RUNNING HEAD: ATTITUDE IMPORTANCE PREDICTS COSTLY INTERGROUP
BEHAVIOR

Personal relevance of attitude importance predicts costly intergroup behavior

Joshua S. Cetron^a, Onyul Haque^{b,1}, Patrick Mair^a, *Mina Cikara^a

^aDepartment of Psychology, Harvard University,
33 Kirkland St., Cambridge, Massachusetts 02138

^bDepartment of Psychology, University of Michigan,
1004 East Hall, 530 Church St., Ann Arbor, Michigan 48109

*Corresponding author:

Mina Cikara

William James Hall 1420

33 Kirkland St., Cambridge, MA, USA 02143

Phone: +1.617.495.3819

Email: mcikara@fas.harvard.edu

This document is a preprint that has not yet been published in a peer-reviewed journal.

¹Present address: Department of Psychology, University of Southern California,
3620 South McClintock Ave., Los Angeles, CA 90089

Abstract

What aspects of attitudes toward social out-groups predict behavior toward those groups? Social attitudes research typically links favorability toward social out-groups to engagement in intergroup behaviors: the stronger your (un)favorability toward a group, the more costly behaviors you should engage in, concerning that group. But many people make strongly-valenced favorability statements about minoritized out-groups without engaging in corresponding costly actions. We investigate subjective attitude importance as a competing predictor of costly intergroup behaviors. When white respondents rate their attitudes toward minoritized racial out-groups as more personally important to them, they are more likely to give up real money to: (1) prevent biased attitude signaling, (2) preserve their reputations, and (3) increase charitable donation to an out-group-supporting nonprofit. Finally, we find that attitude measurements that centered personal relevance better explained variance in costly behavior than generic favorability or importance measurements. These results indicate that measuring subjective attitude importance, with emphasis on the personal relevance of the attitude, improves prediction of supportive and discriminatory behaviors toward minoritized racial out-groups.

Keywords: Intergroup processes, Intergroup Relations, Attitudes, Prejudice/Stereotyping, Bayesian Hierarchical Modeling

Personal relevance of attitude importance predicts costly intergroup behavior

Which aspects of what people *say* about other social groups predict whether they will engage in costly behavior in support of or opposition toward those groups? Predicting what people will do from what they say about others is the classic challenge of social attitudes research. Such predictions are essential to navigating the social world, which perhaps explains why social group attitudes—or prejudice¹—are among the most widely studied and measured attitudes in the social sciences. Social scientists have carefully developed myriad ways of measuring social group attitudes. We have numerous explicit, self-report measures^{2,3}, well-validated scales⁴⁻⁶, implicit measures⁷⁻⁹, and physiological and behavioral measures¹⁰⁻¹². When it comes to studying prejudice, however, the use of these measures focuses almost exclusively on an attitude's *valence extremity* (or “favorability”) as the key predictor of prejudicial behavior: “How warmly do you feel about Democrats or Republicans?” on a thermometer rating; “How favorably would you feel toward a relative who married a person of another race?”; “How closely do you implicitly associate “white” with “good” and “Black” with “bad?” Responses to questions like these are then used to predict engagement in intergroup behaviors, where the typical assumption is that people with more favorable or unfavorable attitude ratings toward a particular group should engage more in (or in more costly) behaviors that affect that group.

But can one dimension—favorability—capture all the nuances of prejudicial behavior? For example, many people make strongly-valenced *statements* without taking corresponding *costly actions* (e.g., meta-analysis by Kurdi et al., 2019¹³). Compare, for example, someone who only tweets a supportive message versus someone who donates money to an out-group's cause. Both may report feeling equally favorably toward minoritized racial out-groups on a feelings thermometer, leaving the difference in their behaviors unexplained by the extremity of their

attitudes alone. In the general attitudes literature ^{14,15}, this would potentially serve as evidence of a difference in attitude *strength*: the former's attitude is weak, whereas the latter's is strong. But what measure best quantifies this particular difference in attitude strength? There remains little consensus among social scientists as to how well existing prejudice measures do or do not reflect genuine, personally-held psychological constructs ^{16,17}, and significant concerns persist about how well these measures predict intergroup behavior ¹⁸. By reducing attitudes to just degrees of favorability, these measures become blunt instruments, collapsing the multiplex of a social group attitude onto a single dimension.

Which other features of social group attitudes might we assess to generate better predictions of intergroup behavior? The broader psychological literature on attitudes outside the domain of prejudice has enumerated several psychological features of attitudes and attitude strength that are theorized to link attitudes and behavior ^{14,15,19} (e.g., Fazio's MODE model ²⁰). One essential feature discussed in the general attitudes literature is attitude *importance*: the subjective response to the question, "how important is this attitude to you?" (for a review see Howe & Krosnick, 2017 ¹⁴). This construct differentiation—separating favorability from importance—has produced notable insight into the way people act on their attitudes, promoting acquisition and accessibility of attitude-relevant information, and increasing engagement in attitude-consistent behavior and stated behavioral intentions ^{14,21–26}. Similar relationships have been observed in political science as well: when people rate their attitudes toward political issues as more personally important to them, they tend to seek out more information about those issues, process that information more deeply and attentively, and even participate more politically as a result ^{27,28}. (Note that measures of identity centrality, the importance or psychological attachment that individuals place on their identities²⁹ are a related construct; however, identity centrality

specifically refers to the importance one places on one's own identity rather than the importance of one's attitudes toward issues or social others.)

Despite the strong evidence that attitude importance is an informative feature of attitude strength for predicting attitude-consistent behavior outside the domain of prejudice, intergroup researchers studying social group prejudice have overlooked this feature, relying instead on favorability to predict complex intergroup behaviors. Here we import insights from the general attitudes literature into the domain of prejudice research. We predict that attitude importance may be a good candidate for assessing the strength of social group attitudes above and beyond baseline favorability-based measures of prejudice. Attitude importance may help us to explain why individuals who make equally extreme statements about their favorability toward minoritized racial out-groups often take different actions, with some engaging in *more costly* intergroup behaviors than others. We test this relationship in two ways: first in a series of willingness-to-pay tasks in which participants prepare to give up real money to affect prejudice-associated outcomes, and then in a series of charitable donation studies in which people *do* give up real money to directly affect the out-group target of their stated attitude.

Across four studies, we use favorability and attitude importance toward minoritized racial out-groups to predict engagement in costly behaviors concerning those out-groups. Note that the central behavior of interest is always how much money people are willing to give up (either to affect the transmission of bias or to provide/withhold aid to a particular group).³⁰ Our first three studies investigate how much money people will give up to: prevent their social group attitudes from being misrepresented when informing others about “typical” attitudes or norms in the sample (Study 1); prevent individuals, with whom they expect to interact, from receiving a misrepresentation of their personal social group attitude (Study 2); or affect a minoritized racial

out-group directly through charitable giving or withholding (Study 3). Across all three studies, we found that increased attitude importance consistently predicted increased amounts of money offered up by participants to transmit their stated attitudes. The models in Studies 1-3 also include a third social group attitude measurement, attitude moral conviction relevance (i.e., how related an attitude is to a participant's sense of right and wrong ³¹), as well as covariates for respondents' overall motivation to respond without prejudice ³² oriented toward "minorities" in general. We include these measures to test the specificity of the importance measure (i.e., does any additional prejudice-related feature increase variance explained?) and its predictive capacity above and beyond multi-item, well-established measures of prejudice, respectively.

In a follow-up experiment (Study 4), we manipulated a theoretical antecedent of attitude importance, personal relevance^{33,34}, across both attitude importance and favorability measurements in a between-subjects design. This experiment was designed to test whether personal relevance might serve as a possible source of the predictive power of importance for costly behavior observed in Studies 1-3. We investigated this hypothesis by varying the degree of personal relevance elicited by the attitude measurement in both favorability and importance measurements across four between-subjects conditions. Our findings from Study 4 indicate that the effect of attitude measurements on costly behavior (i.e., charitable donations affecting the out-group) is driven by personal relevance, an effect that applied to both favorability and importance measurements.

We conducted three studies and one experiment (Studies 1-4). After we collected and analyzed Studies 1 and 2, we preregistered Study 3 on the Open Science Framework (OSF; https://osf.io/a5uhp?view_only=60e1327bd698404d8ff6cf48fe8661e2). Study 4 was designed to follow up on Studies 1-3, and was also preregistered on OSF

(https://osf.io/j9xqf/?view_only=29cc2b5f8d464620b63ccb91c893911c). All procedures were approved by the university's IRB. All data, data analysis code, and materials are available for download at the following URL:

https://osf.io/3ujfg/?view_only=b29d9884cc9a44af94b6bb6dc448db2e.

Note that the analysis reported here for Study 3 deviates slightly from the analysis implied in the preregistration: regression modeling in Study 3 was computed under a Bayesian inferential framework rather than the frequentist framework implied by the preregistered power analysis. This distinction is spelled out explicitly in the preregistration for Study 4, where again the power analysis was computed for an equivalent frequentist model, even though we intended to compute the actual analyses under a Bayesian framework. The frequentist concept of statistical power does not apply to Bayesian models, but it is a useful benchmarking process for determining reasonable sample sizes, hence our usage of frequentist power analyses in the preregistrations for Studies 3-4.

The reason we used Bayesian regression models for Studies 3-4 was to match the inferential framework used in Studies 1-2. Note that the pattern of results across studies does not differ between Bayesian and frequentist versions of the models. Bayesian modeling improved model estimation for the complex multilevel ordinal regression models we fit in Studies 1 and 2, following the tutorial provided by Bürkner and Vuorre (2018)³⁵. The frequentist estimators for complex generalized mixed-effects models (especially cumulative link models, which we used here) often face logistical hurdles during estimation, namely convergence errors requiring manual optimizer adjustments. By contrast, Bayesian estimation of these models circumvents issues with convergence, allowing for a more general and reproducible modeling framework, which we could use consistently across models and analyses.

For all four studies, all measures, manipulations, and exclusions are disclosed. All study participants gave their informed consent before participating, and all procedures were conducted in accordance with the guidelines of the Committee on the Use of Human Subjects at the corresponding author's institution. We aimed for 150 usable participants in Studies 1 and 2 (all groups are rated within subjects) based on estimates of correlation stability by Schönbrodt and Perugini (2013; see Table 1), where $N = 150$ indicates the point of stability for a correlation coefficient of $\rho = .2$ with 90% confidence on a half-width corridor stability of .15³⁶ (We fell short of $N = 150$ after exclusions in Study 1, but exceeded that threshold in Study 2.) For Study 3, we aimed for 330 usable participants per the preregistered power analysis (discussed above and detailed in Study 3 Methods). For the between-group experimental manipulation in Study 4, we aimed for a minimum of 300 usable participants per condition across four conditions, for a total of $N = 1200$ (based on the preregistered power analysis discussed above and detailed in Study 4 Methods). In all cases sample size was determine prior to data collection and analysis and all samples were restricted to participants located in the U.S. Finally and critically, although we use favorability as our baseline attitude measurement predictor that incorporates both valence and extremity throughout the paper (i.e., goes from negative to positive favorability), our results show the same patterns when we instead use the *absolute value* on the favorability scale (i.e., pure attitude extremity such that higher values indicate more extremity, irrespective of whether the attitude is positive or negative). All analyses using extremity instead of favorability are reported in Supplementary Materials (Section 6).

Study 1: How much are people willing to pay to prevent transmission of prejudice?

In Study 1, we investigated how much participants would be willing to pay to prevent their attitudes about minoritized racial out-groups from being misrepresented to unknown future

participants. Specifically, we informed participants that we might use their favorability ratings as examples of “representative responses” from the study to show to future participants, but that there was a risk we might misreport their attitudes in the process, reversing the valence and extremity before reporting them. We hypothesized that participants would be willing to pay more to prevent this misrepresentation for attitudes that were more important to them, not just for attitudes that were more (un)favorable.

Methods

All data, analysis code, and materials are publicly available on OSF at:

https://osf.io/3ujfg/?view_only=b29d9884cc9a44af94b6bb6dc448db2e

Participants. We recruited 299 participants via Amazon Mechanical Turk for financial compensation as well as a one-dollar bonus. We excluded 112 participants who failed attention checks ($n = 30$ failed general attention checks; $n = 50$ failed task-specific comprehension checks; see Procedure for details) or who reported attitudes that were below the extremity threshold necessary for completing the willingness-to-pay task described below ($n = 32$). This yielded a usable total sample of $N = 187$ participants. From there, because we were interested in attitudes toward minoritized U.S. racial out-groups, we chose to focus on monoracial White (MRW) respondents evaluating non-White out-group targets. MRW respondents made up the largest portion of the sample that could be used to efficiently analyze data on multiple minoritized racial out-group attitudes (evaluating “Black people in the U.S.,” “Asian people in the U.S.,” and “Latinx people in the U.S.” as out-group targets). Thus we analyzed a final sample of MRW respondents evaluating out-group targets with $N_{MRW} = 137$ participants.

Procedure. Participants rated their favorability toward four social groups in the United States in random order using a sliding scale from -10 (“Extremely unfavorable”) to 10

(“Extremely favorable”): Black people, Latinx people, Asian people, and White people. We specifically used the language of intergroup attitude assessment which is used in the social sciences and polling data (e.g., Gallup) precisely because it is so widely reported and shared as the basis for crafting policy and strategy.

Participants then used a Becker-DeGroot-Marschak (BDM) mechanism³⁷ to indicate their willingness-to-pay (WTP) to prevent the presentation of distorted attitude norms (see Table 1 for an example). In the BDM assessment, participants considered multiple possible costs of protecting their attitudes from distortion (explained in detail below). Participants were told that one of the outcomes on the table would be selected at the end of the task, and that their choice for that outcome (either paying the stated cost or allowing their attitude to be distorted) would be realized. Consequently, participants’ choices on the BDM task can be considered to reflect the losses they would be willing to tolerate in exchange for attitude preservation—in other words, a monetary indicator of attitude-behavior consistency. Ultimately, we compensated all participants the full bonus amount, performed no misrepresentation, and informed all participants about the minor deception.

Before engaging in the task, participants completed a thorough comprehension check for using the BDM tables with a maximum of two attempts. Participants who failed both comprehension checks were excluded from analysis ($n = 50$).

Table 1: Example Becker-DeGroot-Marschak Table for Studies 1 and 2

	I would prefer to keep my attitude rating unchanged and pay the stated cost	I would prefer to allow my attitude rating to be changed and keep the full bonus payment
Keep my attitude rating unchanged at a cost of \$1.00		X
Keep my attitude rating unchanged at a cost of \$0.90		X
Keep my attitude rating unchanged at a cost of \$0.80		X
Keep my attitude rating unchanged at a cost of \$0.70		X
Keep my attitude rating unchanged at a cost of \$0.60		X
Keep my attitude rating unchanged at a cost of \$0.50	X	
Keep my attitude rating unchanged at a cost of \$0.40	X	
Keep my attitude rating unchanged at a cost of \$0.30	X	
Keep my attitude rating unchanged at a cost of \$0.20	X	
Keep my attitude rating unchanged at a cost of \$0.10	X	

Table 1: Example BDM table completed for a hypothetical participant indicating that they would be willing to pay as much as \$0.50 to prevent their attitude rating from being misrepresented.

We told participants who successfully completed the attention checks that their favorability ratings toward each social group might be shown to future participants as representative example responses. However, we indicated that in doing so, we might *misreport* their reported favorability by shifting it 12 points in the *opposite* direction. For example, if a participant rated that they felt favorably toward Latinx people in the U.S. with an attitude rating of 5, we indicated we might tell future participants that a typical respondent felt negatively

toward Latinx people in the U.S. with an attitude rating of -7. The 12 point difference was used to allow reversals of both valence and extremity—for example, a strongly positive attitude would be distorted into a weakly negative attitude. This kept the degree of “treatment” constant across participants. Participants could then give up some of their bonus payment to *prevent* us from misrepresenting their attitude (i.e., communicating norms counter to their position), using the BDM table to specify their preferences on the matter.

Participants filled out, in random order, a BDM table for each of the social groups for which they had given a favorability rating greater than 1 (for favorable ratings) or less than -1 (for unfavorable ratings). Participants who reported favorability ratings of -1, 0, or 1 for any target group ($n = 32$) were not given BDM tables to fill out for that group, because these rating values were not compatible with the hypothetical misrepresentation (altering a value of -1 or 1 by 12 points would place the final value beyond the -10 to 10 scale, and a value of 0 has no “opposite direction” in which to be altered).

Each BDM table gave tradeoff options in 10-cent increments, where participants indicated their preference as to whether they would rather: A) give up the stated amount from their bonus in order to prevent their attitude from being misrepresented; or B) allow their attitude to be misrepresented and keep the entire bonus amount. We told participants that at the end of the study, we would randomly select one of the tradeoff options across all of the BDM tables they filled out and award their bonus pay according their preferences for the selected options.

After the BDM portion of the study, participants answered two questions in random order: 1) “How important to you is your attitude toward [social group]?” (0 = “Not at all important”, 10 = “Extremely important”); and 2) “How deeply is your attitude toward [social group] connected to your fundamental beliefs about right and wrong?” (0 = “Not at all

connected”, 10 = “Very deeply connected”). We took the language of our moral relevance item from work by Skitka and colleagues (see Skitka et al., 2021 for a comprehensive review ³⁸) with minor adjustments to make the wording as parallel as possible to our importance item. We thought this would be the most conservative approach to comparing their influence on intergroup behavior. Note also that Skitka and colleagues have published papers with single item measures of moral relevance ³¹.

Then participants completed a modified version of the internal and external motivation to respond without prejudice scale (IMS and EMS ³²), adjusted to ask about “minorities” in general instead of a specific target group (i.e., Black people in the original scale items). Because one could make the case that our attitude valence measures were psychometrically weak (i.e., assessed via single items) we included IMS/EMS as a more psychologically-informed, psychometrically validated measure of prejudice. IMS and EMS items were presented on a single page using a 7-point Likert scale, with items presented in random order between participants. (See Supplementary Materials for a complete list of the items used.)

At the conclusion of the study, participants answered an attention check question (“How many fatal heart attacks have you had?”) and reported their political affiliation and demographic information. We did not analyze political affiliation or demographic information here except to use participants’ self-identified race to identify the MRW respondents for our analyzed sample.

Analysis. The primary analysis for Study 1 was a Bayesian hierarchical proportional odds ordinal regression model using a cumulative link function, computed in R ³⁹ using the “brms” package ^{35,40} with default (uninformative) priors and 10,000 sampling iterations. This model estimated the likelihood of each WTP threshold value using the following parameters as predictors: 1) the out-group favorability rating, 2) subjective attitude importance, 3) attitude

moral conviction relevance, 4) average IMS score, and 5) average EMS score. We included varying (“random”) intercepts associated with participant ID since many respondents completed BDM tables for multiple out-group targets.

Results and Discussion

Attitude importance is a better predictor of willingness-to-pay to prevent misrepresentation of norms than favorability. To assess the relationship between social group attitudes and WTP behavior in Study 1, we computed a Bayesian hierarchical multiple ordinal regression model, where each attitude feature was included as a predictor in a model predicting the cumulative likelihood of increased WTP (see Methods for details). Across all three studies, models including all two-way interactions between attitude feature variables (favorability, importance, and moral conviction relevance) were additionally fit, but no interaction models yielded substantive improvements to model fit (see Supplementary Materials Section 5 for details). Thus, we retained the simpler (no-interaction) models reported here.

Results from the ordinal regression model for Study 1 (Table 2) indicated that participants were willing to pay more to prevent attitude misrepresentation during future norm signaling for attitudes that were more important to them. Increased attitude importance was the strongest predictor of WTP with a positive conditional effect, predicting an increased WTP likelihood (posterior mean $\beta = 1.01$, 95% Bayesian uncertainty interval UI = [0.60 1.44], SE = 0.21). Taking the exponent of this estimate provides the following odds ratio interpretation: for each one point increase in attitude importance (holding the other predictors constant), the likelihood that someone was willing to pay more money to prevent distorted norm signaling was approximately 2.7 times greater. With attitude importance included in the model, increasing favorability ratings did not positively predict WTP values (posterior mean $\beta = -0.27$, 95% UI = [-

0.45, -0.09], SE = 0.09). The moral relevance predictor's 95% uncertainty interval also did not contain 0, however the effect was smaller and negative (note also that this effect did not replicate in the next two studies). We observed a very similar pattern of results when we examined the model that included attitude importance and attitude extremity (the absolute value of the valence rating) in the model instead of favorability (see Supplementary Materials Section 6, Table S10 and Figures S16-S17).

Table 2: Study 1 Bayesian ordinal regression parameter estimates predicting WTP

Effect	Estimate	SE	l-95% UI	u-95% UI
Favorability	-0.27	0.09	-0.45	-0.09
Importance	1.01	0.21	0.6	1.44
Moral Conviction	-0.36	0.16	-0.67	-0.07
IMS.Avg	0.91	0.71	-0.41	2.38
EMS.Avg	0.43	0.58	-0.71	1.6

Table 2: Regression coefficient estimates for the slope effects from the Bayesian ordinal regression model predicting WTP in Study 1. Estimates are posterior means for each slope parameter in the model, and they are reported alongside standard errors and Bayesian 95% uncertainty interval (UI) lower (l-) and upper (u-) bounds. Note that a logit link function was used in the estimation of the ordinal models, so parameter estimates should be interpreted accordingly (see main text for exponentiated parameter interpretations for parameters of interest). To underscore differences in inferential statistics for these estimates, effects in Table 2 have been bolded wherever the 95% UI for the effect does not contain 0. For complete plots of the posterior distributions for each effect see Supplemental Materials (Figure S3).

The effects plots from Study 1 are shown in Figure 1, beginning with descriptive visualizations of the conditional distributions of WTP values by favorability and importance observed in the raw data (Figure 1A-B). Model results are visualized in Figure 1C-D using effects plots that show the predicted probabilities for each WTP value based on favorability and importance respectively. (Complete descriptive results for all studies are reported in Supplementary Materials, Table S1-S2, and detailed regression tables, model diagnostics, and complete effects plots for the Study 1 regression model are reported in Table S3 and Figures S2-S4.)

Figure 1: As attitude importance increased, the probability of greater WTP values increased for preventing the misrepresentation of prejudice norms. Top row (A-B) shows descriptive visualizations of the distributions of WTP values by favorability (A) and importance (B). Bottom row (C-D) shows the model-based predictions for WTP values as a function of favorability (C) and importance (D), with effects computed jointly in the same model. As shown in C, predicted probabilities of increased WTP values did not increase as a function of increased favorability ratings when importance was also included in the same model.

Study 2: How much are people willing to pay to prevent prejudice misrepresentation to future interaction partners?

The purpose of Study 2 was to replicate and extend the results from Study 1 in a context with more concrete interpersonal stakes. We told participants that we would share their self-reported social group attitudes with another participant with whom they would interact directly in an online chatroom. Critically, they could give up real money to ensure we represented their

true attitudes in their chatroom profile, which would be viewed other chatroom members.

Participants were not ultimately placed into a chatroom, and they were compensated their full bonus amounts and informed of the minor deception.

Methods

Participants. We recruited 406 participants for Study 2 via Amazon Mechanical Turk to ensure we would have at least 150 usable participants after exclusions. Exclusion criteria were identical to those used in Study 1, and resulted in the exclusion of 144 participants ($n = 61$ failed both BDM comprehension checks; $n = 43$ had favorability ratings of -1, 0, or 1 for any target group; $n = 40$ failed general attention checks). This yielded a total usable sample of $N = 262$ participants. After selecting for MRW participants evaluating non-White out-group targets, our final analyzed sample was $N_{MRW} = 186$ participants.

Procedure. The procedure was identical to the procedure of Study 1 except for one deviation: we told participants that after they were done with filling out the BDM table, we would then put them in a chat room with other survey participants to discuss race relations in America. Participants were informed that the other participants in the chat would be able to see their attitude ratings or our misrepresentations of their attitude ratings (“In order to make sure your attitude rating is reported accurately in your chat room profile, you can sacrifice a portion of your \$1.00 bonus”).

Analysis. Analysis for Study 2 proceeded exactly as for Study 1.

Results and Discussion

Attitude importance outperforms favorability predicting willingness-to-pay to prevent misrepresentation of attitudes to future conversation partners. Results from the ordinal regression model for Study 2 replicated results from Study 1 albeit with a smaller effect,

indicating that participants were willing to pay more to prevent attitude misrepresentation to future conversation partners for attitudes that were more important to them (Table 3). Increased attitude importance was again the strongest positive predictor of WTP in the Bayesian hierarchical cumulative link model, predicting increased WTP likelihood (posterior mean $\beta = 0.47$, 95% UI = [0.19, 0.77], SE = 0.15). Taking the exponent of this estimate provides the odds ratio interpretation that for each one point increase in attitude importance (holding all other predictors constant), the likelihood that someone was willing to pay more money to prevent reputation distortion was approximately 1.6 times greater. As in the case of Study 1, increased favorability ratings did not positively predict unique variance in WTP values (posterior mean $\beta = -0.15$, 95% UI = [-0.32, 0.01], SE = 0.08) above and beyond attitude importance. All other predictors' 95% uncertainty intervals contained 0. Again, we observed a very similar pattern of results when we examined the model that include attitude importance and attitude extremity (absolute value of the valence rating) in the model instead of favorability (see Supplementary Materials Section 6, Table S11 and Figures S18-S19).

Table 3: Study 2 Bayesian ordinal regression parameter estimates predicting WTP

Effect	Estimate	SE	l-95% UI	u-95% UI
Favorability	-0.15	0.08	-0.31	0.01
Importance	0.47	0.15	0.19	0.77
Moral Conviction	0.12	0.13	-0.13	0.39
IMS.Avg	0.46	0.71	-0.9	1.86
EMS.Avg	0.86	0.68	-0.49	2.23

Table 3: Regression slope coefficient estimates for the Bayesian ordinal regression model predicting WTP in Study 2. Estimates are posterior means for each slope parameter in the model, and they and are reported alongside standard errors and Bayesian 95% uncertainty interval (UI) lower (l-) and upper (u-) bounds. As in Study 1 (Table 2), effects in Table 3 have been bolded wherever the 95% UI for the effect does not contain 0. For complete plots of the posterior distributions for each effect see Supplemental Materials (Figure S6).

Descriptive visualizations of the raw data from Study 2 (conditional distributions of WTP values by favorability and importance) are shown in Figure 2A-B, and effects plots with predicted WTP value probabilities for Study 2 are shown in Figure 2C-D. (Detailed regression tables, model diagnostics, and complete effects plots for Study 2 are reported in Supplementary Materials Section 4, Table S4 and Figures S5-S7.)

Figure 2: As attitude importance increased, the probability of greater WTP values increased for preventing reputation distortion. As in Figure 1, top row (A-B) shows descriptive visualizations of the conditional distributions of WTP values by favorability (A) and importance (B). Bottom row (C-D) shows the model-based predictions for WTP values as a function of favorability (C) and importance (D), with effects computed jointly in the same model. As in Study 1, predicted probabilities of increased WTP values in Study 2 did not increase as a function of favorability ratings when importance was also included in the same model.

Study 3: How much are people willing to pay to help or harm the out-group directly?

The purpose of Study 3 was to test the generalizability of the results from Studies 1 and 2 in a context involving real-world consequences directly for a minoritized racial out-group (as opposed to oneself). Unlike in Studies 1-2, the costly behavior in Study 3 was a charitable donation task for a specific minoritized racial out-group (Asian Americans), where participants' donations were realized as actual monetary loss (i.e., they gave up real money from their bonus), and the money they gave up was incorporated into an actual donation to a charity supporting that group.

Critically, to address the concern that WTP evaluations made before importance judgments might have influenced ratings on the importance scale in Studies 1-2, we randomized the order of the key measurements in Study 3: charitable donation and attitude importance. We did not observe any effect of measurement order on the relationship between donation and importance (see Supplementary Materials Section 4, Table S7 and Figures S12-S13).

Methods

Participants. In order to estimate sample size for Study 3, we used a preregistered power analysis as a guide, based on effect size of Cohen's $f^2 = .04$, $\alpha = .05$, and 80% power (estimated from the results of Studies 1 and 2). This analysis provided a benchmark sample size of 330 MRW participants or more for Study 3. It is important to note that this power analysis was computed for a frequentist inferential framework, whereas the actual analysis was performed using Bayesian inference—a framework which does not have the same concept of statistical power. Statistical power is defined based on a frequentist conception of probability (i.e., the probability of observing data given a theoretical replication of an experiment many times and under the assumption that the null hypothesis is true), which does not apply Bayesian statistics.

This sample size determination for Study 3 is therefore considered as a benchmark here based on an analogous Frequentist regression model, rather than a strict constraint.

We collected 623 participants from Prolific Academic, out of which 60 participants were excluded for failing the attention checks ($n = 4$) and made non-zero attitude ratings ($n = 56$). This yielded a total usable sample of $N = 563$ participants, which we subsetting to a final analyzed sample of $N_{MRW} = 393$ after selecting for MRW participants.

Procedure. In this study, participants only rated their favorability toward one minoritized U.S. racial group, Asian people in the United States, using the same scale from the previous studies. Then, three items were presented in random order: 1) a charitable donation opportunity; 2) the attitude importance question from the previous studies; and 3) the moral conviction relevance question from the previous studies.

For item 1 above, participants who reported a positive attitude toward Asian people in the United States were given the opportunity to give up money from their \$0.50 bonus pay in order to *add* that amount of money to a charitable donation being made to an organization supporting Asian people in the United States (the target group). Participants who reported a negative attitude were instead given the opportunity to give up money from their bonus pay in order to *subtract* that amount of money from the charitable donation being made. The final donation pool, after summing across both additive and subtractive donations, was donated to the indicated charity after the study was completed. After participants completed items 1-3, Study 3 proceeded identically to the remainder of Studies 1 and 2.

Preregistered analyses. The primary analysis for Study 3, which was preregistered on OSF (https://osf.io/a5uhp/?view_only=60e1327bd698404d8ff6cf48fe8661e2), was a linear regression model predicting the amount of money given up (either to add to or subtract from the

charitable donation) as a function of the measured variables from Studies 1 and 2 (i.e., extremity, importance, moral relevance, average IMS, and average EMS). Though not specifically preregistered as a Bayesian model (preregistration included a frequentist power analysis to determine sample size, as discussed previously), we fit a Bayesian model here to maintain a consistent inferential framework across all the present studies. The model was fitted as a Bayesian Gaussian regression model using the “brms” package with the default (uninformative) priors and 10,000 sampling iterations.

As an additional preregistered exploratory analysis, we used multidimensional scaling (MDS; see ⁴¹ to explore the relationships among the variables from Studies 1-3 separately from the willingness-to-pay response. MDS provided an exploratory tool that showed the relationships between the measured variables were consistent across studies. Details and results of this analysis are reported in Supplemental Materials, Section 3 (Figure S1).

Results and Discussion

Attitude importance better predicts charitable donations than favorability. Results from the Gaussian regression model in Study 3 indicated that participants were willing to pay more to directly affect Asian people in the U.S. when their attitude toward Asian people was more important to them, a finding consistent with the results of Studies 1 and 2 (Table 4). The Gaussian model predicting increased donation behavior from favorability, attitude importance, attitude moral conviction relevance, average IMS, and average EMS found attitude importance to be a strong positive predictor of increased donation behavior (posterior mean $\beta = 1.33$, 95% UI = [0.59, 2.06], SE = 0.37). In other words, a one-point increase in attitude importance (holding the other predictors constant) predicted an increase in donation behavior of 1.26 cents out of the 50-cent bonus, or about a 2.5% increase. The strength of this predictor is illustrated clearly using

standardized beta parameters (standardized posterior median $\beta_{\text{std.}}$) as a measure of effect size:

attitude importance std. posterior median $\beta_{\text{std.}} = .21$ (95% UI [0.09, 0.32]).

Table 4: Study 3 Bayesian regression parameter estimates predicting donation behavior

Effect	Estimate	SE	l-95% UI	u-95% UI
Favorability	0.67	0.34	-0.01	1.34
Importance	1.33	0.37	0.59	2.06
Moral Conviction	-0.27	0.31	-0.87	0.33
IMS.Avg	3.6	1.02	1.57	5.58
EMS.Avg	-0.2	0.66	-1.49	1.09

Table 4: Regression slope coefficient estimates for the Bayesian Gaussian regression model predicting donation behavior in Study 3. Estimates are posterior means for each slope parameter in the model, and they and are reported alongside standard errors and Bayesian 95% uncertainty interval (UI) lower (l-) and upper (u-) bounds. The response variable in this model is absolute-valued donation behavior—in other words, the amount of money participants gave up from their bonus, whether it was to add to or subtract from the donation made. Effects in Table 4 have been bolded wherever the 95% UI for the effect does not contain 0. For complete plots of the posterior distributions for each effect see Supplemental Materials (Figure S8).

The only other notable predictor in this model was participants' average internal motivation to respond without prejudice score (IMS; posterior mean $\beta = 3.60$, 95% UI = [1.57, 5.58], SE = 1.02), a respondent-specific covariate which had a similar effect size to the importance variable (std. posterior median $\beta_{\text{std.}} = .19$, 95% UI = [0.08, 0.30]). The attitude favorability and importance effects are plotted in Figure 3. (Complete regression results for Study 3, including model diagnostics, all effects plots, and an alternative Bayesian model using zero-one-inflated beta regression are reported in Supplementary Materials Section 4, Tables S5-S6 and Figures S8-S11.)

Figure 3: As attitude importance increased, donations to directly affect the out-group increased. Donation behavior also increased as a function of increased favorability, but to a lesser degree. 95% Bayesian uncertainty interval bands are plotted for each regression line (shaded regions). Raw data points are overlaid on each plot showing observed donation behavior by favorability (left) and importance (right) respectively.

Study 4: Does personal relevance drive the attitude importance effect on costly behavior?

Directly manipulating attitude importance is difficult or impossible to do in a controlled experimental context^{34,42}. This is not because the importance of an attitude is assumed to be unchangeable; rather, to change the importance of an attitude to an individual requires complex, potentially protracted efforts at persuasion, which is difficult to introduce in an experimental paradigm without also introducing several confounds. Following the guidance of Wegener et al.

(1995)³⁴, we turned to the political attitudes literature to identify which theoretical *antecedent* of attitude importance likely imbues it with predictive power for costly intergroup behavior.

In Study 4, we manipulated the salience of the *personal relevance* of the attitude in question^{33,42,43}. Specifically, we tested whether framing an attitude measure—either importance or favorability—in terms of its personal relevance to the participant influenced the degree to which that attitude measure predicted charitable donation behavior. Personal relevance is a broad construct that may reflect determinants including self-interest, social identification with the relevant group, personal goals or values, or some combination thereof^{33,34}. Given how many different inputs there are to personal relevance, we operationalized it in a global manner: by including “TO YOU” in the measurement item.

In Study 4, participants were randomly assigned to report their attitude importance or favorability toward the target group with an item that explicitly highlighted (or not) the personal relevance of their position. Our 2 (attitude measure: favorability or importance) x 2 (framing: personally relevant vs. generic) between-subjects design experimentally isolated the effect of personal relevance in comparing attitude importance and favorability for predicting costly behavior. We predicted that personal relevance would increase both favorability and importance items’ relationship to donation behavior (relative to the same items where we did not emphasize the personal relevance of the attitude). Moreover, we predicted that the personal importance condition would show the strongest relationship between item measurement and donation behavior relative to the other three conditions.

Methods

Participants. As in Study 3, we benchmarked our sample size for Study 4 via a preregistered power analysis for a frequentist model containing an interaction effect. As in all

previous analyses, we still compute a Bayesian regression model for our analysis in Study 4, but we use the frequentist power analysis to approximate the necessary sample size. (A complete explanation of this approach to sample size estimation is provided in the text of the preregistration, and R code for the power simulation is included in the OSF project repository.)

The power analysis determined that a sample size of 300 participants per group would be sufficient to detect an interaction effect size of correlation coefficient $r = .2$ with at least 90% power in an analogous Frequentist regression model. The simulation used correlations of $r = .2$ between each predictor and the donation response, and assumed independence between the predictors (because in this experiment, condition is randomly assigned). 1000 simulations were performed for a sequence of sample sizes between $N = 50$ and $N = 300$ in intervals of 50. After identifying $N = 300$ as a candidate sample size per group for power of at least 90%, the power simulation was recomputed for just $N = 300$ with 10,000 iterations for a more precise estimate. Final estimated power for $N = 300$ per group was 94.61%.

We collected data from 1833 participants via Prolific Academic between May 4 and May 10 of 2022. Due to a bug in early data collection, 7 participants were accidentally allowed to participate more than once in the experiment, and all observations from these repeat participants were excluded entirely from the final analysis ($n = 14$ observations). One additional participant ($n = 1$) was excluded for completing the survey from outside the United States (their IP address indicated an international location during survey completion). No participants were excluded for failing the attention check. This yielded a total usable sample of $N = 1818$ participants, which we subsetting to a final analyzed sample of $N_{MRW} = 1354$ after selecting for MRW participants. The final distribution of MRW participants into each condition of the experiment was as follows:

Attitude Importance: Generic, $N_{MRW} = 322$; Attitude Importance: Personal, $N_{MRW} = 353$;

Favorability: Generic, $N_{MRW} = 335$; Favorability: Personal, $N_{MRW} = 344$.

Procedure. Study 4 consisted of four between-subject experimental conditions to which participants were randomly assigned. In this study, as in Study 3, all participants only evaluated one minoritized U.S. racial group: Asian people in the United States. The *generic* favorability item read, "Consider your attitude toward Asian people in the United States. How favorable or unfavorable are Asian people in the U.S.?" where the response scale ranged from -10 ("Extremely unfavorable") to 10 ("Extremely favorable"). By contrast the *personal* favorability item read, "Consider your attitude toward Asian people in the United States. How favorable or unfavorable TO YOU are Asian people in the U.S.?" where the response scale ranged from -10 ("Extremely unfavorable to me") to 10 ("Extremely favorable to me").

The *generic* attitude importance item read, "Consider your attitude toward Asian people in the United States. How important is your attitude toward Asian people in the U.S.?" where the response scale ranged from 0 ("Not at all important") to 10 ("Extremely important"). By contrast the *personal* attitude importance item read, "Consider your attitude toward Asian people in the United States. How important TO YOU is your attitude toward Asian people in the U.S.?" where the response scale ranged from 0 ("Not at all important to me") to 10 ("Extremely important to me").

After the responding to one of the four attitude items, each participant then completed the donation task (a modification of the task from Study 3). This task consisted of making a donation on a sliding scale from -50 cents to +50 cents. If a participant chose a positive scale value (between +1 and +50 cents), the chosen amount of money was deducted from their bonus pay and *added* to a charitable donation supporting the target out-group. If a participant chose a

negative scale value (between -1 and -50 cents), the chosen amount of money was deducted from their bonus pay and *subtracted* from a charitable donation supporting the target out-group. If a participant chose 0 cents on the scale, no money was deducted from their bonus pay and the total charitable donation was unaffected. (Once data collection was complete we computed the total charitable donation amount made the corresponding donation to the stated charity.)

After the donation task, participants completed demographic items (so that we could subset the sample to monoracial White respondents) and attention checks, as in Studies 1-3.

Preregistered analyses. The primary analysis for Study 4, which was preregistered on OSF (https://osf.io/j9xqf/?view_only=29cc2b5f8d464620b63ccb91c893911c), was a Bayesian linear regression model predicting the amount of money participants gave up (either to add to or subtract from the charitable donation) from the interaction between the condition (generic favorability, personal favorability, generic importance, personal importance) and the absolute value of the measurement rating the participant made. We refer to this as the “condition-level” model because its level of analysis compares the individual conditions (a reference condition—generic favorability—vs. each of the three other conditions). Unlike in Studies 1-3, the model in Study 4 was constrained to the use of absolute values of donation behavior and favorability rating scales to ensure the different measured variables across conditions were compared on compatible scales. We conducted a Bayesian Gaussian regression model using the “brms” package with the default (uninformative) priors and 10,000 sampling iterations. The preregistered comparison of interest was between personal and generic items, and more specifically, each condition against the personal attitude importance condition (which we predicted would have the strongest relationship to donations); however, we also analyzed the full set of pairwise comparisons between conditions.

Additional analyses. We computed two extensions to the preregistered main analysis in order to more completely detail the information contained in the main model. These analyses merely refactor the results of the condition-level model to examine specific comparisons within the data.

The first of these extension analyses was a comparison of the full set of pairwise contrasts between conditions for the effect of increased scale value on donation behavior. These contrasts were computed and visualized directly from the condition-level model using the “emmeans” package ⁴⁴, which applies inferential corrections as needed for multiple comparisons (i.e., bias-adjusted Bayesian highest posterior density intervals, or HPDs).

Second, an alternative “factor-level” version of the main regression model was computed to compare the main effects of the 2x2 design more directly. This factor-level model was computed by refactoring the dataset with categorical variables indicating *relevance level* (generic vs. personal) and *measurement type* (attitude valence vs. attitude importance). This allowed us to analyze the results at the level of the experimental design’s two factors (relevance level and measurement type), rather than at the level of its four conditions. We explored the factor-level comparison using a 3-way interaction between measurement value, relevance level, and measurement type. Effects and plots for the factor-level model were computed using “emmeans”.

Results and Discussion

Personally relevant attitude measurements were more highly correlated with donation behavior. Consistent with our first preregistered hypothesis, results from both the condition-level and factor-level models in Study 4 show a stronger relationship between attitude measurement and donation behavior for personal relevance items compared to generic items (Table 6). The effect of personal relevance overall was substantial enough that the personally

relevant favorability condition performed as well as the personally relevant attitude importance condition. This result rules against our second preregistered hypothesis, where we predicted the personal importance condition would best predict donation behavior among the individual conditions.

Table 5 shows the pairwise contrasts from the “emmeans” package for each condition’s interaction with rating value to predict donation behavior in the preregistered condition-level regression model. The interaction effects plot for this model is shown in Figure 4A. Results indicate that the two *personal* conditions outperform the importance, generic condition at approximately the same level (but not favorability, generic): ([Importance, Generic] x Rating value vs. [Importance, Personal] x Rating value: posterior median $\beta = -1.18$, 95% HPD = [-2.21, -0.17]; [Importance, Generic] x Rating value vs. [Favorability, Personal] x Rating value: posterior median $\beta = -1.28$, 95% HPD = [-2.23, -0.34]).

Table 5: Study 4 condition-level pairwise contrasts for Bayesian regression predicting donation

Contrast	Estimate	l-95% HPD	u-95% HPD
[Favorability, Generic] > [Importance, Generic]	0.5123	-0.428	1.503
[Favorability, Generic] > [Importance, Personal]	-0.6673	-1.692	0.275
[Favorability, Generic] > [Favorability, Personal]	-0.7673	-1.664	0.132
[Importance, Generic] > [Importance, Personal]	-1.1794	-2.167	-0.139
[Importance, Generic] > [Favorability, Personal]	-1.2765	-2.169	-0.293
[Importance, Personal] > [Favorability, Personal]	-0.0994	-1.059	0.857

Table 5: Pairwise contrasts of regression slope coefficient estimates for the condition-level Bayesian Gaussian regression predicting donation behavior in Study 4. These effects are interactions between each condition and the rating value (the continuous numeric value participants provided in response to the attitude variable question they received). Ratings were absolute-valued to maintain comparability between favorability and attitude importance conditions. The response variable in this model is absolute-valued donation behavior: the amount of money participants gave up from their bonus, whether it was to add to or subtract from the donation made. Estimates for these contrasts come from the “emmeans” package, and show posterior medians for each contrast reported alongside adjusted Bayesian 95% Highest Posterior Density (HPD) lower (l-) and upper (u-) bounds. Effects in Table 5 have been bolded wherever the 95% HPD for the contrast does not contain 0.

We also observed the preregistered main overall effect that increased absolute ratings values (i.e., greater favorability or greater importance) on any of the conditions' attitude scales predicted donation behavior, controlling for these interactions (posterior mean $\beta = 0.8$, 95% UI = [0.13, 1.46], SE = 0.34; see Table S8). This corresponds to a standardized effect size std. posterior median $\beta_{\text{std.}} = .12$ (95% UI [0.02, 0.23]). Complete slope parameters from the condition-level regression model where effect estimates are compared to the reference condition (generic favorability) are reported along with model diagnostics in Supplemental Materials Section 4, Table S8, as well as posterior distribution plots in Figure S14.

One interpretation of these results is that removing the personal framing from the attitude importance measure weakens its ability to predict behavior compared to its typical framing (which evokes personal relevance, i.e., “importance TO YOU” in the preceding studies). Similarly, adding the personal framing to the favorability measure strengthens it such that it performs similarly to the personal importance measurement. Based on these observations, we computed the factor-level model to compare the *personal* and *generic* conditions grouped together (while also comparing the favorability and importance conditions grouped together). The results from the factor-level model are reported in Table 6, and interaction effects plots for the factor-level model are reported in Figure 4B-C. (Complete model diagnostics, detailed results, and posterior distribution plots for the factor-level model are available in Supplemental Materials Section 4, Table S9 and Figure S15.)

Figure 4: Attitude measurements, both importance and favorability, were more predictive of donation behavior when they were framed in terms of personal relevance. The condition-level model (A) shows the relative slopes for each condition in Study 4 predicting donation behavior from measurement rating values. The two personal relevance conditions performed best, showing the strongest relationships between measurement ratings and donation behavior. This difference is exposed more directly in the factor-level model (B-C). Comparing the two generic conditions to the two personal relevance conditions (B), the personal relevance conditions show a stronger relationship between measurement ratings and donation behavior. However, after controlling for personal relevance, the favorability and importance conditions do not show substantive differences from each other (C).

The primary result from the factor-level model is that even controlling for attitude measurement type (i.e., favorability vs. importance), the personal relevance conditions show a stronger positive relationship between measurement ratings and donation behavior than the generic conditions (posterior mean $\beta = 1.18$, 95% UI = [0.17, 2.19], SE = 0.52). This corresponds to a standardized effect size std. posterior median $\beta_{\text{std.}} = .18$ (95% UI [0.03, 0.34]).

Table 6: Study 4 factor-level Bayesian regression parameter estimates predicting donation

Effect	Estimate	SE	l-95% UI	u-95% UI
Rating value	0.29	0.36	-0.42	1
Personal > Generic	-3.29	3.39	-9.92	3.37
Favorability > Importance	-2.37	3.35	-8.96	4.25
Rating value x [Personal > Generic]	1.18	0.52	0.17	2.19
Rating value x [Favorability > Importance]	0.5	0.49	-0.47	1.46
[Personal > Generic] x [Favorability > Importance]	-1.73	4.67	-10.94	7.34
Rating value x [Personal > Generic] x [Favorability > Importance]	-0.41	0.69	-1.75	0.96

Table 6: Regression coefficient estimates for the factor-level Bayesian Gaussian regression predicting donation behavior in Study 4. For the factor-level model, the dummy-coded reference level for Relevance is “Generic” and the dummy-coded reference level for Measurement Type is “Attitude Importance”. Reference levels are written explicitly into the labels in Table 6 for maximal interpretability, since there are two categorical variables each with their own reference levels. The “Rating value” predictor corresponds to the continuous numeric value participants provided in response to the attitude variable question they received in their condition of the experiment. As in the condition-level model, ratings in the factor-level model were absolute-valued to maintain comparability between attitude valence and attitude importance measurements. The response variable in this model is the same absolute-valued donation behavior modeled in the condition-level model. Estimates are posterior means for each slope parameter in the model, and they are reported alongside standard errors and Bayesian 95% uncertainty interval (UI) lower (l-) and upper (u-) bounds. Effects in Table 6 have been bolded wherever the 95% UI for the effect does not contain 0. For complete plots of the posterior distributions for each effect see Supplemental Materials (Figure S15).

General Discussion

The results of the present studies indicate that across a variety of intergroup contexts, engagement in costly intergroup behaviors was best predicted by how much importance participants placed on their attitudes toward minoritized racial out-groups. Participants’

favorability ratings, the conventional direct measurement of their attitudes toward the target out-group, consistently mattered less for explaining variance in costly behavior than the importance of the attitude. Additionally, the results from Study 4 suggest that this difference is driven by the personal relevance that is invoked in the measurement of attitude importance. Removing personal relevance language from importance measurements weakens their predictive power for behavior, and adding personal relevance language to favorability measurements increases their predictive power for behavior.

The utility of improving predictive validity by adding a personally-relevant attitude importance item to prejudice questionnaires and surveys should not be underestimated. Across Studies 1-3, where models included multiple additive predictors of behavior, the one case where a competing predictor performed comparably to attitude importance was participants' average internal motivation to respond without prejudice (IMS) in Study 3. IMS is a participant-level summary statistic across five items from prejudice scale. It is therefore impressive that subjective attitude importance—a single, question-specific item—consistently predicted intergroup behavior as well as (or better than) a multi-item, participant-level scale across multiple studies. This result is particularly striking given the relative neglect of attitude importance measurement (and personal relevance language more generally) in large-scale social attitudes research, including in the largest social surveys administered in the United States—e.g., the General Social Survey (GSS) and the American National Election Study (ANES). Our findings here indicate that the inclusion of single-item personal attitude importance measures on these large-scale social surveys has the potential to improve our predictive power for intergroup behavior from social attitude information.

Moreover, results from Studies 3-4 where costly behaviors directly affected the target out-group indicate that conventional measures of attitude importance and favorability may explain different, unique variance in intergroup behavior. Specifically, results from Study 4 suggest that the personal relevance component of typical attitude importance measures captures the component of an attitude's strength value that best tracks engagement in donation behavior. By comparison, the standard practice in studies of such behaviors, which is to measure attitude extremity without a personalized measure such as attitude importance, may limit the explanatory power of measured attitudes for costly prosocial behavior toward minoritized groups.

More broadly, a better understanding of social group attitude importance has the potential to expand our understanding of when implicit and explicit attitudes toward minoritized racial out-groups will or will not correlate with one another. Karpinski and colleagues (2005)⁴⁵ found that attitude importance can moderate the implicit-explicit attitude relationship in political and commercial contexts, such that more important attitudes tend to produce greater correlations between implicit and explicit attitude measures. Although the present studies do not measure implicit attitudes, future work building upon our findings could help identify the role of the importance people place on their attitudes about social groups in implicit-explicit attitude convergence, to better understand the implicit and explicit structures of prejudice.

Finally, the present findings have important implications for the social justice movements of today, namely the ongoing struggle for justice for Black Americans and other marginalized groups in the United States. Since the murder of George Floyd in the summer of 2020, the issue of race-based inequality and violence has returned to the front and center in the United States. That summer alone, actions on behalf of racial justice manifested in streets across the country at a historic scale, with more than 10,000 peaceful protests taking place across all 50 states between

the death of George Floyd on May 25 and the end of the summer in August ⁴⁶. As action mounted in the streets, discussion regarding racial justice pervaded both social and commercial media. But not every Instagram post was accompanied by donations to bail funds that summer or in the months that followed. A groundswell of businesses and organizations issued statements affirming the value of Black lives and committing to anti-racism, but the same businesses did not necessarily demonstrate costly changes in policies or actions. The results we have presented here underscore that the work of equity and social justice cannot consist of rhetoric alone: Simply stating one's favorability toward racial justice is not necessarily a strong predictor that a person or an organization will engage in more *costly actions* to support minoritized groups. A high degree of personal importance on prosocial attitudes toward marginalized groups may be a better indicator of which individuals we can count on for donations, policy changes, and other actions that can greatly impact the fight against prejudice.

References

1. Allport, G. W. The Nature of Prejudice. *Am. Q.* **1**, 58–82 (1949).
2. Lavrakas, P. J. *Encyclopedia of Survey Research Methods*. (SAGE Publications, 2008).
3. Nelson, S. C. Feeling Thermometer. *Encycl. Surv. Res. Methods* 277 (2008).
4. McConahay, J. B. Modern racism, ambivalence, and the Modern Racism Scale. in *Prejudice, discrimination, and racism* 91–125 (Academic Press, 1986).
5. Pettigrew, T. F. & Meertens, R. W. Subtle and blatant prejudice in western Europe. *Eur. J. Soc. Psychol.* **25**, 57–75 (1995).
6. Swim, J. K., Aikin, K. J., Hall, W. S. & Hunter, B. A. Sexism and racism: Old-fashioned and modern prejudices. *J. Pers. Soc. Psychol.* **68**, 199–214 (1995).

7. Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. 17 (1998).
8. Macrae, C. N., Stangor, C. & Milne, A. B. Activating Social Stereotypes: A Functional Analysis. *J. Exp. Soc. Psychol.* **30**, 370–389 (1994).
9. Payne, B. K., Cheng, C. M., Govorun, O. & Stewart, B. D. An inkblot for attitudes: Affect misattribution as implicit measurement. *J. Pers. Soc. Psychol.* **89**, 277–293 (2005).
10. Hendricks, M. & Bootzin, R. Race and Sex as Stimuli for Negative Affect and Physical Avoidance. *J. Soc. Psychol.* **98**, 111–120 (1976).
11. Weitz, S. Attitude, voice, and behavior: A repressed affect model of interracial interaction. *J. Pers. Soc. Psychol.* **24**, 14–21 (1972).
12. Word, C. O., Zanna, M. P. & Cooper, J. The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *J. Exp. Soc. Psychol.* **10**, 109–120 (1974).
13. Kurdi, B. *et al.* Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *Am. Psychol.* **74**, 569–586 (2019).
14. Howe, L. C. & Krosnick, J. A. Attitude Strength. *Annu. Rev. Psychol.* **68**, 327–351 (2017).
15. Petty, R. E. & Krosnick, J. A. *Attitude strength: antecedents and consequences*. (Lawrence Erlbaum Associates, 1995).
16. Crandall, C., Eshleman, A. & O'Brien, L. Social Norms and the Expression and Suppression of Prejudice: The Struggle for Internalization. *J. Pers. Soc. Psychol.* **82**, 359–78 (2002).
17. Payne, B. K. *et al.* Implicit and explicit prejudice in the 2008 American presidential election. *J. Exp. Soc. Psychol.* **46**, 367–374 (2010).
18. Forscher, P. S. *et al.* A Meta-Analysis of Procedures to Change Implicit Measures. *J. Pers. Soc. Psychol.* **117**, 522 (2019).

19. Eaton, A. A., Majka, E. A. & Visser, P. S. Emerging perspectives on the structure and function of attitude strength. *Eur. Rev. Soc. Psychol.* **19**, 165–201 (2008).
20. Fazio, R. H. Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework. in *Advances in Experimental Social Psychology* (ed. Zanna, M. P.) vol. 23 75–109 (Academic Press, 1990).
21. Eaton, A. A. & Visser, P. S. Attitude Importance: Understanding the Causes and Consequences of Passionately Held Views. *Soc. Personal. Psychol. Compass* **2**, 1719–1736 (2008).
22. Kokkinaki, F. & Lunt, P. The relationship between involvement, attitude accessibility and attitude–behaviour consistency. *Br. J. Soc. Psychol.* **36**, 497–509 (1997).
23. Visser, P. S., Krosnick, J. A. & Simmons, J. P. Distinguishing the cognitive and behavioral consequences of attitude importance and certainty: A new approach to testing the common-factor hypothesis. *J. Exp. Soc. Psychol.* **39**, 118–141 (2003).
24. Visser, P. S., Krosnick, J. A. & Norris, C. J. Attitude Importance and Attitude-Relevant Knowledge. 43 (2016).
25. Holbrook, A., Berent, M. K., Krosnick, J. A., Visser, P. S. & Boninger, D. S. Attitude importance and the accumulation of attitude-relevant knowledge in memory. *J. Pers. Soc. Psychol.* **88**, 749–769 (2005).
26. Westerwick, A., Kleinman, S. B. & Knobloch-Westerwick, S. Turn a Blind Eye If You Care: Impacts of Attitude Consistency, Importance, and Credibility on Seeking of Political Information and Implications for Attitudes. *J. Commun.* **63**, 432–453 (2013).

27. Lavine, H., Borgida, E. & Sullivan, J. L. On the Relationship Between Attitude Involvement and Attitude Accessibility: Toward a Cognitive-Motivational Model of Political Information Processing. *Polit. Psychol.* **21**, 81–106 (2000).
28. Holbrook, A. L., Sterrett, D., Johnson, T. P. & Krysan, M. Racial Disparities in Political Participation Across Issues: The Role of Issue-Specific Motivators. *Polit. Behav.* **38**, 1–32 (2016).
29. Sellers, R. M., Rowley, S. A. J., Chavous, T. M., Shelton, J. N. & Smith, M. A. Multidimensional Inventory of Black Identity: A preliminary investigation of reliability and construct validity. *J. Pers. Soc. Psychol.* **73**, 805–815 (1997).
30. McGraw, A. P., Tetlock, P. E. & Kristel, O. V. The Limits of Fungibility: Relational Schemata and the Value of Things. *J. Consum. Res.* **30**, 219–229 (2003).
31. Skitka, L. J., Bauman, C. W. & Sargis, E. G. Moral Conviction: Another Contributor to Attitude Strength or Something More? *J. Pers. Soc. Psychol.* **88**, 895–917 (2005).
32. Plant, E. A. & Devine, P. G. Internal and external motivation to respond without prejudice. *J. Pers. Soc. Psychol.* **75**, 811–832 (1998).
33. Boninger, D. S., Krosnick, J. A. & Berent, M. K. Origins of attitude importance: Self-interest, social identification, and value relevance. *J. Pers. Soc. Psychol.* **68**, 61–80 (1995).
34. Wegener, D. T., Downing, J., Krosnick, J. A. & Petty, R. E. Measures and manipulations of strength-related properties of attitudes: Current practice and future directions. *Attitude Strength Antecedents Consequences* **4**, 455–487 (1995).
35. Bürkner, P.-C. & Vuorre, M. Ordinal Regression Models in Psychology: A Tutorial. (2018) doi:10.31234/osf.io/x8swp.

36. Schönbrodt, F. D. & Perugini, M. At what sample size do correlations stabilize? *J. Res. Personal.* **47**, 609–612 (2013).
37. Becker, G. M., Degroot, M. H. & Marschak, J. Measuring utility by a single-response sequential method. *Behav. Sci.* **9**, 226–232 (1964).
38. Skitka, L. J., Hanson, B. E., Morgan, G. S. & Wisneski, D. C. The Psychology of Moral Conviction. *Annu. Rev. Psychol.* **72**, 347–366 (2021).
39. R Core Team. R: A language and environment for statistical computing. (2020).
40. Bürkner, P.-C. brms: Bayesian Regression Models using ‘Stan’. (2019).
41. Borg, I. & Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*. (Springer-Verlag, 2005). doi:10.1007/0-387-28981-X.
42. Sherif, M. & Hovland, C. I. *Social judgment: Assimilation and contrast effects in communication and attitude change*. xii, 218 (Yale Univer. Press, 1961).
43. Petty, R. E., Cacioppo, J. T. & Haugtvedt, C. P. Ego-Involvement and Persuasion: An Appreciative Look at the Sherif’s Contribution to the Study of Self-Relevance and Attitude Change. in *Social Judgment and Intergroup Relations: Essays in Honor of Muzafer Sherif* (eds. Granberg, D. & Sarup, G.) 147–174 (Springer, 1992). doi:10.1007/978-1-4612-2860-8_7.
44. Lenth, R. V. emmeans: Estimated Marginal Means, aka Least-Squares Means. (2021).
45. Karpinski, A., Steinman, R. B. & Hilton, J. L. Attitude Importance as a Moderator of the Relationship Between Implicit and Explicit Attitude Measures. *Pers. Soc. Psychol. Bull.* **31**, 949–962 (2005).

46. Kishi, R. & Jones, S. *Demonstrations & Political Violence in America: New Data for Summer 2020*. <https://acleddata.com/2020/09/03/demonstrations-political-violence-in-america-new-data-for-summer-2020/> (2020).

Supplement to “Personal relevance of attitude importance predicts costly intergroup behavior”

Contents

1	Complete Study Descriptions	3
2	Descriptive Results	4
3	Exploratory MDS analysis	5
4	Complete Regression Model Details & Results	6
4.1	Study 1 Results	8
4.2	Study 2 Results	12
4.3	Study 3 Results	16
4.4	Study 4 Results	24
5	Checking for Interaction Effects	29
5.1	Study 1	29
5.2	Study 2	29
5.3	Study 3	29
6	Extremity analysis	30
6.1	Study 1 Extremity Model Results	30
6.2	Study 2 Extremity Model Results	33
6.3	Study 3 Extremity Model Results	35
7	Supplementary References	37

- - - -

List of Figures

S1	Multidimensional scaling of measured variables	6
S2	Study 1 descriptive visualizations of conditional WTP distributions	8
S3	Study 1 ordinal regression posterior distributions and trace plots	10
S4	Study 1 predicted WTP probability plots	11
S5	Study 2 descriptive visualizations of conditional WTP distributions	12
S6	Study 2 ordinal regression posterior distributions and trace plots	14
S7	Study 2 predicted WTP probability plots	15
S8	Study 3 Gaussian regression posterior distributions and trace plots	16
S9	Study 3 Gaussian model effects plots predicting donation behavior	17
S10	Study 3 zero-one-inflated beta regression posterior distributions and trace plots	19
S11	Study 3 zero-one-inflated beta effects plots predicting donation behavior	20
S12	Study 3 regression testing order effects: posterior distributions and trace plots	22
S13	Study 3 Gaussian model effects plots predicting donation behavior including potential order effect	23
S14	Study 4 condition-level regression posterior distributions and trace plots	25
S15	Study 4 factor-level regression posterior distributions and trace plots	27

S16	Study 1 extremity model posterior distributions and trace plots	31
S17	Study 1 extremity model predicted WTP probability plots	32
S18	Study 2 extremity model posterior distributions and trace plots	33
S19	Study 2 extremity model predicted WTP probability plots	34
S20	Study 3 extremity model posterior distributions and trace plots	35
S21	Study 3 extremity model effects plots predicting donation behavior	36

List of Tables

S1	Descriptive results from Studies 1-3	4
S2	Descriptive results from Study 4	4
S3	Study 1 Bayesian ordinal regression predicting WTP	9
S4	Study 2 Bayesian ordinal regression predicting WTP	13
S5	Study 3 Bayesian Gaussian regression predicting donation behavior	18
S6	Study 3 Bayesian zero-one-inflated beta regression predicting donation behavior	20
S7	Study 3 Gaussian regression predicting donation behavior including potential order effect . .	23
S8	Study 4 condition-level regression predicting donation behavior	26
S9	Study 4 factor-level regression predicting donation behavior	28
S10	Study 1 extremity model	30
S11	Study 2 extremity model	33
S12	Study 3 extremity model	35

1 Complete Study Descriptions

In Study 1, participants made ratings about their favorability toward each of four social groups in the U.S. (White people, Black people, Asian people, and Latinx people) on a scale from -10 (extremely unfavorable) to 10 (extremely favorable). After making their favorability ratings, they completed a Becker-DeGroot-Marschak (BDM) task, where they rated how much of a \$1 bonus they would be willing to give up in order to prevent us from misrepresenting their favorability by 12 points in the opposite direction (e.g., a rating of +8 being transformed to a rating of -4). this willingness-to-pay (WTP) value was our response variable of interest.

Note that participants were only able to complete the full study if they reported favorability ratings at least as extreme as -2 or +2, which is why there are no favorability values between -1 and 1 in either of these datasets.

The only difference between Study 1 and Study 2 was the framing of the BDM task. In order to raise the stakes of the task, the frame for Study 2 was that participants would enter a chatroom with other online participants. We asked how much of their \$1 bonus pay they would be willing to give up in order to prevent us from misrepresenting each of their stated favorability ratings toward target groups in their chatroom bio, which other participants would see.

For Study 3, Participants made favorability ratings toward just one racial group: Asian people in the U.S. After making their favorability ratings, they completed a donation task instead of the BDM task from Studies 1 and 2.

Donation Paradigm: participants were given the opportunity to sacrifice some of their \$0.50 bonus pay (from none of it to all of it) in order to affect a donation to a charity supporting Asian people in the U.S.: Asian Americans Advancing Justice. If the participant previously made a positive (> 0) favorability rating toward Asian people in the U.S., they were given the opportunity to ADD to the total donation being made to the charity. If the participant previously made a negative (< 0) favorability rating, they were given the opportunity to SUBTRACT from the total donation being made to the charity. Either the addition or the subtraction would be made in the exact amount participants opted to give up from their bonus pay. (Participants who made a favorability rating of 0 were not given this opportunity, and did not receive bonus pay.)

After the task in each study, we collected participants' responses to questions about their previously reported attitudes. (Note: for Study 3, the donation task was randomized along with the questions below, unlike in Studies 1 and 2 where the BDM task came before these ratings.) - Importance (0 to 10): "How important to you is your attitude toward _____ people in the U.S.?" - Moral Conviction (0 to 10): "How deeply is your attitude toward _____ people in the U.S. connected to your fundamental beliefs about right and wrong?"

We also administered a modified version of the internal and external motivation to respond without prejudice scale (IMS and EMS; Plant & Devine, 1998), adjusted to ask about minorities in general instead of a specific group. All items (below) were rated on a 7-point likert scale from "Strongly disagree" (1) to "Strongly agree" (7):

1. "Because of today's PC (politically correct) standards I try to appear nonprejudiced toward minorities."
2. "I try to hide any negative thoughts about minorities in order to avoid negative reactions from others."
3. "If I acted prejudiced toward minorities, I would be concerned that others would be angry with me."
4. "I attempt to appear nonprejudiced toward minorities in order to avoid disapproval from others."
5. "I try to act nonprejudiced toward minorities because of pressure from others."
6. "I attempt to act in nonprejudiced ways toward minorities because it is personally important to me."
7. "According to my personal values, using stereotypes about minorities is OK."
8. "I am personally motivated by my beliefs to be nonprejudiced toward minorities."
9. "Because of my personal values, I believe that using stereotypes about minorities is wrong."
10. "Being nonprejudiced toward minorities is important to my self-concept."

Note that for IMS and EMS items, items 1-5 are external (EMS) and 6-10 are internal (IMS). IMS and EMS average scores are in `IMS.Avg` and `EMS.Avg` columns. Also note that item 7 is worded for reverse-coding.

This is already accounted for in `IMS.Avg` (reverse-coding was done before averaging), but the data columns named `IMS.EMS_7` in each dataset contain the raw response values (i.e., prior to reverse-coding).

Finally, in Study 4, we attempted to manipulate one of the theorized antecedent features of attitude importance, personal relevance, in order to better characterize the distinction between the importance and favorability effects on donation behavior in Study 3. Four groups of participants were collected online via Prolific to form four between-subjects conditions with a 2[favorability vs. attitude importance] x 2[personal vs. generic] structure. The four conditions’ measurement questions were introduced with the phrased as follows:

“Consider your attitude toward Asian people in the United States...” - “...How favorable or unfavorable to YOU are Asian people in the United States?” (Favorability: Personal); - “...How favorable or unfavorable are Asian people in the United States?” (Favorability: Generic); - “...How important to YOU is your attitude toward Asian people in the U.S.?” (Attitude Importance: Personal); OR - “...How important is your attitude toward Asian people in the U.S.?” (Attitude Importance: Generic).

Participants saw exactly one of these questions depending on which condition they were in, and then subsequently completed the donation task. Unlike in Study 3, we gave all participants in Study 4 the option to either ADD TO or SUBTRACT FROM the donation, or to give up no money at all. Participants simply selected a value on a slider scale from -50 to +50, indicating how many cents they would like to give up and whether they would like the money to add to or subtract from the donation (indicated by a positive or negative scale value respectively).

For Study 4, no additional covariates were measured (e.g., moral conviction relevance, IMS/EMS), in order to keep the experiment parsimonious and focus on the attitude importance/favorability distinction vis-a-vis personal relevance.

2 Descriptive Results

Descriptive statistics for Studies 1-3 are reported in Table S1, and descriptive statistics for Study 4 are reported in Table S2. Note that for all studies, the analyzed participants are monoracial white (MRW) respondents evaluating non-white out-group targets.

Table S1: Descriptive results from Studies 1-3

Study	Target	N.MRW	WTP	Favorability	Importance	MoralConviction	IMS.Avg	EMS.Avg
1	Asian	133	16.77 (27.65)	6.55 (2.67)	5.67 (3.05)	5.56 (3.15)	NA	NA
1	Black	129	19.69 (30.23)	5.10 (4.57)	6.09 (3.02)	6.22 (3.02)	NA	NA
1	Latinx	132	18.26 (28.78)	5.32 (4.06)	5.64 (3.08)	5.94 (3.08)	NA	NA
1	All out-groups	137	18.22 (28.84)	5.66 (3.88)	5.80 (3.05)	5.90 (3.09)	5.54 (1.34)	3.18 (1.46)
2	Asian	176	16.08 (28.40)	6.40 (3.03)	5.80 (2.88)	5.81 (3.24)	NA	NA
2	Black	171	19.47 (31.61)	4.93 (4.92)	6.44 (2.91)	6.36 (3.12)	NA	NA
2	Latinx	171	16.67 (28.94)	5.34 (4.43)	6.01 (2.95)	6.18 (3.08)	NA	NA
2	All out-groups	186	17.39 (29.65)	5.57 (4.23)	6.08 (2.92)	6.11 (3.15)	5.46 (1.52)	3.21 (1.56)
3	Asian	393	20.98 (19.68)	6.96 (3.19)	5.69 (3.09)	5.72 (3.47)	5.94 (1.07)	3.69 (1.40)

Table S2: Descriptive results from Study 4

Study	Target	N.MRW	Condition	Donation (raw)	Donation (abs)	Rating (raw)	Rating (abs)
4	Asian	322	AttitudeImportanceGENERIC	10.58 (24.42)	17.48 (20.06)	5.70 (3.05)	5.70 (3.05)
4	Asian	353	AttitudeImportancePERSONAL	11.93 (27.14)	21.28 (20.62)	5.95 (2.82)	5.95 (2.82)
4	Asian	335	FavorabilityGENERIC	12.57 (24.66)	18.53 (20.55)	6.25 (3.59)	6.43 (3.25)
4	Asian	344	FavorabilityPERSONAL	11.41 (23.98)	17.95 (19.56)	5.86 (3.84)	6.09 (3.47)

For Studies 1 and 2, the WTP values in Table S1 are the results from the BDM task. All WTP values are in U.S. cents (¢). Response values reported are means, with standard deviation (SD) values reported

in parentheses. Means and SDs are reported for each target out-group as well as for the full out-group sample (“All out-groups”). Studies 3 and 4 involved only one target out-group (Asian), and the response values reported for Studies 3-4 (“WTP” in Table S1, “Donation (abs)” in Table S2) are the absolute-valued results from the charitable donation task. Additionally, the raw average donation amount (where subtractions are evaluated as negative values and additions are evaluated as positive values) is reported in Table S2 as “Donation (raw)”. Both raw and absolute-valued measurement ratings are also reported for Study 4, since “Favorability” conditions involve a bipolar measurement that could be analyzed either as raw valence scores (on a -10 to 10 scale) or absolute-valued “extremity” scores (on a 0 to 10 scale). All models computed for this study use the absolute-valued donation response and the absolute-valued ratings to maintain comparability between the conditions; descriptive results of raw values are reported here simply for completeness.

In Studies 1 and 2, not every participant rated every target out-group; responses were collected only for target out-groups where favorability ratings were greater than 1 or less than -1, as described in Methods. As a result, different target group averages correspond to different N_{MRW} values from the total N_{MRW} who were analyzed in each study. N_{MRW} values for the “All out-groups” rows and for Study 3 correspond to the total analyzed N_{MRW} for each study. Additionally, across Studies 1-3, IMS and EMS are assessed with respect to “minorities” generally, not a specific target group; this is why the IMS and EMS averages have no values for the individual target out-groups in Studies 1 and 2.

3 Exploratory MDS analysis

As an exploratory analysis, we performed multidimensional scaling (MDS; see Borg & Groenen, 2005) on the measured variables (i.e., attitude rating, importance, moral relevance, average IMS, and average EMS) to explore the relationships among the variables from Studies 1-3 separately from willingness-to-pay. MDS allows us to explore the consistency in the relationships between measured variables across studies. Using the “smacof” package in R (Leeuw & Mair, 2009; Mair et al., 2021), we fit ordinal MDS configurations in two dimensions for each study. Figure S1 shows the two-dimensional multidimensional scaling (MDS) configurations for all four studies projected into a common space and aligned to one another using Procrustes rotation, with the configurations from Studies 2 and 3 each aligned to the configuration from Study 1.

For Studies 1 and 2, we used the INDSCAL method (Carroll & Chang, 1970) to compute a single configuration across all out-groups. Study 3 only had one target group, so that MDS configuration was computed using a standard 2-way MDS.

Figure S1: Multidimensional scaling of measured variables

MDS shows consistency in the relationships between measured variables across studies. A three-way MDS configuration of the measured variables from Study 1 (S1) in two dimensions is plotted using circles. The corresponding three-way MDS configuration of the measured variables from Study 2 (S2) is plotted using squares, aligned to the S1 configuration via Procrustes rotation. Finally, a standard two-dimensional MDS configuration of the measured variables from Study 3 (S3) is plotted using triangles, aligned to the S1 configuration via Procrustes rotation. Across the three studies' configurations, attitude extremity is color coded purple, attitude importance is color coded pink, attitude moral conviction relevance is color coded brown, and average scores for internal and external motivation to respond without prejudice are color coded black.

4 Complete Regression Model Details & Results

To predict WTP values for Studies 1 and 2, we use Bayesian hierarchical proportional odds cumulative logistic regression models with a cumulative link function, including varying (“random”) intercepts associated with subject ID. Note that random intercepts are not included for target group because the variance in WTP associated with target group is approximately zero (as estimated in separate models including only the random intercept for target group). The regression equation below defines a hierarchical proportional odds model with p predictors X and an ordinal response Y with k category levels, for i grouping levels (i.e., participant IDs) observed on j occasions (i.e., BDM trials):

$$\text{logit}(P(Y_{ij} \leq k) = \beta_{k0} - \beta_1 X_{1ij} - \dots - \beta_p X_{pij} - b_{0i}$$

Taking the exponent of the posterior β parameters estimated in these models provides an odds-ratio interpretation of the response $P(Y \leq k)$, which is the probability that the value of Y is less than or equal to a specific category level k .

The Bayesian hierarchical proportional odds models for Studies 1 and 2 were computed using the **brms** package, using the default (uninformative) priors and a flexible threshold. Models were fit with uninformative priors and a flexible threshold, using 4 sample chains with 10^4 iterations each (5000 warmup for 2×10^4 total post-warmup samples).

For Study 3 we fit a Bayesian multiple linear regression model predicting donation behavior from the same features as Studies 1 and 2. There was only one target group in Study 3, so we did not have repeated measurements from each participant, meaning no random effects were necessary in this model. Additionally, because donation amounts were chosen from a continuous set of values (unlike the discrete WTP choice options in Studies 1 and 2), our primary model for the donation behavior response for Study 3 was a Gaussian regression model.

Due to the nature of the donation behavior response variable distribution – continuous, bounded between 0 cents and .5 cents, and with inflated observations at the boundaries – a more sophisticated model choice might be a Bayesian zero-one-inflated beta regression, fit to the response variable after transforming it to a proportion. The parameters from such a model are difficult to interpret, however, and that model does not produce a substantively different pattern of effects, so we have retained the Gaussian model for the main analysis. The zero-one-inflated beta model is included here as well for completeness.

Finally, for Study 4, the main analysis was a Bayesian linear regression model predicting absolute donation behavior from the interaction between condition and absolute-valued measurement. Absolute values were used in this analysis to align valence measurements and importance measurements to a common scale. The default preregistered analysis set the “generic favorability” condition as the reference variable via dummy coding, however a follow-up analysis was also performed to look at all the pairwise comparisons between conditions.

A further additional model was also computed to compare the higher-level groupings of conditions to better represent the 2x2 factorial design of the study. This model was computed using a refactored version of the Study 4 data that indicated relevance level and measurement type separately. It therefore contains no new information beyond the original preregistered model, but it allows us to analyze the 3-way interaction between rating value, relevance level, and rating type.

Effects plots for the attitude and importance effects from all four studies are reported in the main manuscript. Here we include the full set of effects plots for the measured variable predictors for each study (i.e., attitude ratings, attitude importance, attitude moral conviction relevance, and the IMS and EMS covariates). For all studies, the plots show conditional effects: predicted probabilities of WTP or donation behavior values for each predictor holding the other predictors fixed at their mean values.

The next portion of the supplement contains the following results from each study: 1) the descriptive visualizations of the data showing conditional distributions of WTP values by each predictor, and 2) regression results and effects plots for each Bayesian hierarchical ordinal regression model we fit on the data. Note that the black bars on the conditional distribution plots for attitude extremity, importance, and moral conviction relevance indicate mean WTP values, and the gray boxes the standard error of the mean.

4.1 Study 1 Results

Figure S2: Study 1 descriptive visualizations of conditional WTP distributions

Figure S2 shows the descriptive conditional distributions of WTP values by each predictor for Study 1. Conditional distributions are plotted as scatterplots with jittered points to better show the variation in point density.

Table S3: Study 1 Bayesian ordinal regression predicting WTP

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	8.790	1.1500	6.810	11.3000	1	2600	6090
Intercept[1]	8.150	4.5000	-0.504	17.4000	1	1720	3450
Intercept[2]	10.600	4.5300	1.970	19.9000	1	1720	3430
Intercept[3]	13.200	4.5800	4.520	22.7000	1	1750	3510
Intercept[4]	15.600	4.6300	6.870	25.3000	1	1790	3450
Intercept[5]	16.500	4.6500	7.720	26.2000	1	1790	3590
Intercept[6]	18.900	4.7100	10.100	28.7000	1	1830	3660
Intercept[7]	19.600	4.7300	10.700	29.5000	1	1840	3640
Intercept[8]	20.500	4.7600	11.600	30.4000	1	1840	3560
Intercept[9]	21.000	4.7800	12.100	30.9000	1	1850	3620
Favorability	-0.270	0.0927	-0.456	-0.0922	1	7820	11700
Importance	1.000	0.2150	0.588	1.4400	1	4850	8260
MoralConviction	-0.360	0.1580	-0.674	-0.0507	1	5390	8730
IMS.Avg	0.945	0.7040	-0.369	2.4200	1	1690	3610
EMS.Avg	0.452	0.5870	-0.701	1.6000	1	1530	3220
disc	1.000	0.0000	1.000	1.0000	NA	NA	NA

Table S3 shows the Bayesian hierarchical ordinal regression model results for Study 1. Figure S3 shows the posterior distributions for each slope parameter as well as the corresponding MCMC trace plots. (For brevity, posterior distributions and trace plots are not shown for the threshold or subject-level parameters.) The effects from this model are visualized in the predicted WTP probability plots for each fixed effect in Figure S4.

Figure S3: Study 1 ordinal regression posterior distributions and trace plots

Figure S4: Study 1 predicted WTP probability plots

Figure S4 shows the complete list of conditional predicted WTP probability plots for the fixed effects from the ordinal regression model for Study 1. Attitude importance shows a clear positive effect of increased WTP value probabilities as the predictor value increases. All predicted effects are conditioned on the other predictors being fixed at their average values.

4.2 Study 2 Results

Figure S5: Study 2 descriptive visualizations of conditional WTP distributions

Figure S5 shows the descriptive conditional distributions of WTP values by each predictor for Study 2. WTP distributed by attitude extremity, importance, and moral relevance are plotted in the top row using “pirate plots” (vertical smoothed histograms for each distribution with a black bar for the mean and a gray box for standard errors). Conditional distribution plots for WTP by average IMS and EMS are plotted as scatterplots with translucent points to better show the variation in point density.

Table S4: Study 2 Bayesian ordinal regression predicting WTP

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	12.100	1.6000	9.310	15.60000	1	3060	6940
Intercept[1]	9.550	4.7100	0.566	19.30000	1	2670	4360
Intercept[2]	11.700	4.7500	2.670	21.60000	1	2700	4380
Intercept[3]	13.800	4.8000	4.710	23.70000	1	2760	4520
Intercept[4]	16.100	4.8400	6.990	26.20000	1	2820	4570
Intercept[5]	17.900	4.8700	8.760	28.00000	1	2850	4610
Intercept[6]	20.200	4.9100	11.000	30.40000	1	2880	4750
Intercept[7]	21.300	4.9400	12.000	31.60000	1	2890	4700
Intercept[8]	22.600	4.9800	13.300	33.00000	1	2910	4560
Intercept[9]	23.200	5.0100	13.900	33.60000	1	2940	4650
Favorability	-0.151	0.0824	-0.318	0.00692	1	14600	12300
Importance	0.473	0.1510	0.185	0.78200	1	11900	11700
MoralConviction	0.120	0.1330	-0.135	0.38300	1	12400	12700
IMS.Avg	0.480	0.7180	-0.908	1.92000	1	2270	4300
EMS.Avg	0.844	0.6920	-0.495	2.25000	1	2260	4170
disc	1.000	0.0000	1.000	1.00000	NA	NA	NA

Table S4 shows the Bayesian hierarchical ordinal regression model results for Study 2. Figure S6 shows the posterior distributions for each slope parameter as well as the corresponding MCMC trace plots. (For brevity, posterior distributions and trace plots are not shown for the threshold or subject-level parameters.) The effects from this model are visualized in the predicted WTP probability plots for each fixed effect in Figure S7.

Figure S6: Study 2 ordinal regression posterior distributions and trace plots

Figure S7: Study 2 predicted WTP probability plots

Figure S7 shows the complete list of conditional predicted WTP probability plots for the fixed effects from the ordinal regression model for Study 2. Attitude importance shows the clearest positive effect of increased WTP value probabilities as the predictor value increases, although the effect is slightly smaller than in Study 1. All predicted effects are conditioned on fixing the other predictors at their average values.

4.3 Study 3 Results

4.3.1 Study 3 Gaussian Regression

Figure S8: Study 3 Gaussian regression posterior distributions and trace plots

Figure S8 shows the posterior distributions and MCMC trace plots for the Gaussian regression model for Study 3.

Figure S9: Study 3 Gaussian model effects plots predicting donation behavior

Figure S9 shows the descriptive conditional distributions of donation behavior values by each predictor for Study 3, with predicted effects from the Gaussian model of donation behavior overlaid as a regression line. Donation behavior refers either to the amount of money given up to donate to the charitable organization supporting Asian people in the U.S. (for respondents who reported favorable attitudes toward that group) or the amount of money given up to subtract from that total donation (for respondents who reported unfavorable attitudes toward that group).

Table S5: Study 3 Bayesian Gaussian regression predicting donation behavior

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-10.300	5.890	-21.70000	1.180	1	28700	16700
Favorability	0.669	0.342	-0.00622	1.330	1	27400	14600
Importance	1.330	0.369	0.60500	2.060	1	25500	15000
MoralConviction	-0.279	0.302	-0.87000	0.311	1	30000	15700
IMS.Avg	3.610	1.010	1.62000	5.590	1	29000	15900
EMS.Avg	-0.205	0.654	-1.48000	1.080	1	32800	14400
sigma	18.300	0.658	17.10000	19.700	1	33100	14800

Table S5 shows the Gaussian regression results for Study 3. Attitude importance, raw attitude rating, and internal motivation to respond without prejudice (average IMS) were positive predictors of donation behavior. The strongest predictors were attitude importance and average IMS, while raw attitude rating was a slightly weaker predictor.

4.3.2 Study 3 Zero-One-Inflated Beta Regression

Figure S10, Figure S11, and Table S6 below show the results from the zero-one-inflated beta regression model for Study 3. These results reflect a similar overall pattern of effects to the Gaussian regression model, so the Gaussian regression is used in the main analysis for ease of interpretation. The zero-one-inflated beta model is reported here for completeness. Note that the response scale for this model is squeezed into a (0, 1) range, and cannot be interpreted as cents donated anymore.

Figure S10: Study 3 zero-one-inflated beta regression posterior distributions and trace plots

Figure S11: Study 3 zero-one-inflated beta effects plots predicting donation behavior

Table S6: Study 3 Bayesian zero-one-inflated beta regression predicting donation behavior

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-1.0400	0.5950	-2.2300	0.1140	1	28200	15600
Favorability	0.0468	0.0306	-0.0121	0.1070	1	30800	16500
Importance	0.0412	0.0278	-0.0129	0.0960	1	27800	15700
MoralConviction	-0.0234	0.0243	-0.0703	0.0237	1	29200	15900
IMS.Avg	0.0876	0.0921	-0.0932	0.2690	1	32700	15700
EMS.Avg	-0.1170	0.0548	-0.2260	-0.0100	1	32700	16200
phi	3.4700	0.3220	2.8800	4.1300	1	34300	15700
zoi	0.5060	0.0252	0.4580	0.5560	1	37500	15300
coi	0.4680	0.0353	0.3990	0.5370	1	35300	15100

4.3.3 Study 3 Investigation of Potential Order Effects

Because the order of donation task, attitude importance measurement, and moral conviction measurement were randomized in Study 3, we were able to test for any potential order effects that might influence the attitude importance effect we observed. To address this possibility, we re-fit the Gaussian model for Study 3 including an interaction effect between attitude importance and a logical variable indicating whether the importance measurement or the donation task was presented first to a given participant. This variable (`Order_Importance_First`) evaluates to `TRUE` if the importance measurement came before the donation task, or `FALSE` if the donation task came before the importance measurement. (These indicators are agnostic to when the moral conviction measurement occurred.)

The results from this analysis are reported in Figure S12, Figure S13, and Table S7. Overall, results indicate that there were no discernible order effects for the attitude importance measurement. As a result, we retained the original, simpler model for Study 3, which does not include the presentation order predictor.

Figure S12: Study 3 regression testing order effects: posterior distributions and trace plots

Figure S13: Study 3 Gaussian model effects plots predicting donation behavior including potential order effect

Table S7: Study 3 Gaussian regression predicting donation behavior including potential order effect

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-12.000	6.210	-24.20000	0.247	1	25300	16500
Favorability	0.656	0.342	-0.00646	1.330	1	21800	14900
Importance	1.690	0.466	0.77900	2.610	1	17700	14600
Order_Importance_FirstTRUE	3.120	3.830	-4.37000	10.600	1	16100	15200
MoralConviction	-0.277	0.308	-0.88400	0.328	1	23000	13700
IMS.Avg	3.680	1.020	1.67000	5.660	1	24200	13900
EMS.Avg	-0.217	0.666	-1.51000	1.080	1	24300	15200
Importance:Order_Importance_FirstTRUE	-0.830	0.588	-1.98000	0.313	1	15400	14500
sigma	18.300	0.655	17.00000	19.600	1	23000	14300

4.4 Study 4 Results

Interaction plots showing the results of Study 4 can be found in the main text in Figure 4. Here, we show the complete Bayesian regression results for the condition-level model (i.e., the primary preregistered model) as well as the factor-level model. The factor-level model is constructed from the same data as the condition-level model, but with conditions grouped into their factorial categories (rating type: favorability or attitude importance; and relevance level: personal or generic).

4.4.1 Study 4 Condition-level Regression Model

Figure S14 shows the posterior distributions and MCMC trace plots for the condition-level regression model for Study 4. Table S8 lists the corresponding regression table for the condition-level model.

Figure S14: Study 4 condition-level regression posterior distributions and trace plots

Table S8: Study 4 condition-level regression predicting donation behavior

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	13.400	2.420	8.620	18.100	1	6500	9920
conditionAttitudeImportanceGENERIC	2.370	3.410	-4.370	9.060	1	7420	10400
conditionAttitudeImportancePERSONAL	-0.864	3.420	-7.600	5.890	1	7450	10700
conditionFavorabilityPERSONAL	-5.010	3.270	-11.400	1.480	1	7430	10900
rating_abs	0.793	0.336	0.138	1.450	1	6340	8980
conditionAttitudeImportanceGENERIC:rating_abs	-0.496	0.500	-1.470	0.492	1	7680	10600
conditionAttitudeImportancePERSONAL:rating_abs	0.675	0.499	-0.315	1.660	1	7500	11300
conditionFavorabilityPERSONAL:rating_abs	0.775	0.460	-0.137	1.680	1	7310	10400
sigma	19.900	0.382	19.200	20.700	1	19700	14600

4.4.2 Study 4 Factor-level Regression Model

Figure S15 shows the posterior distributions and MCMC trace plots for the factor-level regression model for Study 4. Table S8 lists the corresponding regression table for the factor-level model.

Figure S15: Study 4 factor-level regression posterior distributions and trace plots

Table S9: Study 4 factor-level regression predicting donation behavior

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	15.800	2.360	11.100	20.400	1	8280	12100
rating_abs	0.293	0.364	-0.420	1.000	1	8160	12100
self.relevancePERSONAL	-3.250	3.400	-9.970	3.370	1	7470	10500
rating.typeFavorability	-2.400	3.410	-8.970	4.210	1	7550	10900
rating_abs:self.relevancePERSONAL	1.170	0.520	0.147	2.180	1	7320	10400
rating_abs:rating.typeFavorability	0.502	0.500	-0.471	1.460	1	7390	10600
self.relevancePERSONAL:rating.typeFavorability	-1.760	4.740	-11.100	7.500	1	7220	9900
3-wayInteraction	-0.399	0.694	-1.750	0.971	1	6990	10100
sigma	19.900	0.386	19.200	20.700	1	17200	13000

5 Checking for Interaction Effects

To assess whether the attitude features measured in the present studies interacted in a way that might have suggested alternative patterns of results, we fit the models from each study again, this time including all two-way interactions between the item-specific variables (i.e., Favorability, Importance, and MoralConviction). In each case, the models including interaction terms did not substantively alter model fit over the simpler models reported in the main analysis, as determined by model comparison via LOOIC and ELPD (Vehtari, Gelman, & Gabry, 2017). These results led us to prefer and retain the simpler models for the main analysis of the present studies.

5.1 Study 1

- Base model LOOIC: 553.355 (SE 50.752)
- Base model ELPD: -276.678 (SE 25.376)
- Interaction model LOOIC: 563.196 (SE 51.527)
- Interaction model ELPD: -281.598 (SE 25.763)

5.2 Study 2

- Base model LOOIC: 563.55 (SE 51.034)
- Base model ELPD: -281.775 (SE 25.517)
- Interaction model LOOIC: 573.136 (SE 51.956)
- Interaction model ELPD: -286.568 (SE 25.978)

5.3 Study 3

- Base model LOOIC: 3404.7 (SE 20.171)
- Base model ELPD: -1702.35 (SE 10.085)
- Interaction model LOOIC: 3408.845 (SE 20.63)
- Interaction model ELPD: -1704.422 (SE 10.315)

6 Extremity analysis

In addition to the analysis of raw attitude ratings in our models above, we computed models that used the absolute value of the attitude ratings as a predictor instead of the original ratings themselves. Analyses using this “extremity” predictor show the same overall pattern of results: attitude importance remains the strongest, most consistent predictor of behavior. Effects plots, regression tables, and model diagnostics for those “extremity” models are included here for completeness.

6.1 Study 1 Extremity Model Results

Table S10: Study 1 extremity model

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept[1]	6.150	4.330	-2.250	14.9000	1	2070	4250
Intercept[2]	8.540	4.350	0.180	17.3000	1	2080	4220
Intercept[3]	11.000	4.380	2.630	20.0000	1	2110	4250
Intercept[4]	13.300	4.420	4.910	22.3000	1	2150	4290
Intercept[5]	14.100	4.430	5.730	23.1000	1	2160	4330
Intercept[6]	16.400	4.490	7.990	25.6000	1	2190	4350
Intercept[7]	17.100	4.510	8.620	26.3000	1	2220	4310
Intercept[8]	18.000	4.540	9.470	27.4000	1	2240	4310
Intercept[9]	18.500	4.570	9.940	27.8000	1	2250	4370
Extremity	-0.252	0.147	-0.543	0.0386	1	8630	12600
Importance	1.020	0.223	0.594	1.4700	1	5090	8740
MoralConviction	-0.257	0.150	-0.550	0.0387	1	6640	11000
IMS.Avg	0.493	0.661	-0.782	1.8400	1	2070	4160
EMS.Avg	0.413	0.562	-0.686	1.5300	1	1690	3550
disc	1.000	0.000	1.000	1.0000	NA	NA	NA

Figure S16: Study 1 extremity model posterior distributions and trace plots

Figure S17: Study 1 extremity model predicted WTP probability plots

6.2 Study 2 Extremity Model Results

Table S11: Study 2 extremity model

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept[1]	8.540	4.620	-0.290	17.700	1	3060	5070
Intercept[2]	10.600	4.660	1.810	20.000	1	3110	5090
Intercept[3]	12.700	4.700	3.810	22.200	1	3150	5320
Intercept[4]	15.000	4.740	6.060	24.600	1	3180	5380
Intercept[5]	16.800	4.770	7.800	26.400	1	3190	5420
Intercept[6]	19.000	4.800	9.970	28.700	1	3220	5330
Intercept[7]	20.100	4.830	11.100	29.800	1	3240	5560
Intercept[8]	21.300	4.860	12.300	31.300	1	3260	5740
Intercept[9]	22.000	4.890	12.900	31.900	1	3280	5580
Extremity	-0.123	0.134	-0.387	0.135	1	15200	15100
Importance	0.495	0.160	0.191	0.822	1	16300	14400
MoralConviction	0.125	0.132	-0.134	0.388	1	14700	14100
IMS.Avg	0.251	0.690	-1.090	1.640	1	2610	5200
EMS.Avg	0.856	0.654	-0.424	2.180	1	2790	5440
disc	1.000	0.000	1.000	1.000	NA	NA	NA

Figure S18: Study 2 extremity model posterior distributions and trace plots

Figure S19: Study 2 extremity model predicted WTP probability plots

6.3 Study 3 Extremity Model Results

Table S12: Study 3 extremity model

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-13.5000	6.080	-25.300	-1.610	1	27000	17600
Extremity	1.0700	0.419	0.257	1.900	1	30200	16000
Importance	1.2600	0.368	0.532	1.980	1	25800	16200
MoralConviction	-0.3460	0.310	-0.949	0.259	1	28400	15200
IMS.Avg	3.6700	0.985	1.730	5.580	1	29700	15400
EMS.Avg	-0.0749	0.662	-1.380	1.230	1	31100	14600
sigma	18.2000	0.660	17.000	19.600	1	29800	14400

Figure S20: Study 3 extremity model posterior distributions and trace plots

Figure S21: Study 3 extremity model effects plots predicting donation behavior

7 Supplementary References

- Borg, I., & Groenen, P. J. F. (2005). *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.). *Springer-Verlag*. <https://doi.org/10.1007/0-387-28981-X>
- Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3), 283–319. <https://doi.org/10.1007/BF02310791>
- Leeuw, J. de, & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, 31(1), 1–30. <https://doi.org/10.18637/jss.v031.i03>
- Mair, P., Groenen, P. J. F., & de Leeuw, J. (2021). More on Multidimensional Scaling and Unfolding in R: smacof Version 2. *Journal of Statistical Software*, 47.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>