypotheses 4 and 5 were tested as Level 1 x Level 2 interactions. As such, the results supported the inference that *within the population* judgments were associated with the population's true implicit, but not explicit, preferences. Similarly, the results from Models 4 and 5 suggest that an increase in the population's implicit attitudes was not associated with an increase in the association between judgments and the population's true preferences. It would *not* be valid to infer from Models 4 and 5 that there was no association between an individual's implicit attitudes and the accuracy of their individual judgments.

4.2. Supported and unsupported hypotheses

The finding in support of Hypothesis 2, but not Hypothesis 1a, that judgments of the population's true preferences were associated with the population's true implicit, but not explicit, preferences suggests that individuals have meaningful insight into the distribution of implicit attitudes within the population. Such accuracy is akin to normative or stereotypic accuracy (Furr, 2008; Jussim, Crawford, & Rubinstein,

Table 4
Linear mixed-effects models predicting judgments of the population's preferences.

	Model 4			Model 5		
Predictors	*Estimates*	*95% CI*	*p*	*Estimates*	*95% CI*	*p*
(Intercept)	−0.26	−0.43 to −0.10	**0.002**	−0.34	−0.56 to −0.13	**0.002**
True Explicit Preference	0.18	−0.19–0.56	0.342	0.03	−0.43–0.50	0.897
True Implicit Preference	0.64	0.09–1.19	**0.025**	0.70	0.15–1.26	**0.015**
Explicit Attitude	0.08	0.06–0.10	**<0.001**	0.08	0.06–0.10	**<0.001**
Implicit Attitude (D)	0.04	0.01–0.07	**0.017**	0.04	0.01–0.07	**0.015**
True Explicit:D	−0.04	−0.11–0.04	0.351	−0.04	−0.11–0.04	0.355
True Implicit:D	−0.02	−0.13–0.10	0.782	−0.01	−0.12–0.10	0.866
True Explicit:Expl_Att	0.01	−0.04–0.06	0.728	0.01	−0.04–0.06	0.728
True Implicit:Expl_Att	−0.03	−0.10–0.04	0.455	−0.03	−0.10–0.04	0.454
True Explicit Preferencesˆ2				0.31	−0.25–0.88	0.280
Implicit Attitudeˆ2				0.01	−0.02–0.04	0.471
Random Effects						
σ^2	1.70			1.70		
τ_{00}	0.01_{user_id}			0.01_{user_id}		
	0.64_{domain}			0.63_{domain}		
τ_{11}	$0.02_{domain.imp\_att}$			$0.02_{domain.imp\_att}$		
	$0.01_{domain.exp\_att}$			$0.01_{domain.exp\_att}$		
ρ_{01}	$0.06_{domain.imp\_att}$			$0.07_{domain.imp\_att}$		
	$0.20_{domain.exp\_att}$			$0.21_{domain.exp\_att}$		
ICC	0.29			0.29		
N	95_{domain}			95_{domain}		
	$68{,}418_{user\_id}$			$68{,}418_{user\_id}$		
Observations	97,176			97,176		
Marginal R^2 / Conditional R^2	0.038 / 0.315			0.041 / 0.315		

p-values less than 0.05 are bolded

Table 5
Linear mixed-effects model predicting judgments of the population's preferences.

	Model 6		
Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	−0.26	−0.45 to −0.07	**0.010**
True Explicit Preference	0.18	−0.19–0.56	0.341
True Implicit Preference	0.64	0.09–1.19	**0.025**
Explicit Attitude	0.08	0.06–0.10	**<0.001**
Implicit Attitude (D)	0.04	0.01–0.07	**0.015**
True Explicit:D	−0.04	−0.11–0.04	0.355
True Implicit:D	−0.01	−0.12–0.10	0.864
True Explicit:Expl_Att	0.01	−0.04–0.06	0.727
True Implicit:Expl_Att	−0.03	−0.10–0.04	0.454
True Implicit Preference^2	−0.06	−0.97–0.84	0.891
Implicit Attitude^2	0.01	−0.02–0.04	0.472
Random Effects			
σ^2	1.70		
$\tau_{00\ user\_id}$	0.01		
$\tau_{00\ domain}$	0.64		
$\tau_{11\ domain.imp\_att}$	0.02		
$\tau_{11\ domain.exp\_att}$	0.01		
$\rho_{01\ domain.imp\_att}$	0.06		
$\rho_{01\ domain.exp\_att}$	0.20		
ICC	0.29		
N_{domain}	95		
N_{user_id}	68,418		
Observations	97,176		
Marginal R^2 / Conditional R^2	0.038 / 0.315		

p-values less than 0.05 are bolded

2015) for implicit attitudes, and parallels research suggesting individuals do have insight into their own implicit attitudes (Hahn et al., 2014; Hahn & Gawronski, 2019). One possible explanation for why individuals may have knowledge of the population's true implicit, but not explicit, preferences may relate to how individuals infer the preferences of others through behavior. If implicit attitudes manifest at behaviors that actors are relatively unaware of but observers can perceive relatively easily, for example a white person seeking greater physical distance from a black stranger compared to a white stranger, then observers may be able to infer preferences which are captured by implicit attitude measures. Another possibility is that the aggregated explicit attitude measures captured socially desirable responses which did not reflect true (explicit) preferences. This is unlikely however, as socially desirable responding should make individuals' responses to explicit attitude measures closer to the true population mean, not further.

One further consideration is that participants were not asked explicitly to estimate the implicit attitudes of the population, they were simply asked "Does the [average person/culture you live in/most people] prefer X or Y." Considered in the context of Model 2's findings, this suggests the observed judgment accuracy represents lay perceptions of social preferences, and that people's lay perceptions were better aligned with others' true implicit preferences than true explicit preferences. This finding bolsters the argument that aggregated implicit attitudes are useful tools for predicting social psychological outcomes (Hehman et al., 2019; Schimmack, 2021).

Exploratory Hypothesis 1b predicted that judgments would exhibit mean-level biases distributed across the attitude domains, and the results bore out that pattern. Most attitude domains exhibited some level of directional bias, and across domains the average bias (the model intercept) was to underestimate the strength of the population's true relative preferences. There is one qualitative pattern worthy of note: most of the attitude domains that represented measures of intergroup prejudice ("African-Americans - European Americans," "Jews - Christians," "Gay People - Straight People," etc.) tended to exhibit overestimation. While individuals are relatively accurate in judging the prejudices of specific others (Alaei & Rule, 2019; Hehman, Leitner, Deegan, & Gaertner, 2013; LaCosse et al., 2015; Richeson & Shelton, 2005; Rudman & Fetterolf, 2014), the finding that participants tended to overestimate prejudicial attitudes aligns with recent work demonstrating significant overestimation of intergroup prejudices in political contexts (Armaly & Enders, 2020; Moore-Berg, Ankori-Karlinsky, Hameiri, & Bruneau, 2020; Ruggeri et al., 2021).

The findings in support of Hypotheses 3a and 3b, that individuals projected their explicit and implicit attitudes, suggest that explicit and implicit attitudes play distinct roles in biasing the accuracy of social judgment. Evidence for the projection of individuals' explicit attitudes when judging others is widely observed in the psychological literature (Ames, 2004; Cho & Knowles, 2013; Cronbach, 1955; Hoch, 1987), yet to my knowledge the findings in support of Hypothesis 3b represents the first documented instance of the unique projection of implicit attitudes in social judgment after controlling for explicit attitude projection (note that the implicit projection effect was not due to the inclusion of explicit attitudes in the model, and in fact implicit projection was stronger if entered into the model prior to modeling explicit projection). The finding that implicit attitudes are projected when judging the population's true preferences is supported by theorizing of projection as an automatic process outside of conscious awareness (Birch & Bloom, 2007; Epley, Keysar, Van Boven, & Gilovich, 2004), although the question of how conscious projection processes are is still empirically debated (Gramzow, Gaertner, & Sedikides, 2001; Van Boven, Judd, & Sherman, 2012). The results here potentially shed light on this debate. We observed simultaneous projection of implicit and explicit attitudes, and this could be interpreted as unconscious and conscious projection processes respectively.

The lack of support for either Hypotheses 4 or 5 potentially speaks to a larger theoretical question regarding the nature of implicit attitudes. Hypothesis 4 was based on the prediction that projection of attitudes in the direction of the population's true preferences (which was observed) would cause an incidental increase in accuracy. Hypothesis 5 was based on the prediction that implicit attitude-strength reflects social knowledge which would confer greater judgment accuracy. While caution is urged in the interpretation of null findings, one potential explanation is that the "knowledge" which implicit attitudes might reflect is too far beyond individuals' conscious awareness to affect judgment accuracy. This may explain why we observed the projection of implicit attitudes (i.e., implicit attitudes biased judgment in a manner individuals may be unaware of), but we did not observe individuals' implicit attitudes conferring greater accuracy when asked to make overt judgments of the population's true preferences. If there is a manner in which implicit attitudes mechanistically affect the accuracy of social judgments of others' preferences, then new theoretical frameworks will need to be developed to understand the nature of such a relationship. Conversely, it may be the case that the null is unconditionally true, and there is no manner in which implicit attitudes affect social judgment accuracy. Another consideration is the level of analysis. It may be the case that implicit attitudes affect social judgment accuracy only at the individual level, and the current results cannot speak directly to that possibility, although they provided no reason to believe this may be the case.

4.3. Limitations and future research

Many of the limitations discussed below relate to the fact that this Registered Report sought to answer a novel research question using data collected from another project with a separate research question. I strongly encourage researchers seeking to replicate and expand upon the findings here to consider how changes to the design, measures, and operationalizations may strengthen one's ability to explore the relationships between social judgment accuracy and implicit attitudes. Several suggestions for such potential improvements are discussed below.

One central limitation across all tested hypotheses is that the dependent variable did not ask participants to discriminate between the explicit and implicit attitudes of others. Future research ought to

consider directly asking perceivers to judge both the explicit and implicit attitudes of their targets to better disentangle the extent to which individuals can meaningfully detect and discriminate between each. Especially given the findings from Model 2 that the population's true implicit but not explicit preferences were associated with judgment, disambiguating participants' judgments would be a crucial methodological step in confirming whether individuals' knowledge of the population's true preferences are solely limited to the domain of the implicit. More broadly, the findings of Model 2 suggest that existing research might be *underestimating* judgment accuracy in any design where the dependent variable is similar to the one examined here, yet the independent variable(s) (the true values) are constrained to explicitly reported true preferences.

Another limitation is that implicit attitudes were exclusively measured using the IAT. Alternative measures of implicit attitudes, including evaluative priming (Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Koppehele-Gossel, Hoffmann, Banse, & Gawronski, 2020), and the Affect Misattribution Procedure (Payne et al., 2005), should be utilized by future research to validate the population-level true values, individual-level implicit attitudes, and their relationships to judgments of the population's true preferences. Moreover, while research suggests that aggregate IAT scores are valid measures of group-level preferences (Hehman et al., 2019), scholars have challenged the measurement validity of individual-level IAT responses for capturing implicit attitudes. In particular, Schimmack (2021) argues "The most promising use of the IAT is to use it as a complementary method and use the shared variance between IAT scores and explicit measures to control for measurement error in both methods" (pg. 411). If one fully accepts this critique, it would suggest that the test here of the Implicit Attitude Projection Hypothesis (Hypothesis 3b) is not capturing the projection of implicit attitudes but rather capturing additional variance in explicit attitude projection. I would not have proposed to test for implicit attitude projection if I were not more confident in the validity of that IAT as a measure of implicit attitudes, yet Schimmack's (2021) critique should at the very least constrain the confidence in the implicit attitude projection effect until further evidence can be gathered in support of its robustness.

Another central limitation of the present data is the crossed levels of analysis (discussed above in Model Interpretations). In order for future research to overcome the inferential limitations inherent in examining hypotheses across levels of analysis, researchers would need to measure multiple judgments with varying true values (i.e., repeated-measures) in order to model participant-level linear relationships between judgments and corresponding true values. Such judgments would ideally be measured across multiple attitude domains, alongside multiple within-participant measures of implicit attitudes across domains, to robustly examine participant-level judgment accuracy and interactions between accuracy and implicit attitudes. Future research ought to also counterbalance the order of explicit and implicit attitude measures, as in the presented data the implicit measures always preceded the explicit measures.

Future research is also encouraged to more formally examine the large biases observed in the test of exploratory Hypothesis 1b. As I did not specify a priori hypotheses regarding which attitude domains would exhibit what, if any, directional biases, such questions are beyond the current scope of this project. Yet such examinations could begin with the data available herein. For example, by performing qualitative or quantitative content analyses of the attitude domains and using those results to inform the development of studies seeking to understand why some domains exhibited greater or less judgment bias.

The lack of support for Hypotheses 4 and 5 also highlight a critical methodological decision in the preregistered analyses: the choice to model random slopes for individuals' own implicit and explicit attitudes (i.e., random slopes for projection within domain). This choice was based on evidence that only modeling random intercepts can bias model predictions and inflate Type 1 error rates (Barr et al., 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). When Model 4 was run without random slopes for participants' own explicit and implicit attitudes (which was not preregistered) I observed several statistically significant interactions which would have supported some of the stated hypotheses. The inclusion of random slopes however rendered those interactions statistically insignificant, and including random slopes in Model 4 significantly improved model fit over a random intercepts-only model, $\chi^2(5) = 1782$, $p < 0.001$.

In simple terms, this means an assumption of the random-intercepts only model, that the implicit and explicit projection effects were invariant across attitude domains, did not hold. Had I opted for model "parsimony" rather than a maximal random structure approach, I might have introduced spurious findings into the literature. This highlights the need for more careful consideration of how hierarchical modeling assumptions may reflect psychological assumptions which are unfounded.

More broadly, the breadth of the results found here reinforce methodological arguments for utilizing componential approaches to studying judgment accuracy (Biesanz, 2010; Lees & Cikara, 2021; West & Kenny, 2011; Wood & Furr, 2016) which are common in the person perception literature but have yet to be adopted more widely in social psychology. I examined judgment accuracy as the linear relationship between two different true values and judgment, mean-level directional biases in judgment relative to the true values, the biasing effects of two distinct projection processes, and the interactive effects of projection on accuracy, all within a single unified hierarchical model framework. Future work seeking to bridge social judgment accuracy research with research on implicit attitudes is highly encouraged to utilize frameworks such as the Truth & Bias Model used here (West & Kenny, 2011) or the Social Accuracy Model (Biesanz, 2010) to guide theorizing, data collection, and analyses.

In conclusion, I found generalizable evidence that implicit attitudes matter for social judgments of others' preference, but do not make those judgments more or less accurate. Participants displayed the capacity to meaningfully detect the distribution of implicit (but not explicit) attitudes within the population, and participants projected their own explicit and implicit attitudes when making judgments of the population's true preferences. These findings suggest implicit attitudes play a key role in social judgment accuracy which we have only begun to understand.

Acknowledgement

I thank Charlie Ebersole and the AIID Team for organizing the call for Registered Reports, J. Rebecca Young for feedback on the manuscript draft, and Steven Worthington and Ista Zahn at Harvard's IQSS for statistical guidence.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2021.104202.

References

Alaei, R., & Rule, N. O. (2019). People can accurately (but not adaptively) judge strangers' antigay prejudice from faces. *Journal of Nonverbal Behavior, 43*(3), 397–409. https://doi.org/10.1007/s10919-019-00305-2.

Ames, D. R. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573–585. https://doi.org/10.1037/0022-3514.87.5.573.

Arcuri, L., Castelli, L., Galdi, S., Zogmaister, C., & Amadori, A. (2008). Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology, 29*(3), 369–387. https://doi.org/10.1111/j.1467-9221.2008.00635.x.

Armaly, M. T., & Enders, A. M. (2020). The role of affective orientations in promoting perceived polarization. *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2020.24.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1–70. https://doi.org/10.1037/h0093718.

Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2004). No place for nostalgia in science: A response to Arkes and Tetlock. *Psychological Inquiry, 15*(4), 279–310. https://doi.org/10.1207/s15327965pli1504.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Barranti, M., Carlson, E. N., & Côté, S. (2017). How to test questions about similarity in personality and social psychology research: Description and empirical demonstration of response surface analysis. *Social Psychological and Personality Science, 8*(4), 465–475. https://doi.org/10.1177/1948550617698204.

Barranti, M., Carlson, E. N., & Furr, R. M. (2016). Disagreement about moral character is linked to interpersonal costs. *Social Psychological and Personality Science, 00*(0), 1–12. https://doi.org/10.1177/1948550616662127.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Berns, G. S., Capra, C. M., Moore, S., & Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *NeuroImage, 49*(3), 2687–2696. https://doi.org/10.1016/j.neuroimage.2009.10.070.

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research, 45*(5), 853–885. https://doi.org/10.1080/00273171.2010.519262.

Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy-motivated perceivers see the personality of others more distinctively but less normatively than perceivers without an explicit goal. *Psychological Science, 21*(4), 589–594. https://doi.org/10.1177/0956797610364121.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*(5), 382–386. https://doi.org/10.1111/j.1467-9280.2007.01909.x.

Brand, R., Heck, P., & Ziegler, M. (2014). Illegal performance enhancing drugs and doping in sport: A picture-based brief implicit association test for measuring athletes' attitudes. *Substance abuse treatment, prevention, and policy, 9*(1), 7. https://doi.org/10.1186/1747-597X-9-7.

Brewer, M. B., & Miller, N. (1984). Beyond the contact hypothesis: Theoretical perspectives on desegregation. In M. B. Brewer, & N. Miller (Eds.), *Groups in contact: The psychology of desegregation (pp. 281–302)*. Academic Press.

Bursztyn, L., Gonzalez, A. L., & Yanagizawa-Drott, D. (2018). *Misperceived social norms: Female labor force participation in Saudi Arabia. National Bureau of Economic Research Working Paper Series*.

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.

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology, 20*(13), 1165–1170. https://doi.org/10.1016/j.cub.2010.04.055.

Carlson, E. N. (2016). Meta-accuracy and relationship quality: Weighing the costs and benefits of knowing what people really think about you. *Journal of Personality and Social Psychology, 111*(2), 250–264. https://doi.org/10.1037/pspp0000107.

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them? *Journal of Personality and Social Psychology, 101*(4), 831–846. https://doi.org/10.1037/a0024297.

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.

Cho, J. C., & Knowles, E. D. (2013). I'm like you and you're like me: Social projection and self-stereotyping both help explain self–other correspondence. *Journal of Personality and Social Psychology, 104*(3), 444–456. https://doi.org/10.1037/a0031017.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1974), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in experimental social psychology, 56*, 131–199. Elsevier https://doi.org/10.1016/bs.aesp.2017.03.001.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*(3), 177–193. https://doi.org/10.1037/h0044919.

Dambrun, M., & Guimond, S. (2004). Implicit and explicit measures of prejudice and stereotyping: Do they assess the same underlying knowledge structure? *European Journal of Social Psychology, 34*(6), 663–676. https://doi.org/10.1002/ejsp.223.

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36*(6), 37.

Enders, A. M., & Armaly, M. T. (2018). The differential effects of actual and perceived polarization. *Political Behavior, 41*(3), 815–839.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*(3), 327–339. https://doi.org/10.1037/0022-3514.87.3.327.

Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of Personality and Social Psychology, 114*(4), 547–571. https://doi.org/10.1037/pspa0000115.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229–238. https://doi.org/10.1037/0022-3514.50.2.229.

Funder, D. C. (1995). On the accuracy of personality judgment:A realistic approach. *Psychological Review, 102*(4), 652–670.

Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality, 76*(5), 1267–1316. https://doi.org/10.1111/j.1467-6494.2008.00521.x.

Goh, J. X., Rad, A., & Hall, J. A. (2017). Bias and accuracy in judging sexism in mixed-gender social interactions. *Group Processes & Intergroup Relations, 20*(6), 850–866. https://doi.org/10.1177/1368430216638530.

Gramzow, R. H., Gaertner, L., & Sedikides, C. (2001). Memory for in-group and out-group information in a minimal group context: The self as an informational base. *Journal of Personality and Social Psychology, 80*(2), 188–205. https://doi.org/10.1037/0022-3514.80.2.188.

Green, P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality & Social Psychology, 74*(6), 1464–1480.

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

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology, 116*(5), 769–794.

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

Hall, J. A., & Goh, J. X. (2017). Studying stereotype accuracy from an integrative social-personality perspective. *Social and Personality Psychology Compass, 11*(11), Article e12357. https://doi.org/10.1111/spc3.12357.

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

Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2013). Facial structure is indicative of explicit support for prejudicial beliefs. *Psychological Science, 24*(3), 289–296. https://doi.org/10.1177/0956797612451467.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*(2), 221–234.

Hofmann, W., & Friese, M. (2008). Impulses got the better of me: Alcohol moderates the influence of implicit attitudes toward food cues on eating behavior. *Journal of Abnormal Psychology, 117*(2), 420–427. https://doi.org/10.1037/0021-843X.117.2.420.

Holtz, R., & Norman, M. (1985). Assumed similarity and opinion certainty. *Journal of Personality and Social Psychology, 48*(4), 890–989.

Human, L. J., Carlson, E. N., Geukes, K., Nestler, S., & Back, M. D. (2018). Do accurate personality impressions benefit early relationship development? The bidirectional associations between accuracy and liking. *Journal of Personality and Social Psychology. doi.* https://doi.org/10.1037/pspp0000214.

Humberg, S., Nestler, S., & Back, M. D. (2019). Response surface analysis in personality and social psychology: Checklist and clarifications for the case of congruence hypotheses. *Social Psychological and Personality Science, 10*(3), 409–419. https://doi.org/10.1177/1948550618757600.

Ivanov, I., Muller, D., Delmas, F., & Wänke, M. (2018). Interpersonal accuracy in a political context is moderated by the extremity of one's political attitudes. *Journal of Experimental Social Psychology, 79*, 95–106. https://doi.org/10.1016/j.jesp.2018.07.001.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior, 29*, 39–69. https://doi.org/10.1016/j.riob.2009.10.001.

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review, 100*(1), 109–128. https://doi.org/10.1037/0033-295X.100.1.109.

Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology, 61*(3), 366–379. https://doi.org/10.1037/0022-3514.61.3.366.

Jussim, L. (2017). Précis of social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy. *Behavioral and Brain Sciences, 40*. https://doi.org/10.1017/S0140525X1500062X.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science, 24*(6), 490–497. https://doi.org/10.1177/0963721415605257.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review, 8*(3), 265–280. https://doi.org/10.1207/s15327957pspr0803_3.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin, 102*(3), 390–402.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (2020). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology, 87*, 103905. https://doi.org/10.1016/j.jesp.2019.103905.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*(4), 596–610.

Krzyzaniak, S. L., Colman, D. E., Letzring, T. D., McDonald, J. S., Biesanz, J. C., et al. (2019). The effect of information quantity on distinctive accuracy and normativity of personality trait judgments. *European Journal of Personality, 33*(2), 197–213. https://doi.org/10.1002/per.2196.

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, 201820238. https://doi.org/10.1073/pnas.1820238116.

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, 201820240. https://doi.org/10.1073/pnas.1820240116.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of statistical software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

LaCosse, J., Tuscherer, T., Kunstman, J. W., Plant, E. A., Trawalter, S., & Major, B. (2015). Suspicion of white people's motives relates to relative accuracy in detecting external motivation to respond without prejudice. *Journal of Experimental Social Psychology, 61*, 1–4. https://doi.org/10.1016/j.jesp.2015.06.003.

Lai, C. K., & Banaji, M. R. (2020). The psychology of implicit intergroup bias and the prospect of change. In D. Allen, & R. Somanathan (Eds.), *Difference without domination: Pursuing justice in diverse democracies*. University of Chicago Press.

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

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour, 4*(3), 279–286. https://doi.org/10.1038/s41562-019-0766-4.

Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B, 376*(1822), 20200143. https://doi.org/10.1098/rstb.2020.0143.

Lewis, K. L., Hodges, S. D., Laurent, S. M., Srivastava, S., & Biancarosa, G. (2012). Reading between the minds: The use of stereotypes in empathic accuracy. *Psychological Science, 23*(9), 1040–1046. https://doi.org/10.1177/0956797612439719.

Li, Q., & Hong, Y.-Y. (2001). Intergroup perceptual accuracy predicts real-life intergroup interactions. *Group Processes & Intergroup Relations, 4*(4), 341–354. https://doi.org/10.1177/1368430201004004004.

MacCoun, R. J. (2012). The burden of social proof: Shared thresholds and social influence. *Psychological Review, 119*(2), 345–372.

Mann, T. C., Kurdi, B., & Banaji, M. R. (2019). How effectively can implicit evaluations be updated? Using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General. doi*. https://doi.org/10.1037/xge0000701.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001.

Miller, D. T., & Nelson, L. D. (2002). Seeing approach motivation in the avoidance behavior of others: Implications for an understanding of pluralistic ignorance. *Journal of Personality and Social Psychology, 83*(5), 1066–1075. https://doi.org/10.1037//0022-3514.83.5.1066.

Monin, B., & Norton, M. I. (2003). Perceptions of a fluid consensus: Uniqueness bias, false consensus, false polarization, and pluralistic ignorance in a water conservation crisis. *Personality and Social Psychology Bulletin, 29*(5), 559–567. https://doi.org/10.1177/0146167203029005001.

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences, 117*(26), 14864–14872. https://doi.org/10.1073/pnas.2001263117.

Murray, S. L., Holmes, J. G., Bellavia, G., Griffin, D. W., & Dolderman, D. (2002). Kindred spirits? The benefits of egocentrism in close relationships. *Journal of Personality and Social Psychology, 82*(4), 563–581. https://doi.org/10.1037/0022-3514.82.4.563.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134).

Neyer, F. J., Banse, R., & Asendorpf, J. B. (1999). The role of projection and empathic accuracy in dyadic perception between older twins. *Journal of Social and Personal Relationships, 16*(4), 419–442. https://doi.org/10.1177/0265407599164001.

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition & Emotion, 22*(4), 553–594. https://doi.org/10.1080/02699930701438186.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (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.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*(2), 243–256.

Reynolds, S. J., Leavitt, K., & DeCelles, K. A. (2010). Automatic ethics: The effects of implicit assumptions and contextual cues on moral behavior. *Journal of Applied Psychology, 95*(4), 752–760. https://doi.org/10.1037/a0019411.

Richeson, J. A., & Shelton, J. N. (2005). Brief report: Thin slices of racial bias. *Journal of Nonverbal Behavior, 29*(1), 75–86. https://doi.org/10.1007/s10919-004-0890-2.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9*(1), 32–47. https://doi.org/10.1207/s15327957pspr0901_3.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X.

Rudman, L. A., & Fetterolf, J. C. (2014). How accurate are metaperceptions of sexism? Evidence for the illusion of antagonism between hostile and benevolent sexism. *Group Processes & Intergroup Relations, 17*(3), 275–285. https://doi.org/10.1177/1368430213517272.

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., … Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01092-x.

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science, 16*(2), 396–414. https://doi.org/10.1177/1745691619863798.

Smith, T. W., Davern, M., Freese, J., & Stephen, M. (2019). General social survey, 1972–2019. NORC at the University of Chicago. Sponsored by the National Science Foundation. gssdataexplorer.norc.org.

Solomon, B. C., & Vazire, S. (2016). Knowledge of identity and reputation: Do people have knowledge of others' perceptions? *Journal of Personality and Social Psychology, 111*(3), 341–366. https://doi.org/10.1037/pspi0000061.

Van Boven, L., Judd, C. M., & Sherman, D. K. (2012). Political polarization projection: Social projection of partisan attitude extremity and attitudinal processes. *Journal of Personality and Social Psychology, 103*(1), 84–100. https://doi.org/10.1037/a0028145.

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology, 95*(5), 1202–1216. https://doi.org/10.1037/a0013314.

Vuletich, H. A., & Payne, B. K. (2019). Stability and change in implicit bias. *Psychological Science*. https://doi.org/10.1177/0956797619844270, 095679761984427.

West, T. V. (2016). Accuracy of judging group attitudes. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately*. Cambridge University Press. http://psych.nyu.edu/westlab/documents/West_2016_Accuracy%20of%20Judging%20Group%20Attitudes.pdf.

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*(2), 357–378. https://doi.org/10.1037/a0022936.

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science, 10*(2), 145–158. https://doi.org/10.1177/1745691615569849.

Wood, D., & Furr, R. M. (2016). The correlates of similarity estimates are often misleadingly positive: The nature and scope of the problem, and some solutions. *Personality and Social Psychology Review, 20*(2), 79–99. https://doi.org/10.1177/1088868315581119.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science, 22*(7), 894–900. https://doi.org/10.1177/0956797611411057.