De-stereotyping Public Performance Evaluation

Yixin Liu[a]* and Chengxin Xu[b]

[a]*Askew School of Public Administration, Florida State University, Tallahassee, U.S.A.;* [b]*Institute of Public Service, Seattle University, Seattle, U.S.A.*

*Corresponding author: Yixin Liu, Askew School of Public Administration, Florida State University, 028 Bellamy, Tallahassee, FL 32306. Email: yl17g@my.fsu.edu.

Author bio:

Yixin Liu is a PhD Candidate in Askew School of Public Administration at Florida State University. His research focuses on collaborative governance, performance measurement, and environmental management.

Chengxin Xu is an assistant professor at the Institute of Public Service at Seattle University. His research focuses on sectoral relations of the social service market, including sector difference, competition, collaboration, and organization hybridization. He receives his PhD from Rutgers University–Newark.

De-stereotyping Public Performance Evaluation

Experimental evidence suggests that citizens' judgments of service quality often rely on prior beliefs about providers' characteristics, such as racial stereotypes. Such a biased judgment process prevents the public from understanding performance information accurately and choosing high-quality service providers. To address this, we studied the relation between performance information and the evaluation mode and propose that presenting information jointly (joint evaluation) rather than separately (separate evaluation) may help people avoid stereotyping and consider actual performance. We compared people's perceived performance and preferences through the separate and joint evaluation modes (SE and JE) in three online experiments (N > 2,000), and obtained similar results in all studies: Subjects used racial stereotype to evaluate school performance in the SE condition, but such stereotyping decreased in the JE condition. Our findings provide an effective tool to de-stereotype performance evaluations, which also has implications for other public management research areas in reducing stereotyping behaviours.

Keywords: performance information; stereotyping; joint evaluation; online experiment

Introduction

People's prior stereotypical knowledge of the public organization may bias their understanding on its performance information. From two online experiments, this study shows that placing performance information jointly for people's evaluation can amplify the importance of performance data and update people's stereotypical judgement in evaluation. This finding enriches our knowledge of strategies that address widely reported cognitive bias in public performance evaluation.

Scholars' efforts in performance management research in public administration in recent decades can be summarized according to two waves, both of which focus on managing performance through performance information. The first wave of performance research focuses on institutional, organizational, and individual factors in the public sector that motivate the use of performance information (e.g., Julnes and Holzer 2001; Moynihan and Pandey 2010). The efficacy of using performance information relies on the assumption that those who use such information understand government performance in rational and consistent ways; however, this holds rarely in reality. Moynihan (2008) suggests that the interpretation of performance information is not an objective process but can be influenced by evaluators' roles in the public policy process. Following such findings, the second wave of performance management research, inspired primarily by psychology and behavioural science, challenges the rational decision-making assumption of performance evaluations and management, and points out various cognitive biases that public officials and the general public hold when processing performance information (Battaglio Jr et al. 2019). By involving the general public in the discussion of government performance evaluations, the second wave of performance management research also reminds public administration scholars of the “last mile problem” in public agencies’ efforts to build a performance-based accountability system.

To retain performance information’s important democratic value, a new wave of performance management research has begun to examine strategies to debias performance evaluations for both public officials, who use performance information to improve decision making, and the public, who rely on performance information to hold the government accountable and make service choices (e.g., Andersen and Guul 2019; James and Van Ryzin 2017; Nagtegaal et al. 2020). However, since cognitive biases stem from different psychological mechanisms, there is no panacea for all types of biases. Among all cognitive biases identified, stereotyping causes substantial problems in public affairs but remains as unsolved. Stereotyping is not only a biased process that people use to evaluate public service providers’ performance (Hvidman and Andersen 2016; Marvel 2016; Meier, Johnson, and An 2019), but also a problematic mechanism through which street-level bureaucrats discriminate against minorities (Andersen and Guul 2019). As the previous literature has demonstrated cases of stereotyping in different public administration issues (e.g. Jilke, Van Dooren, and Rys 2018), de-stereotyping strategies are necessary to create a public sector that fosters social justice and equity.

To reduce stereotypical category-based evaluation, we propose a nudging strategy for the general public: place performance information jointly rather than separately for evaluation. A key element for this strategy is the role of reference points in performance evaluation. Although literature on performance information has widely recognize how reference points matter to people’s evaluation (Olsen 2017), evidence about the efficacy of using simple reference points to improve performance evaluation is still rare. Compared with the separate evaluation mode (SE), the joint evaluation mode (JE) offers more data as reference points that help people to understand the relative performance of the organization. By involving comparable reference points, the JE mode is more compatible with people's behavioural mode of processing performance information, making sense of the performance by comparing and benchmarking. In turn, the new knowledge from the relative performance can update people's stereotypical beliefs. In consequence, people use the performance rather than stereotypical beliefs to make evaluation and relative decisions.

We conducted three online survey experiments on Amazon Mechanical Turk (MTurk) in the context of public high schools in the U.S. Both Study 1 and 2 manipulated the evaluation mode (i.e., JE or SE) and racial majority of students (White vs. Black) on which the public maintains stereotypical beliefs. Both studies showed consistent results: In the SE mode, the negative stereotype of Black students affected people's evaluations of high schools; In the JE mode, the effects of stereotyping diminished and the performance data became the major predictor to explain people's evaluation of high schools. In addition, we conducted Study 3 to check the robustness of our findings from researcher demand effects.

Our research shows the role of the JE mode in offsetting stereotypical judgment and facilitating the use of a data-driven decision-making process when evaluating performance. Practically, it also provides public administrators a cost-efficient tool, simply juxtaposing two or more pieces of information, to communicate public performance information.

Background: Debiasing Performance Evaluation

In recognition of people's cognitive biases in processing public performance information, efforts to identify debiasing strategies are emerging in performance management research in public administration. However, research designed specifically to address ways to correct bias is scant. The current debiasing literature considers the normative model of decision-making based on economic rationality as a benchmark of an unbiased decision-making process (Milkman, Chugh, and Bazerman 2009). Therefore, the goal of debiasing is to correct people's judgments and decisions that violate the rational decision making model, and foster decisions based on outcomes, values, or utilities, and the probability of occurrences. Importantly, as cognitive biases result from human beings' fundamental cognitive system (System 1 and System 2), debiasing strategies are not for eliminating cognitive biases but for encouraging rational reasoning (Battaglio Jr et al. 2019). As a result, people's judgment and decision-making are made based on unbiased process of information. Such a goal coincides with the normative value of performance information to improve decision making in the public sector. Biased understanding of performance data undermines the function of performance information which enables the public to make informed choices about public services and political participation (James and Van Ryzin 2017).

Debiasing strategies can be categorized into those that "modify the decision maker" through educational approaches and "modify the environment" to "...alter the environment to provide a better match for the thinking that people naturally do when unaided" (Soll, Milkman, and Payne 2015, 926). Recent experiments on debiasing in public administration focus on modifying the decision makers, who are largely politicians, public managers, and officials, and the effectiveness of debiasing strategies are mixed. For example, in a study of Danish politicians, Baekgaard et al. (2019) test whether increasing the amount of evidence could correct participants' biased evaluations of performance information attributable to their prior knowledge. In contrast to the theoretical expectation, the study shows that increasing evidence strengthened the role of politicians' prior knowledge. Another large-scale survey experiment on Danish politicians by Christensen and Moynihan (2020) examines justification requirements as a remedy of motivated reasoning and found different effects on debiasing perceptions of non-elite citizens and politicians. Other contributions to debiasing strategies that modify the decision maker include Cantarelli, Belle, and Belardinelli´s (2020) study, which shows that educational approaches can eliminate the band-wagoning and framing effects, and Nagtegaal et al.'s (2020) work, which provide experimental support for the "consider-the-opposite" technique in mitigating public managers and employees' anchoring bias. In comparison, strategies that modify the decision-making environment, such as changing the evaluation mode, gain insufficient attention.

Examinations in correcting biased performance evaluation suggest that there is no single debiasing strategy that can solve all judgment problems effectively, as biases derive from different sources. Therefore, the design of a debiasing strategy should focus specifically on the theoretical mechanism of a certain type of cognitive bias. Accordingly, it is also important to take individual variations into consideration, as people with different identities and backgrounds may react differently to the debiasing strategy. For example, Christensen and Moynihan (2020) find that politicians are more resistant to debiasing interventions than are the general public, probably because politicians maintain a stronger loyalty to their political ideologies and policy preferences.

Theory

Stereotyping in Performance Evaluation

Stereotyping has been recognized as one of people's major cognitive biases when evaluating government performance (Battaglio Jr et al. 2019). Previous evidence shows that people evaluate public organizations' performance based on stereotypical prior knowledge of a variety of organizational characteristics. This bias is similar to the fact that people's performance and merits are usually perceived stereotypically, either in negative or positive ways, based on their gender, race, and ethnicity (Bordalo et al. 2016). The demographic representation of one public organization may influence people's perceived performance. For example, the results of Riccucci, Van Ryzin, and Jackson's (2018) survey experiment indicate that although Whites in the U.S. rate the performance of the police favourably overall, they rate their performance, trustworthiness, and fairness less favourably when the police force include primarily Black officers compared to a force with a small percentage of Black officers, while complaints against the agency increase in the same way in both scenarios. Scholars have suggested also that parents' perceptions of school quality in the U.S. "...may be influenced by the racial and socioeconomic makeup of the school's student body" (Chingos, Henderson, and West 2012, 416). Inaccurate understanding of performance information leads to inefficacy of using performance measurement to enhance public trust toward the government. Indeed, Yang and Holzer (2006) suggests that performance measurement can influence citizen trust toward the government by indirectly improving citizens' perceptions of government performance.

Mechanism of Stereotyping

In social psychology literature, stereotyping is understood well as a cognitive approach that categorizes one object (individual or organization) automatically with a group of similar objects and forms judgments based on a general impression of the category (Fiske and Taylor 2017). Such biases are understood generally as outcomes of people's imbalanced reliance on the dual-system model of judgment, in which System 1 is quick and intuitive, while System 2 is more rational and quantitative and therefore monitors the quality of System 1 judgments (Kahneman and Frederick 2005). However, System 2 usually fails to correct judgments in System 1, as System 2 requires more cognitive effort and thus is difficult to activate. Recent research (e.g., Bordalo et al. 2016) shows that stereotyping is related to the use of heuristics in probability judgments, which is processed by System 1 (Kahneman and Tversky 1972). People's overreliance on System 1 rather than System 2 is a shared cause of other types of cognitive biases in public performance evaluation (Battaglio Jr et al. 2019). Although stereotypes may be based on empirical reality, they may still entail exaggerations (Judd and Park 1993). As a biased cognitive process, stereotypical judgment deviates from the goal of performance information, which is to provide a foundation of evidence-based decision making.

Decision making under uncertainties, such as the lack of relevant information, is an important prerequisite of people's overreliance on the heuristic judgment model in System 1 (Kahneman and Frederick 2005). In our case, people use stereotypes when they lack sufficient diagnostic information about the object and when stereotypes are comparatively more useful for their judgment (Crawford et al, 2011). Such mechanism implies that when people predominantly rely on stereotypes to evaluate public performance even if performance information is presented, it is possible that they consider performance information ambiguous or incompletely useful. Therefore, when performance information is more informative than stereotypes, System 2 will be triggered, and people will use such information to update their judgment based on stereotypical heuristics. To make performance information more diagnostic for the people and activate System 2, we propose the use of joint evaluation mode (JE), which enhances the evaluability of performance information by simply presenting two pieces of information side by side.

Evaluation Mode: Joint (JE) and Separate Evaluation (SE)

Evaluation mode is the cognitive mechanism about how a decision is made, and joint and separate evaluation are two fundamental evaluation modes for people's decision making. Observations in decision making research suggest that people make more reasoned decisions in joint than in separate evaluation modes (Bohnet et al. 2016; Hsee et al. 1999; Li and Hsee 2019). Theoretically, the JE mode makes the performance information more evaluable by providing new reference points and more data than the SE mode. Such evaluation mode is compatible with people's information processing behaviour, which is to make comparisons. Research on public performance evaluation emphasizes the value of comparative public performance. Indeed, public managers and average people make sense of performance information by comparing the data with peer organizations or with historical performance. For example, Meier, Favero, and Zhu (2015)'s Bayesian decision theory of managerial action is centred on the focal organization's performance gap compared to its previous performance and peer organizations' performance. Similarly, Ammons and Roenigk (2015) also suggests that statistical benchmarking is one important way to involve performance information into managerial decision making. In addition to studies on public managers, both Charbonneau and Van Ryzin (2015) and Olsen (2017) find that citizens' evaluation of government performance is substantially influenced by social and historical performance benchmarks. In contrast to the SE mode where people only evaluate the performance of one single organization, the JE mode satisfies people's propensity to make more reasoned, evidence-based judgments on the organizational performance by comparing and benchmarking. Specifically, when evaluating the performance of two organizations simultaneously, information in addition to the target organization's performance enables evaluators to use the organization's relative performance to update their prior (possibly biased) beliefs on organizational characteristics, for example, *Black schools* are generally underperforming. Importantly, when the relative performance information is counter-stereotypical, people tend to update their prior beliefs actively in the JE mode (Bohnet et al. 2016). Therefore, the JE mode will lead to more performance-driven judgment compared to the SE mode. However, it is important to notice that the JE mode only encourage decision makers using performance data to update their stereotypes in short term. Long term effort such as education and social movement is necessary for the change of stereotypes.

The JE mode has shown its effectiveness in mitigating stereotyping in the decision-making process. For example, an ongoing stereotype of females is that "women are bad at math" (Reuben, Sapienza, and Zingales 2014). Evidence from a lab experiment that simulated hiring practices in the labour market shows that participants were more likely to hire candidates for a job that require math skills based on real math performance in the JE mode, while more male candidates were hired in the SE mode (Bohnet et al. 2016). Hence, the hiring decision is merit-based in the JE mode, while the decision is made based upon gender stereotypes in the SE mode. Similar effect has also been reported in other research contexts. For example, studies in consumer behaviour show that brand name was more important than product features and price when participants evaluated products separately rather than jointly (Nowlis and Simonson 1997). Li and Hsee (2019) shows that when judging the sentencing term for a US fighter pilot who mistakenly had fired a missile and killed 18 civilians, participants in SE conditions imposed a significantly more severe punishment on the pilot if the victims are Belgians than if they are Somalians, whereas in the JE condition, participants imposed similar jail terms regardless of the victims' ethnicities. Beyond mitigating stereotyping, studies also showed that the JE mode encourages decision making based more on reasoned analysis than in the SE mode, in which decisions rely more on emotional desires. For example, people were willing to pay more to protect animal species when evaluating separately and to invest in human health when evaluating the two causes jointly (Kahneman et al., 1993). In different policy areas, Milkman et al (2012) shows that evaluating policy alternatives jointly (referred to as policy-bundling technique) can help to mitigate voters' loss aversion and gain support for policies that would create net social benefits but contain salient costs.

Hypotheses

Large amount of empirical evidence in different research context underscores the external validity of the JE mode in encouraging reasoned decision making. In this study, we apply these insights and test the effectiveness of the JE mode in the context of racial disparity of high school performance in the U.S. Examining strategies to improve public perception of performance information in the context of school choice offers important practical implications for public administration, because the general public has two important identities in this context. First, the public is the most important stakeholder of public schools, and accurate interpretation of school performance is necessary for people to hold the public education system accountable. Recent research studies public school performance as an important case of the way performance-based accountability reforms influence citizens' political behaviours. Evidence shows that voters react, by either exiting or voting, in response to public school performance (e.g., Holbein and Hassell 2019). Second, for those who are able to make school choice decisions, an unbiased understanding of high school performance should be the foundation of their decisions. However, on-going policy debates about the school system in the U.S., such as racial disparity (White vs. Black students) in academic performance, have established widely accepted stereotypes of school performance that may influence people's understanding of the actual performance of one particular school.

Black students have long suffered from a negative stereotype about their group academic performance (Steele and Aronson 1995) that mainly result from the unequal distribution of funding to Black public schools (Bifulco 2005). In turn, the negative perception of Black students and schools in general could colour the perceptions of the school's performance when they are evaluated in isolation (SE). Specifically, in the SE mode, people will rate schools with Black students as majority (i.e., Black schools) performance worse than schools with White students as majority (i.e., White schools), even though given the equivalent students' performance. Such negative stereotype of Black schools may be taste-based, which is a type of discrimination against out-group members, or statistical, which consider the race majority as a proxy of indicators related to the school's performance, such as the school district's socioeconomic status. However, both sources of stereotype influences people's decision making through the heuristic judgment process governed by System 1. Thus, we expect such biased evaluation will be addressed in the JE mode when the performance of Black schools and White schools are placed side by side, which triggers System 2. In the JE mode, school performance information is more diagnostic than stereotypes because the joint evaluation provides additional reference points for people to understand the performance better. In turn, people's school performance evaluation will be driven mainly by the school performance rather than student race majority.

In addition to perceived performance, we also expect people's school choice may follow the same pattern. Previous research in psychology already shows that changing the evaluation mode from SE to JE not only influences people's decision making but also actual behaviour such as hiring (Bohnet et al. 2016), purchasing (Nowlis and Simonson 1997), and charitable giving (Kahneman et al., 1993). In addition, people made school choice decisions based on a mix of considerations, including but not limited to school performance, students' socio-economic composition, home-school distance, and school environment. Among these factors, empirical evidence shows that, *ceteris paribus*, people have strong preference for schools' academic performance; meanwhile, school performance is more important than other factors that influence people' preference (Burgess et al. 2014). Since school performance is positively associated with people's school choice, we anticipate that the evaluation modes will affect people's school choice, following the pattern of perceived school performance. In sum, we test two following hypotheses:

H1 In the JE mode, students' racial majority of the high school are less likely to determine people's perceived school performance and choices than in the SE mode.

H2 In the JE mode, performance information is more likely to determine people's perceived performance of the organization and choices than in the SE mode.

Experiments

We conducted three randomized experiments on Amazon Mechanical Turk (MTurk) to examine the effect of JE mode on people's evaluation of high schools. Study 1 & 2 test both hypotheses, while Study 2 applied a more realistic design to enhance the generalizability of findings in Study 1. Both experiments are pre-registered at [anonymous for peer review] (pre-registration reports for both studies are available at: Appendix A1 and Appendix B1). In addition, we conducted Study 3 to explicitly examine the demand effect in our design of JE mode in Study 1 & 2.

We used the same measurements of performance evaluations in all studies: Perceived performance and school choice. We asked: “How well do you think this school is doing?” on a scale from 0 = “Very bad” to 100 = “Very good” to measure perceived performance. Next, we asked: “Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?” on a scale from 0 = “impossible” to 100 = “Very possible” to measure school choice. We included manipulation check and attention test questions after the survey instruments for the dependent variables to detect any treatment noncompliance or random choice behaviour. At the end of each experiment, we asked the same set of demographic questions as the covariates in the analysis, including gender, race, age, location (state), parenthood status, income, education, and ideology. We report the wording of these questions in Appendix D.

Study 1: De-stereotyping Black vs. White

Study 1 exams whether the negative stereotype of the Black high schools’ (the racial majority of the students is Black) performance has less effect on people’s perceived performance of the school and their choice decisions in the JE mode than the SE mode. We anticipate that, when given the same SAT performance[1], people on average will perceive the Black high school performs worse than its White counterpart in the SE mode, while they will judge two schools more equally in the JE mode. Considering the positive association between perceived performance and school choice, we expect people’s school choices to follow the same pattern as their perceived school performance.

[1] Scholastic Assessment Test (SAT) is a standardized exam required for most college admission in the U.S. It includes math and evidence-based reading and writing. Total score of SAT is 1600, 800 for math and 800 for evidence-based reading and writing.

Design

We assigned subjects randomly to three groups: Two SE groups in which subjects were asked to evaluate only one school profile (either a Black or a While school, randomly), and a JE group that paired both Black and White schools' information in the SE groups. After subjects read the school information in each condition, we asked about their perceptions of the schools' performance and school choice (see Appendix A2 for the experimental intervention). Figure 1 shows the experimental procedure. For such design, the unit of analysis of the study is the school profile rather than the individual, and dependent variables are performance ratings for school profiles and people's propensity to choose the school in the profile.

[Figure 1 is here]

Results

Study 1 recruited a total of 988 adult subjects who are located in the U.S. (53% female, 67% White, *M*age = 36). 824 (83%) subjects answered both the manipulation check and attention test correctly. We included the full sample in our main analysis to avoid potential posttreatment effect. We also conducted robustness analysis including only those who passed both check questions, and its result was similar to the main analysis (see Appendix A4, Figure A1). In the full sample, 316 subjects were randomly assigned to evaluate the performance of the Black school, 331 subjects were assigned to evaluate the performance of the White school, and 341 subjects evaluated both the Black and White schools simultaneously. The randomization resulted in 1,329 rated school profiles, in which 647 profiles were rated in SE mode and 682 profiles were rated in JE mode. Appendix A3 (Table A1) provides the more information of subjects' demographic statistics.

Findings from Study 1 support H1. We first conducted a t-test of ratings on Black and White schools in SE and JE modes to examine whether ratings of school profiles are influenced by the racial majority of schools in each mode (Figure 2). In the SE mode, the Black school's performance rating was 6% lower (95% confidence interval, C.I. = [3.37, 9.58], $p = 0.00$) than was the White school's performance rating. However, this difference was less than 1% in the JE mode. The analysis of school choice showed similar results. Subjects in the SE mode were 10% less (C.I. = [6.09, 14.06], $p = 0.00$) likely to send their children to the Black school than to the White school. This intention difference decreased to 4% in the JE mode (C.I. = [0.35, 7.64], $p = 0.03$).

[Figure 2 is here]

Following Li and Hsee (2019), we adopted a regression approach to estimate the treatment effect of evaluation mode change (Equation 1). We used evaluation mode (SE or JE), students' race majority (Black (B) or White (W)) and their interaction to predict the performance rating and the propensity to be chosen for one school profile. The coefficient of the interaction between the evaluation mode and students' racial majority indicates the treatment effect of JE in reducing racial stereotypical bias. This strategy allows us to test the null hypothesis that: $(JE_W - JE_B) - (SE_W - SE_B) = 0$, in which racial stereotypical alternatives (students' race majority) were denoted as W and B. In the following regression model, β_3 is the coefficient of the interaction effect to indicate the treatment effect of evaluation mode shift from SE to JE. C_i' is a matrix of covariates. In addition, we also clustered the standard errors by individuals to control potential non-independence between schools' profiles in the JE mode.

$$Y = \beta_1 \, Alternative + \beta_2 \, Mode + \beta_3 \, Alternative \times Mode + C_i'\gamma + \mu \text{ (Equation 1)}$$

The regression results indicate that the JE mode can effectively mitigate stereotyping on racial bias. Table 1 reports the results of the regression model, which confirmed the *t*-test results. Models 1 and 3 show the results not conditioned on the covariates. The coefficients of the interaction term suggests that a shift from the SE to JE mode reduced the racial difference in perceived performance by 7% (S.E. = 1.67, *p* = 0.00) and that in school choice by 6% (S.E. = 2.33, *p* = 0.01). We also obtained consistent results from the covariate-adjusted models (Models 2 & 4).

[Table 1 is here]

Discussion

Study 1's findings support our hypotheses. It shows that in the SE mode, when given the same average SAT score, people downgraded the performance of the Black high school stereotypically (H1), but the difference attributable to racial stereotyping was mitigated in the JE mode (H2). This indicates that the JE mode updates people's prior stereotypical knowledge about Black and White Schools, which in turn nudges people more likely to make judgment and behave based on performance information.

Study 2: De-stereotyping and Performance Information Use

Study 2 has two purposes. First, it aimed to replicate Study 1 and confirm H1. In Study 1, we provided same performance information (students' average SAT score) for each school in both SE and JE conditions. This design helps us to test the theoretical relations between racial stereotype and evaluation modes in a pure information environment. However, it is artificial to present two schools in JE mode with the same performance. To reduce the chance that subjects may detect researchers' purpose, Study 2 applied random numbers to SAT scores. By doing so, we can assess the robustness of the results in Study 1 in a more realistic information environment. Second, Study 2 also aimed to test H2. We expect that when given varied levels of performance data, people will be more likely to give higher performance ratings to the school with better performance in the JE mode than in the SE mode.

Design

Study 2 followed the similar experimental procedure of Study 1. We randomly assigned subjects to three groups to evaluate school performance: Two SE groups with either a Black or White school (randomly) and one JE group showing both Black and White schools. We then collected data on perceived performance, school choice, and demographics. The only difference between Study 1 and 2 was that in Study 2 we presented the total SAT score as random numbers between 1000 to 1190. This score range is the medium range of all American high school students (33% students locate in this range) (College Board 2020). Therefore, the school performance data we presented were at average level of American high schools, which was similar to the real-world situation. Figure 3 illustrates the experimental procedure. Detailed information about experimental intervention in Study 2 is reported in Appendix B2.

[Figure 3 is here]

Results

We recruited 1002 subjects (47% female, 69% White, Mage = 37), 854 (85%) of whom passed both the manipulation check and attention test. In the same process as Study 1, we include the full sample in our main analysis, and we conducted a robustness analysis including only those who passed both check questions (see Appendix B4, Figure B1, B2). Both analyses show similar results. After randomization, 330 subjects evaluated a Black school, 338 a White school, and 334 evaluated both schools simultaneously. In total, subjects evaluated 1336 school profiles, in which 668 profiles were in the SE mode and 668 profiles were in the JE mode. Detailed information on our sample and the randomization check (F test) is reported in Appendix B3 (Table B1).

To examine the de-stereotyping effect of JE, we employed the same analytical strategy as in Study 1. Figure 4 illustrates the t-test between racial alternatives under both the SE and JE modes. In general, we replicated our findings in Study 1. The subjects rated the White school's performance 7% higher (C.I = [4.50, 10.44], *p* = 0.00) than the Black school's in the SE mode. This difference decreased to 0% (C.I. = [-2.67, 2.42], *p* = 0.92) in the JE mode. We also found similar effects in school choice. The subjects had 12% stronger willingness (C.I. = [8.20, 15.71], *p* = 0.00) to send their children to White rather than Black schools in the SE mode, and this difference decreased to 4% (C.I. = [0.59, 7.51], p = 0.02) in the JE mode.

[Figure 4 is here]

Models in Table 2 without the covariate adjustment (Models 1 & 3) suggest that subjects decreased their perceptual difference attributable to their racial stereotype in the schools' performance evaluations by 8% (S.E. = 1.73, p = 0.00); they also decreased their school choice difference based on racial stereotypes by 8% (S.E. = 2.40, p = 0.00). Similar to Study 1, we also obtained consistent results from the covariate-adjusted models in this study (Models 2 & 4).

[Table 2 is here]

Next, we test H2 by looking at the two evaluation modes separately. For the subjects in each evaluation mode, we regressed students' average SAT score (*Performance*) and students' major race (*Alternative*) on performance ratings and school choice respectively, and then compare the effects of independent variables in SE and JE models (Equation 2). Following our hypotheses, we expect that in the SE mode, the coefficient of race alternative is more powerful in predicting the performance rating and school choice, whereas in the JE mode, the coefficient of SAT score randomly showed in the vignette is more dominant. Again, we clustered standard errors by individuals in the JE mode models.

$$Y = \beta_1\ Alternative + \beta_2\ Performance + C_i'\gamma + \ \mu \text{ (Equation 2)}$$

Model 1 in Table 3 shows that both racial stereotype and SAT score matters to people's perceived school performance in the SE mode (see Model 1 in Table 3), nonetheless the effect of performance information was much smaller. Holding other factors constant, the subjects rated the Black school's performance 7% lower (S.E. = 1.48, *p* = 0.00) than the White school. In contrast, one point increased in students' average SAT score contributed to 0.072% (S.E. = 0.01, *p* = 0.00) increased in school performance rating. In the JE mode, we did not find significant effect from racial alternatives, but the effect of SAT score was similar as it in the SE mode (see Model 3 in Table 3). Consistent results from covariate-adjusted models (Models 2 & 4 in Table 3) show the robustness of the above findings. Above results show that the subjects mainly used racial stereotype to evaluate school performance in the SE mode, but this effect diminished in the JE mode, where they only used performance data to evaluate school performance.

[Table 3 is here]

In Table 4, we obtain similar findings when considering school choice as the dependent variable. The effect of SAT score was consistent across the SE and JE modes, but the effect of racial stereotypical information largely reduced in the JE mode. We also obtained robust results from the covariate-adjusted models (Model 2 & 4 in Table 4). Different than perceived performance, the subjects still had 5% (S.E. = 1.40, *p* = 0.00) higher chance to send their kids to a White rather than Black school, holding other factors as equal.

[Table 4 is here]

Figure 5 visualizes the linear function between students' average SAT score and the two outcomes under the SE and JE modes. For both perceived performance and school choice in the SE mode (the left panels in Figure 5), the fit lines of Black and White schools were nearly paralleled, which suggest that the rating gaps between Black and White schools were significant and nearly constant regardless the school's SAT score. In contrast, rating gaps between Black and White schools in the JE mode were largely reduced (see the right panels in Figure 5). The subjects' evaluation outcomes were almost pure linear functions of the school performance, and the difference of racial majority is trivial.

[Figure 5 is here]

Discussion

Findings of Study 2 support both H1 and H2. In the SE mode, school evaluation was mainly driven by racial stereotype. Although performance data also positively affected school evaluation, its effects were much smaller than the stereotypical information. In the JE mode, the stereotypical information did not seem to matter in perceived performance, and its effect were largely reduced in school choice. Performance information were more likely to determine subjects' perceived performance of the organization. Thus, both Studies 1 and 2 documented the JE mode's effectiveness in reducing biases attributable to stereotyping in school performance evaluations.

Study 3: Robustness Test for Demand Effect in JE

Although Study 2 replicated findings from Study 1, some concerns about the demand effect remain. For example, we explicitly asked subjects to compare performances of two schools with salient race information in the JE mode, which may alert subjects that we preferred them to use performance rather than race in making their decisions. If so, the JE effects from Study 1 & 2 are from researchers' demands but not JE's de-stereotyping mechanism. To test the robustness of our findings from this bias, we supplement Study 3. This experiment is adapted from Mummolo and Peterson (2019), which explicitly manipulate the threat of demand effect to examine whether subjects will change their answers by researchers' preferences.

Design

We recruited 200 subjects from MTurk and randomly assigned them into two groups: the control group and the demand group. Detailed information on our sample and the randomization check (F test) is reported in Appendix C2 (Table C1). Holding all else equal, we explicitly told subjects in the demand group our experimental purpose and indicated that we expected people to prefer the White school rather than the Black school. Then, both groups saw the same information and questions as in the JE mode in Study 2. Theoretically, if the equivalent evaluations between Black and White schools in the JE mode were from the demand effect, we should observe the evaluation gap increase between Black and White schools when we induced them to consider preferring the White school (Mummolo and Peterson 2019). Detailed information about experimental intervention in this study is reported in Appendix C1.

Results

We used the similar regression approach as Study 1 & 2 to estimate the treatment effect of researcher demand. The coefficients of the interaction terms in Table 5 are our key variables of interest, which indicate whether the evaluation gap between Black and White schools are larger in the demand group than the control group. As result, we do not observe demand effect in estimating either performance evaluation or school choice. Although our sample size is smaller than Study 1 & 2, the interaction effect size

(Cohen's r) in the model (1) is 0.01 and in the model (3) is 0.09, which are both smaller than the 0.1 threshold for detecting small effect. Therefore, the demand effect should be negligible in our case, even if we increase the sample size.

[Table 5 is here]

Discussion

The results in this experiment show the robustness of our JE design from the demand effect. Subjects are not likely to align with researchers' preferences even if we explicitly tell them our expectation. Therefore, we have confidence that the JE effects we observed from Study 1 & 2 are from the evaluation mode mechanism that encouraging people to de-stereotype, but not the researcher demand effect.

General Discussion

Cognitive biases in evaluating and understanding public organizations' performance ultimately undermine people's wellbeing. When performance information is perceived insufficient or incomplete for decision making, evaluators will rely on their stereotypical prior knowledge and make inaccurate judgments. This study proposes that the JE mode can be considered as an effective nudge to encourage evidence-based performance evaluation by enhancing the evaluability of performance information and updating stereotypical prior knowledge. Our experimental results support our hypotheses. Our studies demonstrate that people would use racial and sector stereotypes to make performance judgments and school choice decisions in the SE mode, while such a stereotyping process would be dampened in the JE mode. This study has several implications for public management.

First, Study 1 & 2 reaffirm the existence of stereotype-based performance evaluation. In Study 1, Black schools' performance are rated significantly lower than

White schools, given the same SAT score. In Study 2 with a more realistic setting, we observe a consistent racial gap regarding school performance, even though higher SAT score is associated with better perceived performance. This racial gap is corresponding to the statistical reality that Black schools in the U.S. underperform, comparing with White counterparts, because of a segregated school financing system and other structural inequity in the society (Burnette II 2019). However, a stereotype-based performance evaluation may fail Black schools' efforts to improve their performance and gain public trust.

Second, we find that a switch from the SE mode to the JE mode can overcome stereotyping process in public performance evaluation. When stereotyping was observed, the stereotype-based performance gap was about 3 to 5 times smaller in the JE mode than in the SE mode. This finding shows evaluation mode's important role in performance evaluation, which previous literature has rarely explored. However, we argue that the JE mode only provides a starting point of ways to encourage evidence-based judgments. The JE mode only updates stereotypes in short term. People's stereotypical knowledge of social groups is socially constructed, and only by persistent effort can these stereotypes be changed in longer term. Evidence-based decision making cannot address the problem of stereotyping without concerning how evidence is considered.

Third, given the JE mode's effectiveness in de-stereotyping and highlighting performance data, this low-cost strategy can be translated in practice. The JE mode can be useful for public performance with comparable indicators. Beyond only using the JE mode to improve the effectiveness of performance information communication, public managers should present performance information in an evaluation mode that encourages people to process the information. Considering the democratic value of public performance, an appropriate mode that improves people's understanding of information can help the public to make better service choice decision in the public sector and hold the government accountable.

Limitation and Future Research

Although this study contributes to the scholarly understanding of stereotyping and de-stereotyping public performance information, it inevitably has limitations. First, our MTurk sample is more liberal, younger, and has higher educational degrees than does the general public, which may lead them to behave differently from others. Therefore, we welcome future work to replicate our findings in a more representative sample.

Second, the vignette context is hypothetical, which may elicit researcher demand effect, especially when using professional survey takers on MTurk. Although we applied random SAT scores in Study 2 to reduce this risk and explicitly tested the demand effect in Study 3, our design per se may still be too simple to the reality. Future work should test our hypotheses in a more complex scenario with multi-dimensional attribute information (such as conjoint design) to not only confound experimental purpose but also create a more realistic decision-making environment. Furthermore, we look forward to scholars validating the JE strategy in lab or field experiments to investigate further whether changing the evaluation mode leads to actual behavioural change, which is impossible to determine in survey experiments.

Third, given the complexity of performance information in real world, the effectiveness of JE in reducing stereotypical biases should be further investigated. JE can nudge people to focus more on quantifiable outcomes of public services, but its effectiveness may be weakened when performance indicators are multidimensional and ambiguous, which might lead to information overload that discourages System 2 (Jilke et al. 2016; Lee et al. 2020). Therefore, we encourage future work to directly test this assumption. Given the cognitive limitation and time constraint of citizens in making decision in choosing public services, public evaluation systems should reduce administrative burdens for citizens, strategically simplifying the comparison process of performance information (Moynihan 2008).

Fourth, our measurement of racial stereotype, the racial majority of the students, could be seen as proxies for multiple concepts in social construction. The Black and White schools' stereotypes can root from students' socio-economic status, regional distribution, or even teachers' racial composition. Therefore, our evaluation mode approach cannot explain the reasons behind racial serotyping, but it can effectively overcome stereotyping process and encourage people to use performance information in public evaluation. We suggest future studies to further disentangle the reasons of stereotype formation, which can contribute us to reduce stereotypical biases in long-run.

In addition, our design did not allow us to conduct a multi-dimensional subgroup analysis with sufficient statistical power. Thus, we encourage future researchers to parse the SE-JE effect among diverse social groups to detect any heterogeneity in the debiasing process. In particular, given that public servants' prior knowledge differs from that of the general public, the JE mode's effectiveness in de-stereotyping public servants' decision making remains an important question in generalization.

Future research on the evaluation mode could be developed through a variety of journeys. First, it is worthwhile to investigate whether the JE is a strategy to reduce stereotyping behaviours in areas other than performance evaluation. Public servants' unequal treatment of the public has been acknowledged widely as a major problem in the public sector. Evidence has shown a wide range of stereotyping behaviour on the part of public officials, in which officials' responsiveness varies depending on constituents' race, gender, social class, or religious identity (Grohs, Adam,and Knill 2016; Harrits 2019; Pedersen, Stritch, and Thuesen 2018; Pfaff et al. 2020). In addition, stereotyping leads to discrimination against female and minority public servants in working places (Guul, Villadsen, and Wulff 2019). For both problems, the JE mode may be an appropriate strategy to reduce unequal treatments against disadvantaged groups depending on the nature of specific issues. Overall, the JE's effectiveness and reliability is yet to be discussed in even more than the public management areas we listed above. Therefore, we look forward to applying this theory to other administrative behaviours.

The JE mode also has potentials to ameliorate biases which share the cognitive mechanism with stereotyping that influence public officials' decision making. For example, we wonder whether changing the evaluation mode addresses the problem of political motivated reasoning in public officials' interpretation of performance information. It is possible that a switch to the JE mode may not suppress or even encourage political motivated reasoning for those with stronger political knowledge. Such suspicion stems from previous evidence from Christensen and Moynihan (2020) and Baekgaard et al. (2019). Both studies show that debiasing strategies that works for the public might not correct politicians' biased performance evaluation. Unlike the non-elite subjects, public officials have stronger prior political knowledge and beliefs, which often renders political biases more resistant to behaviour interventions that modify the decision maker (Baekgaard et al. 2019; Christensen and Moynihan 2020). It is worthwhile to investigate whether approaches that modify the environment such as the JE mode can be effective alternatives that encourage better decision making.

References

Ammons, David N and Dale J Roenigk. 2015. Benchmarking and interorganizational learning in local government. *Journal of Public Administration Research and Theory* 25 (1):309–335.

Andersen, Simon Calmar and Thorbjørn Sejr Guul. 2019. Reducing minority discrimination at the front line—Combined survey and field experimental evidence. *Journal of Public Administration Research and Theory* 29 (3):429–444.

Andersen, Simon Calmar and Morten Hjortskov. 2016. Cognitive biases in performance evaluations. *Journal of Public Administration Research and Theory* 26 (4):647–662.

Baekgaard, Martin and Søren Serritzlew. 2016. Interpreting performance information: Motivated reasoning or unbiased comprehension. *Public Administration Review* 76 (1):73–82.

Baekgaard, Martin, Julian Christensen, Casper Mondrup Dahlmann, Asbjørn Mathiasen, and Niels Bjørn Grund Petersen. 2019. The role of evidence in politics: Motivated reasoning and persuasion among politicians. *British Journal of Political Science* 49 (3):1117–1140.

Battaglio Jr, R Paul, Paolo Belardinelli, Nicola Bell´e, and Paola Cantarelli. 2019. Behavioral public administration ad fontes: A synthesis of research on bounded rationality, cognitive biases, and nudging in public organizations. *Public Administration Review* 79 (3):304–320.

Bifulco, Robert. 2005. District-level black-white funding disparities in the united states, 1987- 2002. *Journal of Education Finance* 31 (2):172–194.

Bohnet, Iris, Alexandra Van Geen, and Max Bazerman. 2016. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science* 62 (5):1225–1234.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. Stereotypes. *The Quarterly Journal of Economics* 131 (4):1753–1794.

Burgess, S., Greaves, E., Vignoles, A., & Wilson, D. 2015. What parents want: School preferences and school choice. The Economic Journal, 125(587), 1262-1289.

Burnette, Daarel II. 2019. U.S. Spends $23B More on White Districts Than Nonwhite Districts, Report Says. *EducationWeek*. Retrieved on May 14 2021, from https://www.edweek.org/education/u-s-spends-23b-more-on-white-districts-than-nonwhite-districts-report-says/2019/02

Cantarelli, Paola, Nicola Bell´e, and Paolo Belardinelli. 2020. Behavioral public hr: Experimental evidence on cognitive biases and debiasing interventions. *Review of Public Personnel Administration* 40 (1):56–81.

Charbonneau, ´ Etienne and Gregg G Van Ryzin. 2015. Benchmarks and citizen judgments of local government performance: Findings from a survey experiment. *Public Management Review* 17 (2):288–304.

Chingos, Matthew Mark, Michael Henderson, and Martin Raymond West. 2012. Citizen perceptions of government service quality: Evidence from public schools. *Quarterly Journal of Political Science*.

Christensen, Julian and Donald P Moynihan. 2020. Motivated reasoning and policy information: Politicians are more resistant to debiasing interventions than the general public. *Behavioral Public Policy*.

CollegeBoard. 2020. SAT Suite of Assessment Annual Report: Total Group. Retrieved on May 17, 2021, from https://reports.collegeboard.org/pdf/2020-total-group-sat-suite-assessments-annual-report.pdf

Crawford, J. T., Jussim, L., Madon, S., Cain, T. R., & Stevens, S. T. (2011). The use of stereotypes and individuating information in political person perception. *Personality and Social Psychology Bulletin*, 37, 529–542.

Fiske, Susan T. and Shelley E. Taylor 2017. *Social Cognition: From Brains to Culture*. 3rd edition SAGE Publications Inc.

Grohs, Stephan, Christian Adam, and Christoph Knill. 2016. Are some citizens more equal than others? Evidence from a field experiment. *Public Administration Review* 76 (1):155–164.

Guul, Thorbjørn Sejr, Anders R Villadsen, and Jesper N Wulff. 2019. Does good performance reduce bad behavior? Antecedents of ethnic employment discrimination in public organizations. *Public Administration Review* 79 (5):666–674.

Harrits, Gitte Sommer. 2019. Stereotypes in context: How and when do street-level bureaucrats use class stereotypes? *Public Administration Review* 79 (1):93–103.

Holbein, John B and Hans JG Hassell. 2019. When your group fails: The effect of race-based performance signals on citizen voice and exit. *Journal of Public Administration Research and Theory* 29 (2):268–286.

Hsee, Chrisopher K, George F Loewenstein, Sally Blount, and Max H Bazerman. 1999. Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin* 125 (5):576.

Hvidman, Ulrik and Simon Calmar Andersen. 2016. Perceptions of public and private performance: Evidence from a survey experiment. *Public Administration Review* 76 (1):111–120.

James, Oliver and Gregg G Van Ryzin. 2017. Motivated reasoning about public performance: An experimental study of how citizens judge the affordable care act. *Journal of Public Administration Research and Theory* 27 (1):197–209.

Jilke, Sebastian, Wouter Van Dooren, and Sabine Rys. 2018. Discrimination and administrative burden in public service markets: Does a public–private difference exist? *Journal of Public Administration Research and Theory* 28 (3):423–439.

Julnes, Patria de Lancer and Marc Holzer. 2001. Promoting the utilization of performance measures in public organizations: An empirical study of factors affecting adoption and implementation. *Public Administration Review* 61 (6):693–708.

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, 100(1), 109.

Kahneman, D., Ritov, I., Jacowitz, K.E. and Grant, P., 1993. Stated willingness to pay for public goods: A psychological perspective. *Psychological Science*, 4(5), pp.310-315.

Li, Xilin and Christopher K Hsee. 2019. Beyond preference reversal: Distinguishing justifiability from evaluability in joint versus single evaluations. *Organizational Behavior and Human Decision Processes* 153:63–74.

List, John A. 2002. Preference reversals of a different kind: The" more is less" phenomenon. *American Economic Review* 92 (5):1636–1643.

Marvel, John D. 2016. Unconscious bias in citizens' evaluations of public sector performance. *Journal of Public Administration Research and Theory* 26 (1):143–158.

Meier, Kenneth J, Nathan Favero, and Ling Zhu. 2015. Performance gaps and managerial decisions: A bayesian decision theory of managerial action. *Journal of Public Administration Research and Theory* 25 (4):1221–1246.

Meier, Kenneth J, Austin P Johnson, and Seung-Ho An. 2019. Perceptual bias and public programs: The case of the united states and hospital care. *Public Administration Review* 79 (6):820–828.

Milkman, Katherine L, Dolly Chugh, and Max H Bazerman. 2009. How can decision making be improved? *Perspectives on Psychological Science* 4 (4):379–383.

Milkman, Katherine L, Mary Carol Mazza, Lisa L Shu, Chia-Jung Tsay, and Max H Bazerman. 2012. Policy bundling to overcome loss aversion: A method for improving legislative outcomes. *Organizational Behavior and Human Decision Processes* 117 (1):158–167.

Moynihan, Donald P 2008. *The dynamics of performance management: Constructing information and reform*. Georgetown University Press.

Moynihan, Donald P and Sanjay K Pandey. 2010. The big question for performance management: Why do managers use performance information? *Journal of Public Administration Research and Theory* 20 (4):849–866.

Mummolo, J. and Peterson, E., 2019. Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113(2), pp.517-529.

Nagtegaal, Rosanna, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. 2020. Designing to debias: Measuring and reducing public managers' anchoring bias. *Public Administration Review* 80 (4):565–576.

Nowlis, S.M. and Simonson, I., 1997. Attribute–task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research*, 34(2), pp.205-218.

Olsen, Asmus. 2017. Compared to what? how social and historical reference points affect citizens' performance evaluations. *Journal of Public Administration Research and Theory* 27 (4):562–580.

Pedersen, Mogens Jin, Justin M Stritch, and Frederik Thuesen. 2018. Punishment on the frontlines of public service delivery: Client ethnicity and caseworker sanctioning decisions in a Scandinavian welfare state. *Journal of Public Administration Research and Theory* 28 (3):339–354.

Pfaff, Steven, Charles Crabtree, Holger L Kern, and John B Holbein. 2020. Do street-level bureaucrats discriminate based on religion? A large-scale correspondence experiment among American public school principals. *Public Administration Review* 81(2): 244-259.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences* 111 (12):4403–4408.

Riccucci, Norma M, Gregg G Van Ryzin, and Karima Jackson. 2018. Representative bureaucracy, race, and policing: A survey experiment. *Journal of Public Administration Research and Theory* 28 (4):506–518.

Soll, Jack B, Katherine L Milkman, and John W Payne. 2015. A user's guide to debiasing. *The Wiley Blackwell handbook of judgment and decision making* 2:924–951.

Steele, Claude M and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology* 69 (5):797.

Yang, K. and Holzer, M., 2006. The performance–trust link: Implications for performance measurement. *Public Administration Review*, 66(1), pp.114-126.

Figures and Tables

Study 1

Figure 1. Study 1: Experimental Procedure

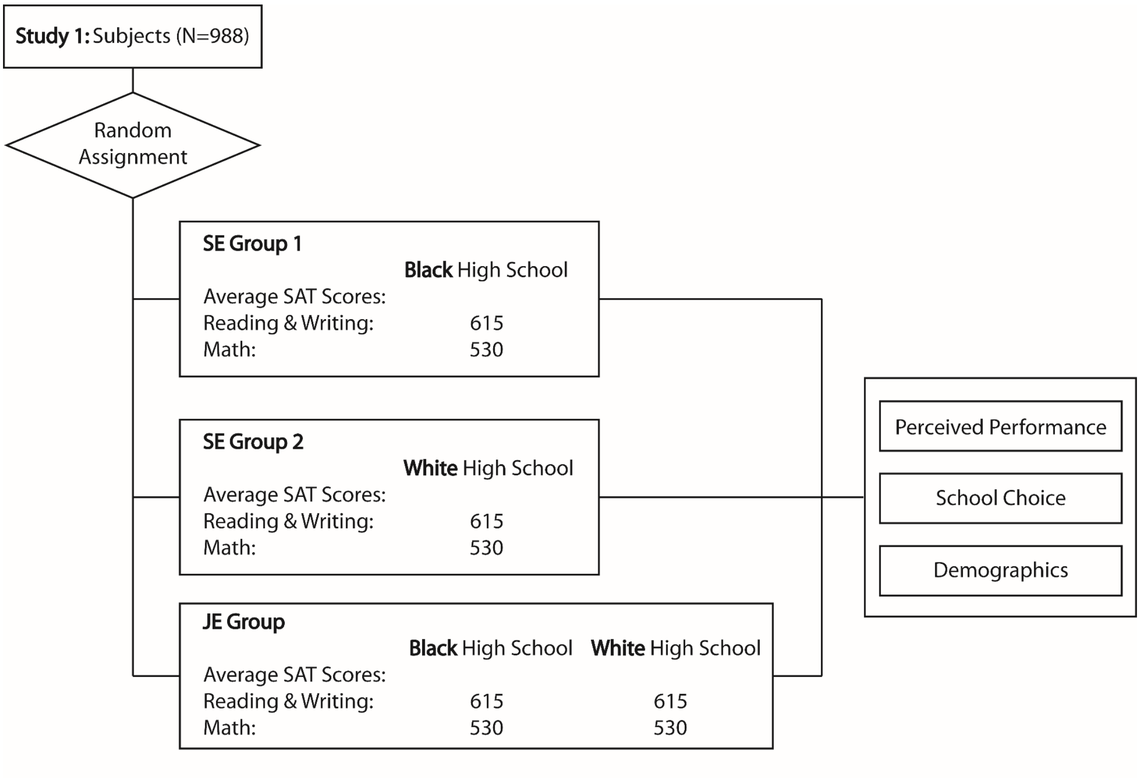


Figure 2. Study 1: Racial Stereotype
Note: Bars are 95% confidence intervals.

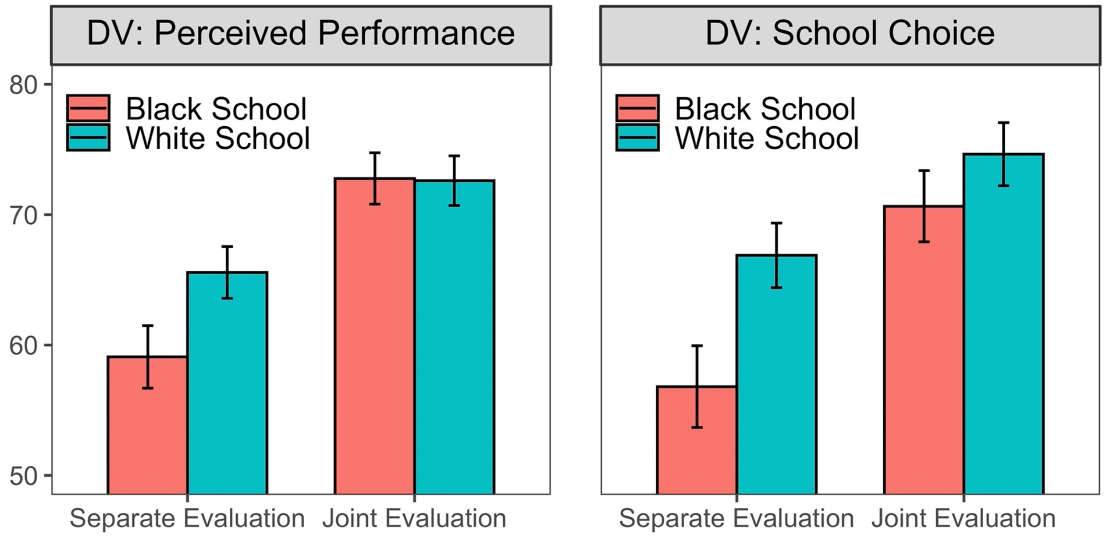


Table 1. Study 1: Racial Stereotype

	Perceived Performance		School Choice	
	(1)	(2)	(3)	(4)
JE × Black	6.645***	7.375***	6.080**	6.354**
	(1.673)	(1.737)	(2.329)	(2.396)
Mode: JE	7.040***	6.584***	7.758***	7.250***
	(1.398)	(1.417)	(1.760)	(1.844)
Students' Major Race: Black	-6.478***	-7.208***	-10.074***	-10.357***
	(1.581)	(1.639)	(2.029)	(2.086)
Constant	65.570***	64.867***	66.884***	73.430***
	(1.009)	(8.060)	(1.259)	(8.928)
Covariates	No	Yes	No	Yes
Observation	1,328	1,320	1,327	1,319
Adjusted R^2	0.078	0.102	0.063	0.085

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Study 2

Figure 3. Study 2: Experimental Procedure

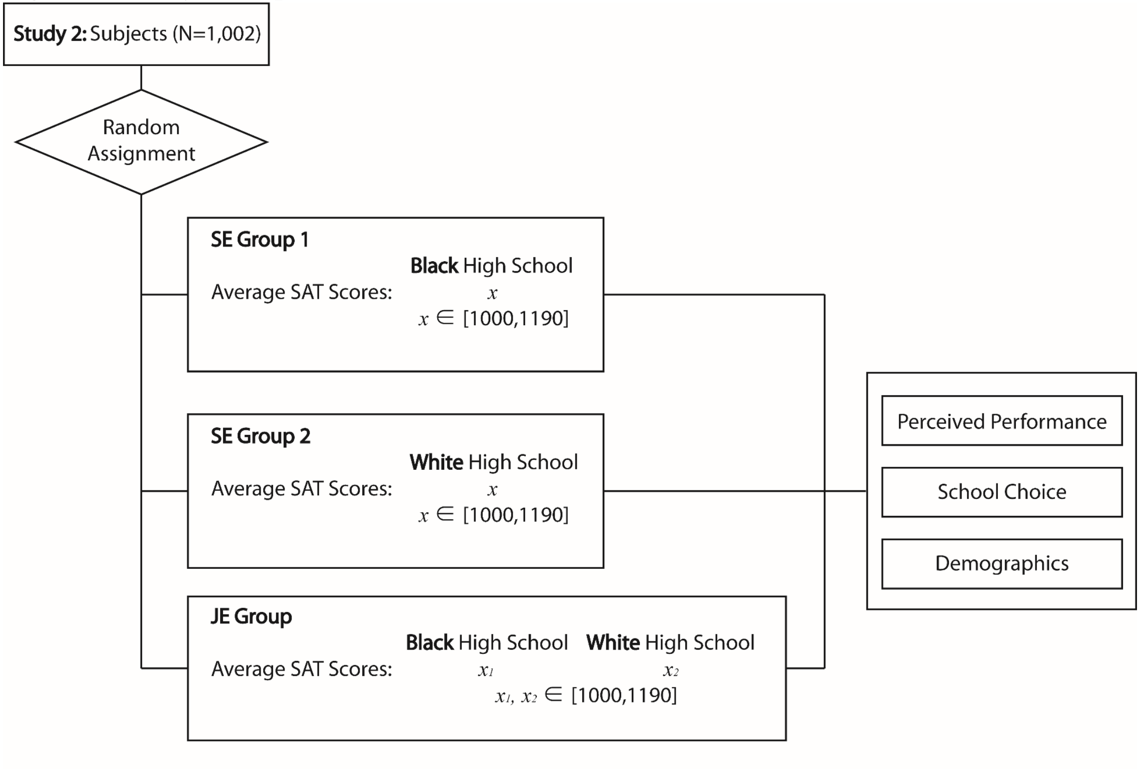


Figure 4. Study 2: Racial Stereotype
Note: Bars are 95% confidence intervals.

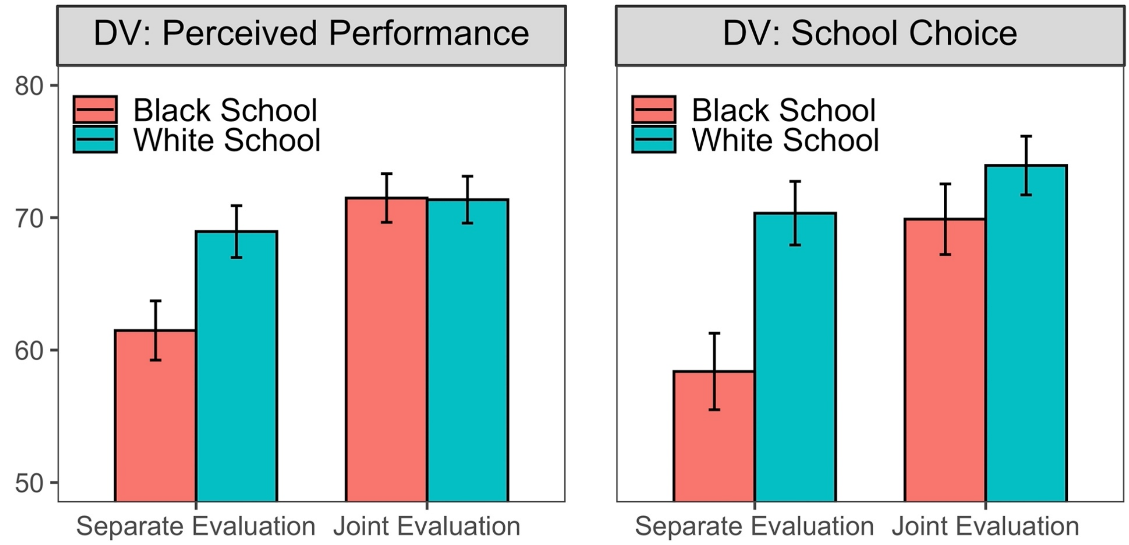


Table 2. Study 2: Racial Stereotype

	Perceived Performance		School Choice	
	(1)	(2)	(3)	(4)
JE × Black	7.598^{***}	6.160^{***}	7.901^{**}	6.183^{**}
	(1.730)	(1.709)	(2.397)	(2.382)
Mode: JE	2.405	2.744^{*}	3.598^{*}	4.126^{*}
	(1.343)	(1.356)	(1.663)	(1.711)
Students' Major Race: Black	-7.472^{***}	-6.619^{***}	-11.955^{***}	-10.830^{***}
	(1.511)	(1.515)	(1.911)	(1.923)
Constant	68.955^{***}	-0.599	70.342^{***}	8.294
	(0.996)	(11.171)	(1.222)	(14.426)
Covariates	No	Yes	No	Yes
Observation	1,336	1,328	1,336	1,328
Adjusted R^2	0.046	0.112	0.055	0.115

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 3. Study 2: Perceived Performance in the SE and JE modes

	Separate Evaluation		Joint Evaluation	
	(1)	(2)	(3)	(4)
Students' Major Race: Black	-7.270***	-6.283***	-0.392	-0.352
	(1.484)	(1.557)	(0.768)	(0.800)
Students' Average SAT	0.072***	0.068***	0.056***	0.052***
	(0.014)	(0.014)	(0.012)	(0.012)
Constant	-10.592	-16.418	9.865	24.817
	(15.001)	(17.029)	(12.849)	(14.902)
Covariates	No	Yes	No	Yes
Observation	668	660	668	668
Adjusted R^2	0.072	0.087	0.031	0.125

*p < .05; **p < .01; ***p < .001

Table 4. Study 2: School Choice in the SE and JE modes

	Separate Evaluation		Joint Evaluation	
	(1)	(2)	(3)	(4)
Students' Major Race: Black	-11.785***	-10.131***	-4.674***	-4.655**
	(1.900)	(1.992)	(1.395)	(1.451)
Students' Average SAT	0.061***	0.058**	0.068***	0.065***
	(0.017)	(0.018)	(0.017)	(0.016)
Constant	3.593	3.716	0.219	20.985
	(19.205)	(21.782)	(18.525)	(20.274)
Covariates	No	Yes	No	Yes
Observation	668	660	668	668
Adjusted R^2	0.069	0.096	0.031	0.089

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Figure 5. Study 2: Students' Major Race, SAT, and their Effects on Outcomes
Note: Shaded regions indicate the 95% confidence intervals.

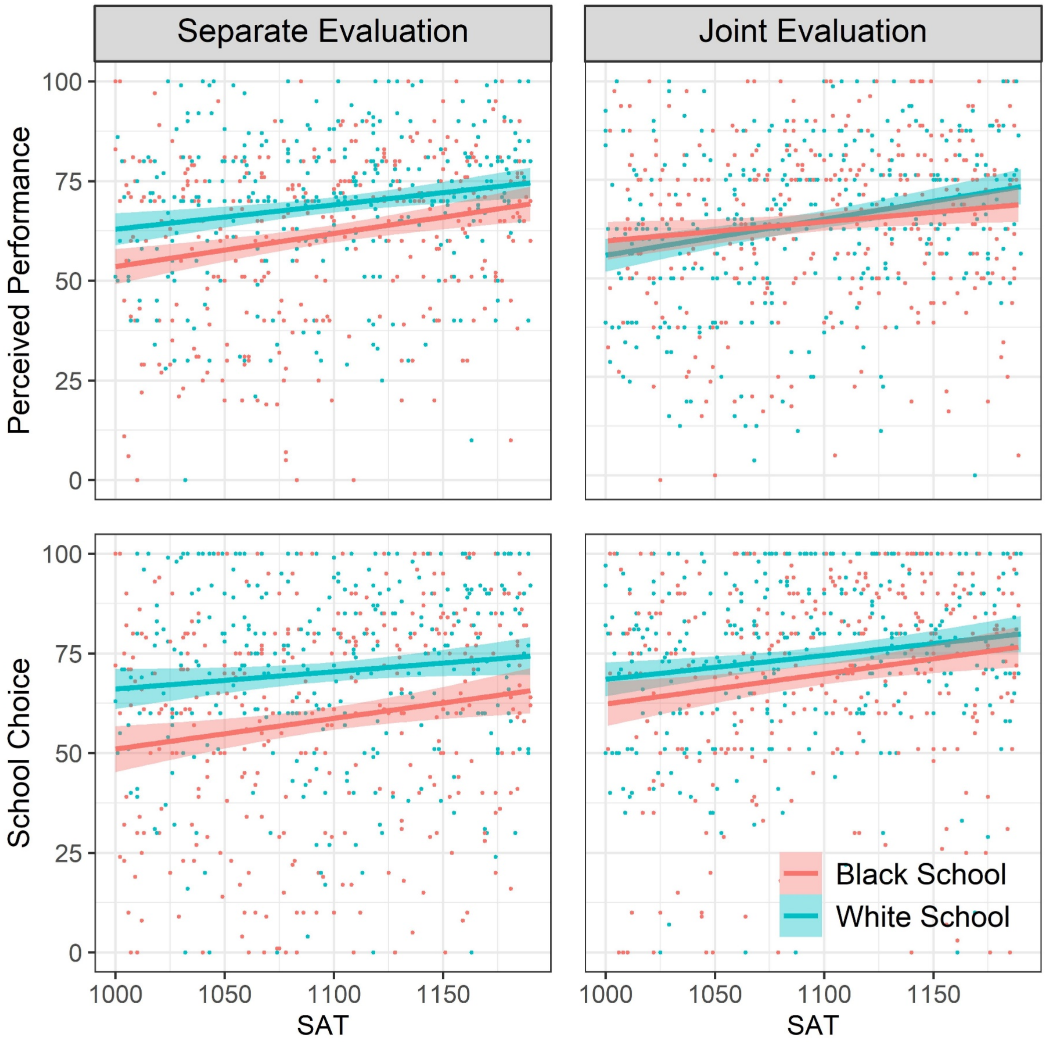


Table 5. Study 3: Racial Stereotype and Demand Effect

	Perceived Performance		School Choice	
	(1)	(2)	(3)	(4)
Demand Group × Black	0.486	1.612	-7.100	-6.295
	(2.126)	(2.095)	(3.639)	(3.958)
Demand Group	-4.910*	-4.160	-3.001	-1.191
	(2.313)	(2.605)	(3.011)	(3.039)
Students' Major Race: Black	-0.529	-0.584	-1.365	-1.472
	(1.363)	(1.220)	(2.528)	(2.632)
Constant	78.094***	-11.272	79.659***	10.048
	(1.505)	(20.271)	(2.149)	(26.662)
Covariates	No	Yes	No	Yes
Observation	398	392	398	392
Adjusted R^2	0.010	0.177	0.029	0.135

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Supplemental Information

Appendix A: Study 1

Appendix A1: Pre-registration report

1. Have any data been collected for this study already?

No, no data have been collected for this study yet.

2. What's the main question being asked or hypothesis being tested in this study?

In this survey experiment, we ask:

Can joint evaluation mode reduce stereotyping in performance perception and encourage people to use rational thinking rather than heuristics?

3. Describe the key dependent variable(s) specifying how they will be measured.

In this study, we have two main dependent variables.

i. Performance perception: how well do you think this school is doing? (Moving a 0-100 scale bar: 0 = very bad; 100 = very good)

ii. Behavioral intention: Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school? (Moving a 0-100 scale bar: 0 = impossible; 100 = very possible)

4. How many and which conditions will participants be assigned to?

We will randomize participants in three groups: two separate evaluation groups (SE) and one joint
evaluation group (JE).

In two SE groups, subjects will either see a a high school that its race majority of students is Black or a high school that its race majority of students is White. We also show that both schools have the same SAT performances. The only difference between SE groups is the students' race information (black or white). In the JE group, subjects will see both schools that SE groups see.

In the both SE groups, subjects will be asked to answer both performance perception and behavioral intention questions after they see the school information. In the JE condition, subjects will be asked to answer both performance perception and behavioral intention questions for both schools they see.

5. Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Analyses will be based on linear regression models using experimental manipulations as the explanatory variables.

6. Any secondary analyses?

We will conduct subgroup analyses by participants' characteristics.

After manipulation, we will ask all subjects a manipulation check question and an attention test question. We will compare results with or without manipulation check and attention test failure samples to see the robustness of our findings.

7. *How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.*

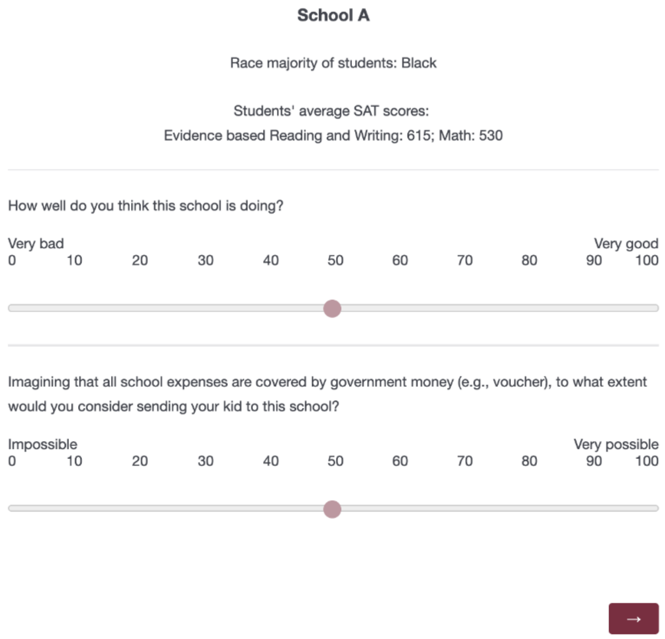
We will stop data collection once 1000 subjects have submitted a responses on MTurk. Deviations from this goal are entirely due to MTurk software and outside of our control.

8. *Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)*

Subjects' demographic information will be collected after they have answered the questions regarding key dependent variables. The information is collected for detecting the heterogeneity of the treatment effect and for the randomization balance check. Since we only recruit adult subjects in the U.S., VPN and proxy identifier will be applied at the beginning to filter out disqualified subjects.

Appendix A2: Experimental intervention

After a brief introduction to the American high school scenario, we randomly assigned subjects into one of the following three conditions.
Group 1: Separate Evaluation Condition: Black School



School A

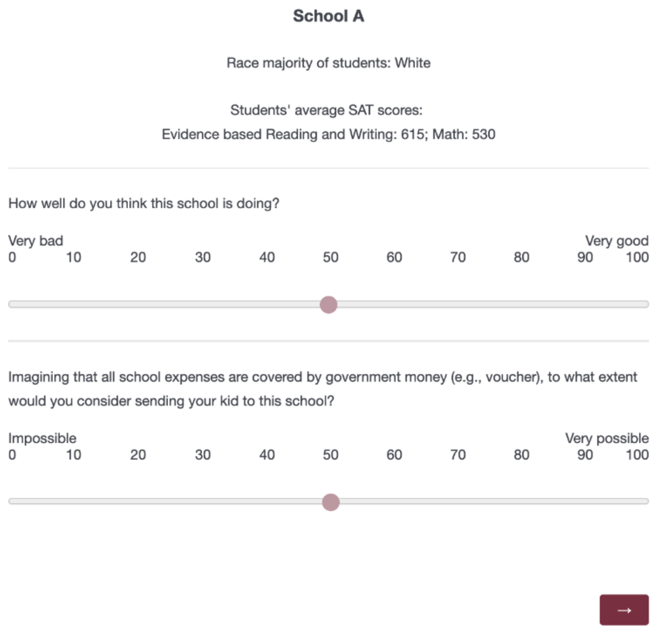
Race majority of students: Black

Students' average SAT scores:
Evidence based Reading and Writing: 615; Math: 530

How well do you think this school is doing?

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

Group 2: Separate Evaluation Condition: White School



School A

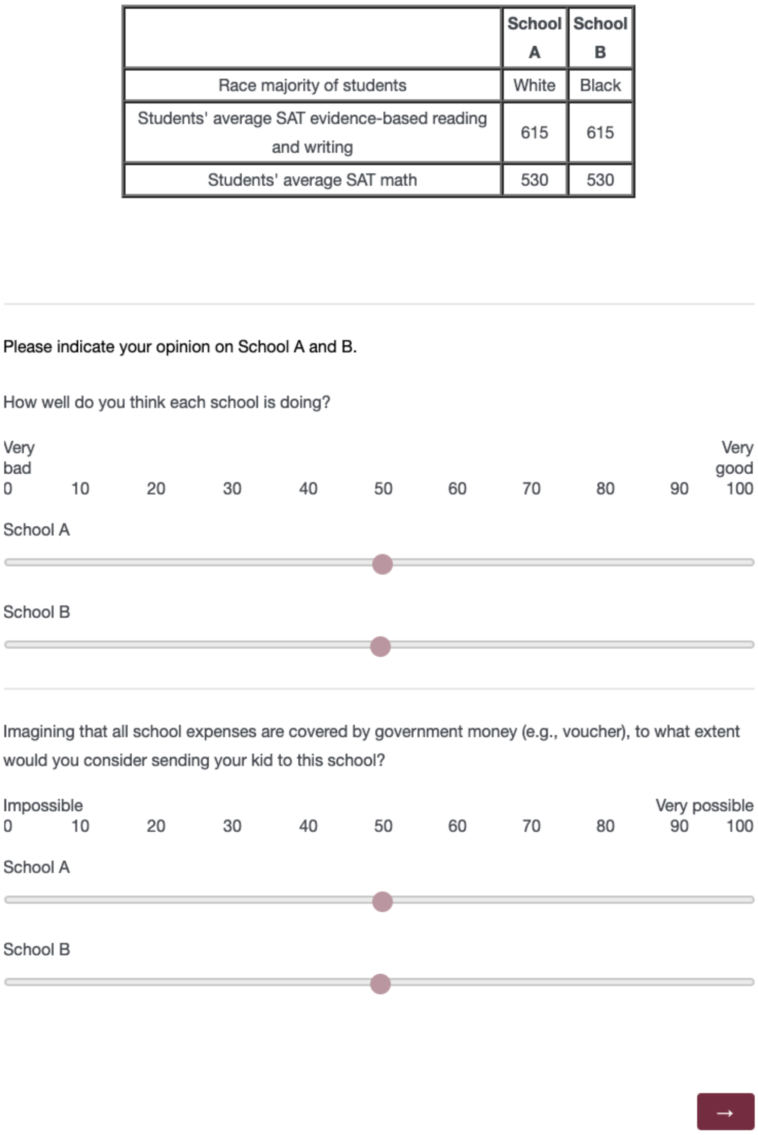
Race majority of students: White

Students' average SAT scores:
Evidence based Reading and Writing: 615; Math: 530

How well do you think this school is doing?

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

Group 3: Joint Evaluation Condition



	School A	School B
Race majority of students	White	Black
Students' average SAT evidence-based reading and writing	615	615
Students' average SAT math	530	530

Please indicate your opinion on School A and B.

How well do you think each school is doing?

Very bad 0 10 20 30 40 50 60 70 80 90 Very good 100

School A

School B

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

Impossible 0 10 20 30 40 50 60 70 80 90 Very possible 100

School A

School B

[Manipulation check] So far, which information have you seen in the previous part of this survey?

- I only saw School A, and its race majority of students was white.
- I only saw School A, and its race majority of students was black.
- I only saw School A, and its race majority of students was Hispanic.
- I saw two schools. The race majority of students in School A was white, and that of School B was black.

Appendix A3: Characteristics of sample

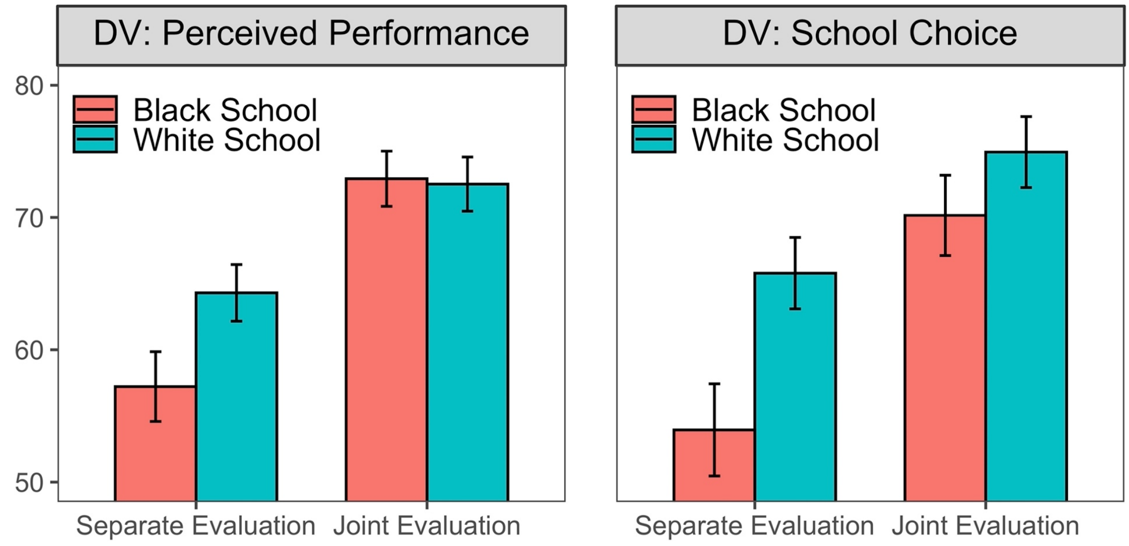
Table A1. Study 1 Sample

Note: P-values are generated from ANOVA *F*-tests.

	Total Sample		Separate Evaluation				Joint Evaluation		
			Black School		White School				
	N = 988		*N* = 316		*N* = 331		*N* =341		
	Frequency	%	Frequency	%	Frequency	%	Frequency	%	*P*-value
Female	520	53	178	57	166	50	176	52	0.25
Male	466	47	137	43	164	50	165	48	0.25
White	656	67	221	70	219	67	216	63	0.20
Black	130	13	37	12	38	12	55	16	0.14
Hispanic	87	9	24	8	31	9	32	9	0.65
Asian	85	9	27	9	27	8	31	9	0.92
Other	28	3	7	2	14	4	7	2	0.17
Age: 18-29	353	36	109	34	116	35	128	38	0.68
30-49	494	50	162	51	168	51	164	48	0.68
≥ 50	141	14	45	14	47	14	49	14	1.00
Income: < $25k	175	18	55	17	57	17	63	19	0.90
$25k to $75k	503	51	164	52	168	51	171	50	0.92
≥ $75k	308	31	97	31	105	32	106	31	0.95
College degree	592	60	189	60	197	60	206	60	0.98
Conservative	219	22	74	23	67	20	78	23	0.58
Liberal	462	47	139	44	151	46	172	50	0.22
Moderate	307	31	103	33	113	34	91	27	0.09
Parenthood	443	45	141	45	152	46	150	44	0.88

Appendix A4: Manipulation check (MC) and attention test (AT)

Figure A1. Study 1: Racial Stereotype
Note: This figure is generated with the sample who passed both MC and AT. Bars are 95% confidence intervals.



Appendix B: Study 2

Appendix B1: Pre-registration report

1. Have any data been collected for this study already?

No, no data have been collected for this study yet.

2. What's the main question being asked or hypothesis being tested in this study?

In this survey experiment, we ask:

Can joint evaluation mode reduce stereotyping in performance perception and encourage people to use rational thinking rather than heuristics?

3. Describe the key dependent variable(s) specifying how they will be measured.

In this study, we have two main dependent variables.

i. Performance perception: how well do you think this school is doing? (Moving a 0-100 scale bar: 0 = very bad; 100 = very good)

ii. Behavioral intention: Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school? (Moving a 0-100 scale bar: 0 = impossible; 100 = very possible)

4. How many and which conditions will participants be assigned to?

We will randomize participants in three groups: two separate evaluation groups (SE) and one joint evaluation group (JE).

In two SE groups, subjects will either see a a high school that its race majority of students is Black or a high school that its race majority of students is White. The only difference between SE groups is the students' race information (Black or White). In the JE group, subjects will see both schools that SE groups see.

We also show students' average SAT performances of each school in every condition. The
SAT scores are random numbers between 1000 to 1190.

In the both SE groups, subjects will be asked to answer both performance perception and behavioral intention questions after they see the school information. In the JE condition, subjects will be asked to answer both performance perception and behavioral intention questions for both schools they see.

5. Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Analyses will be based on linear regression models using experimental manipulations as the explanatory variables.

6. Any secondary analyses?

We will conduct subgroup analyses by participants' characteristics.

After manipulation, we will ask all subjects a manipulation check question and an attention test question. We will compare results with or without manipulation check and attention test failure samples to see the robustness of our findings.

7. *How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.*

We will stop data collection once 1000 subjects have submitted a responses on MTurk. Deviations from this goal are entirely due to MTurk software and outside of our control.

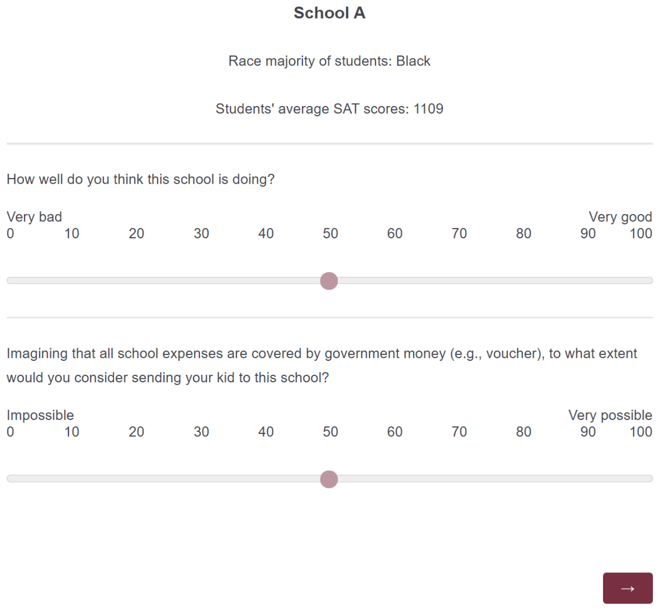
8. *Anything else you would like to pre-register? (e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)*

Subjects' demographic information will be collected after they have answered the questions regarding key dependent variables. The information is collected for detecting the heterogeneity of the treatment effect and for the randomization balance check. Since we only recruit adult subjects in the U.S., VPN and proxy identifier will be applied at the beginning to filter out disqualified subjects.

Appendix B2: Experimental intervention

After a brief introduction to the American high school scenario, we randomly assigned subjects into one of the following three conditions. The students' average SAT scores in below graphics are random numbers between 1000 to 1190.

Group 1: Separate Evaluation Condition: Black School



School A

Race majority of students: Black

Students' average SAT scores: 1109

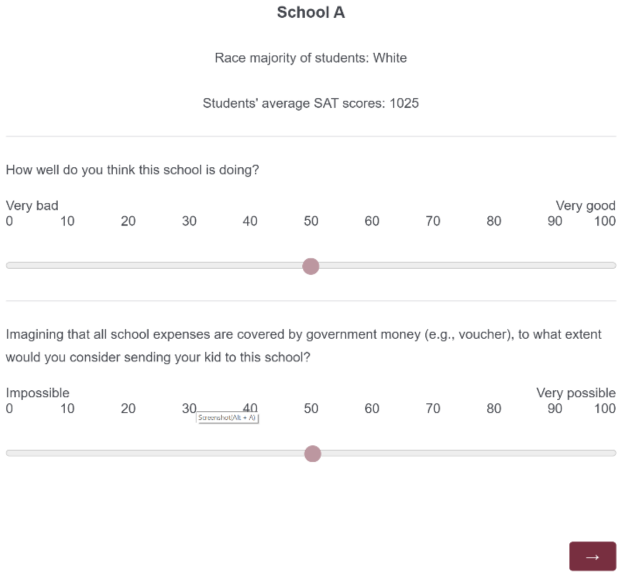
How well do you think this school is doing?

Very bad … Very good
0 10 20 30 40 50 60 70 80 90 100

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

Impossible … Very possible
0 10 20 30 40 50 60 70 80 90 100

Group 2: Separate Evaluation Condition: White School



School A

Race majority of students: White

Students' average SAT scores: 1025

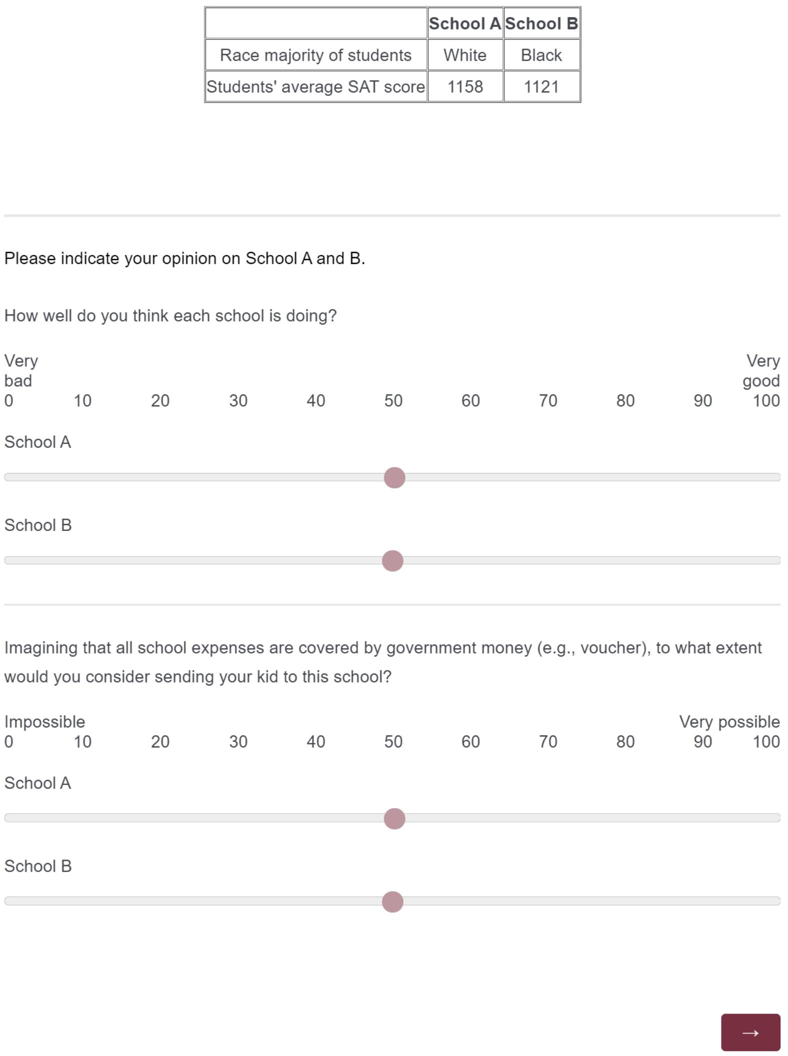
How well do you think this school is doing?

Very bad … Very good
0 10 20 30 40 50 60 70 80 90 100

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

Impossible … Very possible
0 10 20 30 40 50 60 70 80 90 100

Group 3: Joint Evaluation Condition



	School A	School B
Race majority of students	White	Black
Students' average SAT score	1158	1121

Please indicate your opinion on School A and B.

How well do you think each school is doing?

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

[Manipulation check] So far, which information have you seen in the previous part of this survey?

- I only saw School A, and its race majority of students was white.
- I only saw School A, and its race majority of students was black.
- I only saw School A, and its race majority of students was Hispanic.
- I saw two schools. The race majority of students in School A was white, and that of School B was black.

Appendix B3: Characteristics of sample

Table B1. Study 2 Sample

Note: P-values are generated from ANOVA *F*-tests.

	Total Sample		Separate Evaluation				Joint Evaluation		
			Black School		White School				
	N = 1002		*N* = 330		*N* = 338		*N* = 334		
	Frequency	%	Frequency	%	Frequency	%	Frequency	%	*P*-value
Female	473	47	152	45	165	50	156	47	0.42
Male	529	53	186	55	165	50	178	53	0.42
White	688	69	231	68	223	68	234	70	0.80
Black	96	10	31	9	34	10	31	9	0.85
Hispanic	77	8	27	8	22	7	28	8	0.69
Asian	117	12	39	12	43	13	35	10	0.58
Other	23	2	10	3	7	2	6	2	0.59
Age: 18-29	305	31	98	29	106	32	101	30	0.66
30-49	538	54	192	57	179	55	167	50	0.18
≥ 50	156	16	47	14	43	13	66	20	0.04
Income: < $25k	119	12	46	14	27	8	46	14	0.04
$25k to $75k	517	52	162	48	183	56	172	51	0.15
≥ $75k	364	36	129	38	119	36	116	35	0.63
College degree	634	63	211	62	212	64	211	63	0.86
Conservative	250	25	78	23	87	26	85	25	0.61
Liberal	475	47	162	48	160	48	153	46	0.76
Moderate	276	28	97	29	83	25	96	29	0.49
Parenthood	520	52	155	46	189	57	176	53	0.01

Appendix B4: Manipulation check (MC) and attention test (AT)

Figure B1. Study 2: Racial Stereotype
Note: This figure is generated with the sample who passed both manipulation check and attention test. Bars are 95% confidence intervals.

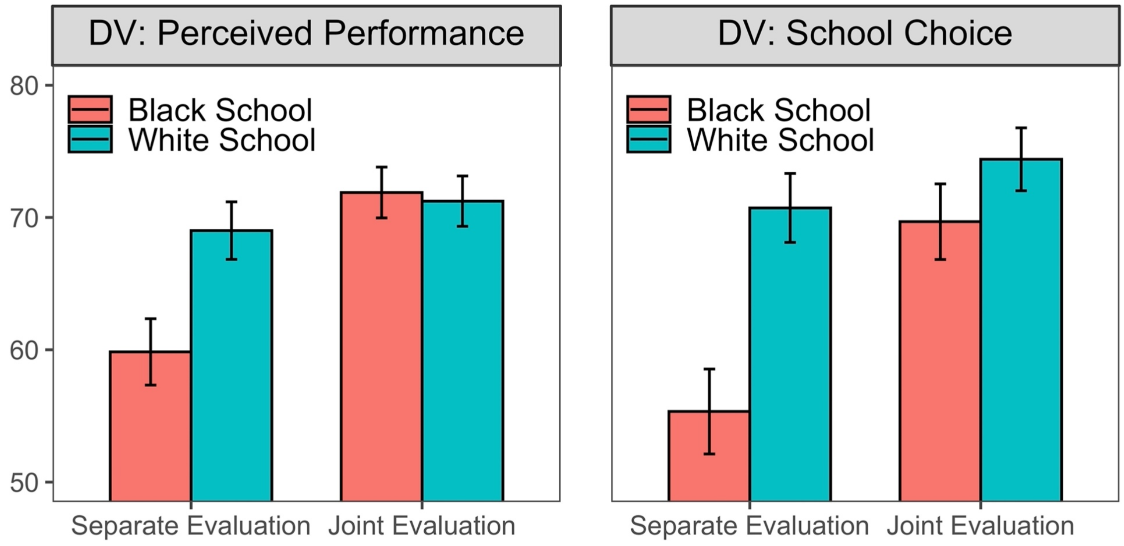
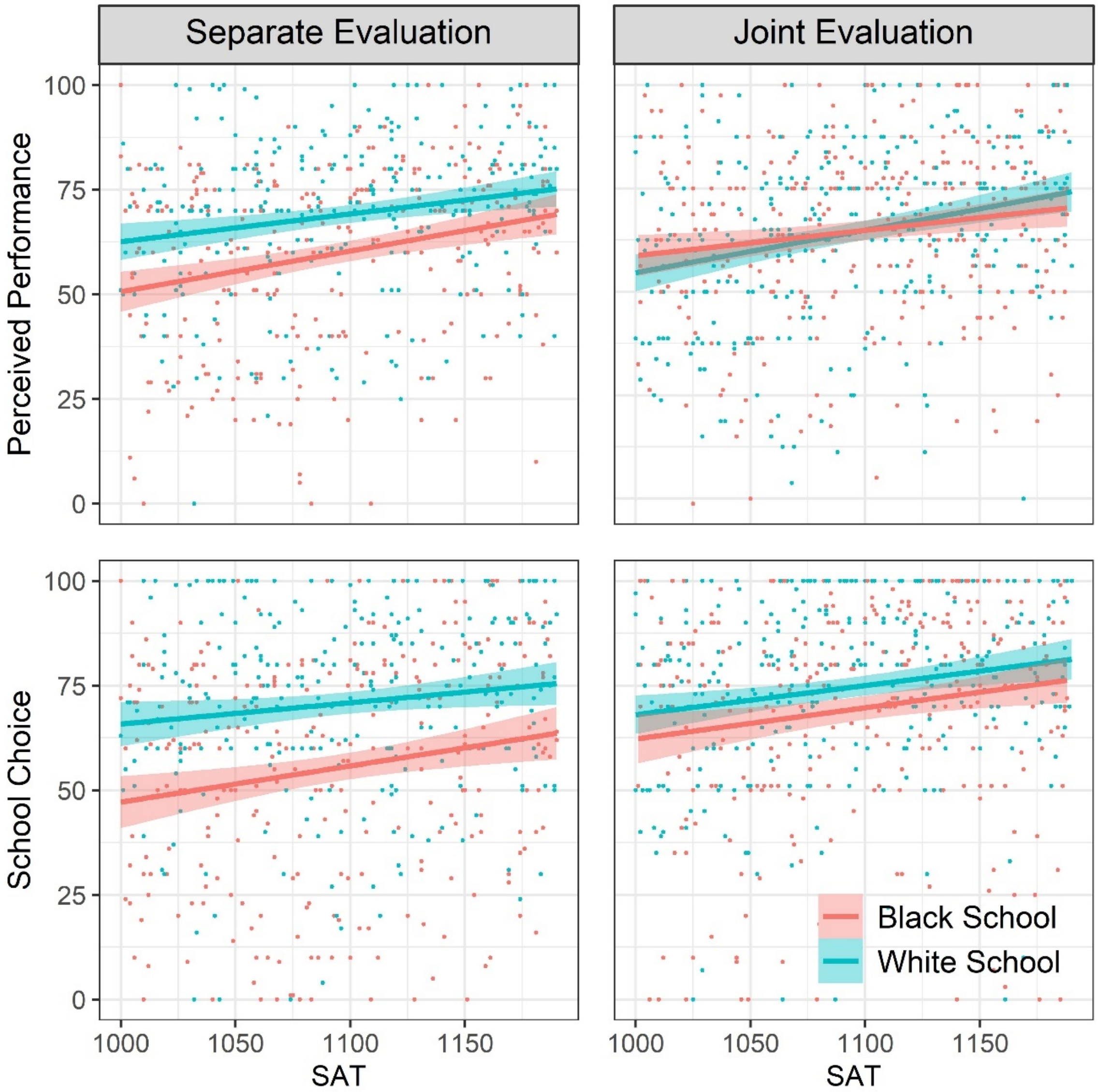


Figure B2. Study 2: Students' Major Race, SAT, and their Effects on Outcomes
Note: This figure is generated with the sample who passed both manipulation check and attention test. Bars are 95% confidence intervals.



Appendix C: Study 3

Appendix C1: Experimental intervention

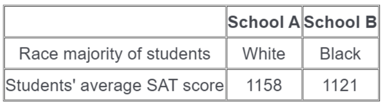
In the introduction section, subjects read:

Now, we invite you to share your opinion of high schools. Please imagine that you are under the situation that you are choosing a high school for your kid. Consider the following information carefully and answer related questions.

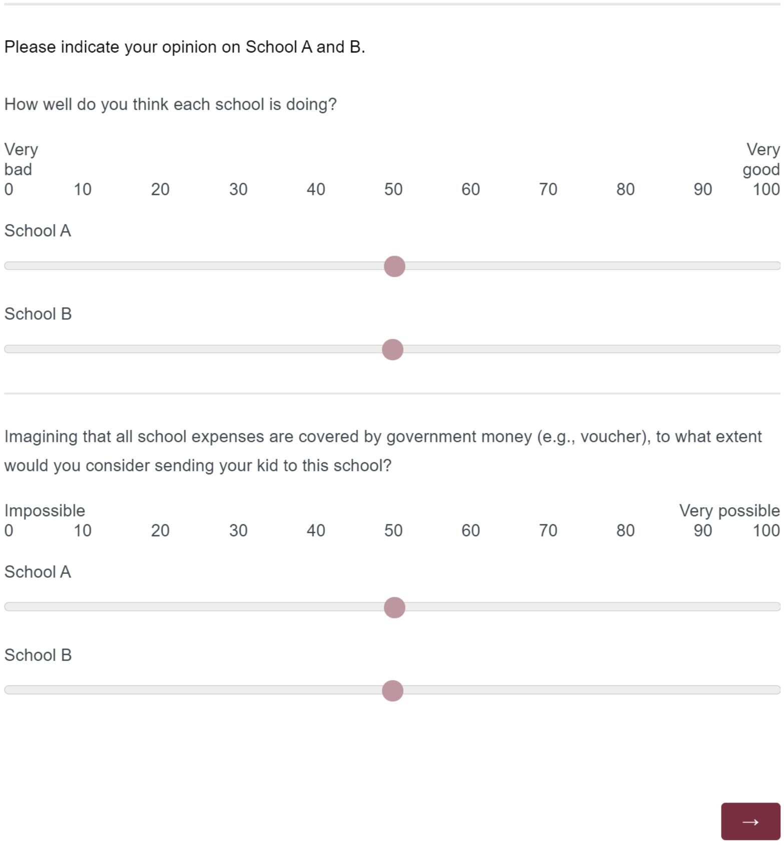
[Treatment] Only subjects in the demand group read the following information: (The purpose of this exercise is so we can measure whether school's race majority of students affects how likely people are to make judgment of a high school. We expect that people prefer schools where majority students are White than schools where majority students are Black because of the historical advantages White students have on education outcomes.)

NOTE: There are no right or wrong answers for these questions.

In the next page, we show the same information as Study 2. The students' average SAT scores in below graphics are random numbers between 1000 to 1190.



	School A	School B
Race majority of students	White	Black
Students' average SAT score	1158	1121



Please indicate your opinion on School A and B.

How well do you think each school is doing?

Very bad 0 10 20 30 40 50 60 70 80 90 100 Very good

School A

School B

Imagining that all school expenses are covered by government money (e.g., voucher), to what extent would you consider sending your kid to this school?

Impossible 0 10 20 30 40 50 60 70 80 90 100 Very possible

School A

School B

[Follow up question] If you had to guess, what do you think the researchers conducting this study are trying to learn by having you state opinions for both schools?
(Randomized question order)

- Whether people favor schools which race majority of students are White
- Whether people favor schools which students' average SAT scores are high
- Whether people favor schools which are tuition fee-free
- I don't know

Appendix C2: Characteristics of sample

Table C1. Supplemental Study Sample

Note: P-values are generated from t-tests.

	Total Sample		Demand Group		Control Group		
	N = 200		*N* = 115		*N* = 85		
	Frequency	%	Frequency	%	Frequency	%	*P*-value
Female	102	51	56	49	46	54	0.45
Male	98	49	59	51	39	46	0.45
White	153	76	91	79	62	73	0.31
Black	12	6	5	4	7	8	0.25
Hispanic	15	8	8	7	7	8	0.74
Asian	16	8	7	6	9	11	0.25
Other	4	2	4	3	0	0	0.08
Age: 18-29	65	33	36	31	29	35	0.63
30-49	111	56	70	61	41	49	0.09
≥ 50	23	12	9	8	14	17	0.05
Income: < $25k	34	17	16	14	18	21	0.19
$25k to $75k	113	57	68	60	45	53	0.35
≥ $75k	52	26	30	26	22	26	0.95
College degree	108	54	61	54	47	55	0.80
Conservative	45	22	28	24	17	20	0.47
Liberal	96	48	54	47	42	49	0.73
Moderate	59	30	33	29	26	31	0.77
Parenthood	106	53	64	56	42	49	0.38

Appendix D Demographic Questions

Study 1-3 shared the same set of demographic questions. These questions were asked after experimental interventions.

Are you…

- Male
- Female

Do you consider yourself to be…

- White, not Hispanic or Latino
- Black, not Hispanic or Latino
- Hispanic or Latino
- Asian, not Hispanic or Latino
- Other

Your age:________

Which state do you live in?

Do you have any children in the following school-age categories? (Check all that apply)

- Pre-school
- Elementary school
- Middle/intermediate school
- High school
- High school graduate/college
- NONE OF THE ABOVE or NO CHILDREN

What was your total household income before taxes during the past 12 months?

- Less than $25,000
- $25,000 to $34,999
- $35,000 to $49,999
- $50,000 to $74,999
- $75,000 to $99,999
- $100,000 to $149,999
- $150,000 or more

What is the highest level of education you have completed?
Less than high school

- High school/GED
- Some college
- 2-year college degree
- 4-year college degree
- master degree

- doctoral degree
- Professional Degree (JD, MD)

When comes to social issues, I am…

- Very liberal
- Liberal
- Moderate
- Conservative
- Very conservative

[Attention test] This is just to screen out random clicking. Please move the slide to the answer of
the following question: 17 + 63 =?