# De-stereotyping Public Performance Evaluation

Yixin Liu and Chengxin Xu[*]

December 16, 2020

## Abstract

Experimental evidence suggests that citizens' judgments of service quality often rely on prior beliefs about providers' characteristics, such as sectoral or 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 ($N > 3,000$), and obtained similar results in these studies: Subjects used sectoral or racial stereotypes to evaluate school performance in the SE condition, but the 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 behaviors.

**Keywords**: Performance information, stereotyping, joint evaluation, online experiment

[*] Yixin Liu is a PhD candidate at Askew School of Public Administration, Florida State University, yl17g@my.fsu.edu. Chengxin Xu is an assistant professor at the Institute of Public Service, Seattle University, cxu1@seattleu.edu. Alphabetical author order based on equal contributions.

We thank Asmus Leth Olsen and Donald Moynihan for feedbacks during different stages of this project. We are also grateful to insightful comments from panel chairs Lars Tummers and Sharon Gilad, and participants at the 2020 Annual Conference of the European Group for Public Administration: Behavioral Public Administration Sessions.

# Introduction

People's prior stereotypical knowledge of the public organization may bias their understanding on its performance information. From three online experiments, this study shows that placing performance information jointly for people's evaluation can discourage stereotyping and amplify the importance of performance data 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 behavioral 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. The findings of cognitive biases in public performance evaluations have indicated that, under certain circumstances (e.g., when information is framed in specific ways or colored by political identities), public performance information fails to ensure an accurate basis for the public to evaluate how well or poorly the public agency or service organization is performing (James et al.

2020). From the perspective of political accountability, such a biased understanding may result in public managers' problematic understanding of citizens' opinion of public services and lead to inaccurate decisions (Andersen and Hjortskov 2016). At the same time, from the perspective of the marketization of public services, the "last mile problem" in performance evaluations may undermine the role of the performance information system for public and social services, which is applied to address the information asymmetry in the public service market. A biased understanding of service quality may lead consumers to choose inappropriate or poor service providers, which may undermine public policies' effectiveness (Weimer 2020).

To retain performance information's important democratic and market 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: place performance information jointly rather than separately for evaluation. We applied this strategy to de-stereotype performance evaluations on the part of the general public. Our strategy is based on extensive observations in decision making literature showing that people are more likely to make reasoned judgments when evaluating two subjects jointly than separately (Bohnet, Van Geen, and

Bazerman 2016; Hsee et al. 1999; Li and Hsee 2019). In the context of performance evaluation, the joint evaluation mode (JE) offers more data as reference points that help people to understand the relative performance of the organization. This advantage over the separate evaluation mode (SE) is compatible with people's behavioral 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's evaluation and relative decisions are made based on the performance rather than stereotypical beliefs.

We conducted three online survey experiments on Amazon Mechanical Turk (MTurk) in the context of public high schools in the U.S., and manipulated the evaluation mode and schools' characteristics on which the public maintains stereotypical beliefs. Studies 1 and 2 showed that people's evaluations of high schools were driven by negative stereotypes of Black high schools and public high schools in the SE mode given the same school performance, while their evaluations were based on performance data in the JE mode. Study 3 employed a conjoint design that showed that the effect of performance data ignored in the SE mode was amplified in the JE mode.

Our study shows the role of the JE mode in offsetting stereotyping 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 eliminate bias is scant. The current debiasing literature considers the normative model of decisionmaking 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. 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). Early experiments on debiasing in public administration focus on modifying the decision makers, who are largely politicians, public managers, and officials. 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. 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 non-elite citizens and politicians. Recent contributions to debiasing strategies also include Cantarelli, Bellé, and Belardinelli (2020) study, which test educational approaches to eliminate the band-wagoning and framing effects, and Nagtegaal et al. (2020) work, which provide experimental support for the "consider-the-opposite" technique in mitigating public managers and employees' anchoring bias.

Examinations in eliminating cognitive bias 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). It 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). Stereotyping leads to biased judgments because it ignores information about the specific organization's features, and the impression of the category, based on which the judgment is made, exaggerates one of that category's representative characteristics (Bordalo et al. 2016).

Previous evidence shows that people evaluate public organizations' performance based on stereotypical understanding 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 (2018) survey experiment indicate that although Whites in the U.S. rate the performance of the police favorably overall, they rate their performance, trustworthiness, and fairness less favorably 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).

Another important signal that triggers stereotyping is the sector in which the organization operates. Given private and nonprofit organizations' increasing participation in areas of public service delivery, public organizations' performance in citizens' eyes are influenced by their negative perceptions of the government, which biases their actual performance evaluations. As Marvel (2016, 143) indicate, "Citizens automatically and unconsciously associate public sector organizations with inefficiency, inflexibility, and other pejoratives, and these automatic associations color their assessments of public sector performance" (See also: Hvidman and Andersen 2016; Meier, Johnson, and An 2019; Xu 2020).

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

One potential strategy for de-stereotyping is to change people's evaluation mode from the separate to the joint mode. Evaluation mode is how a decision is made. 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. 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, public organizations are generally inefficient. 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). 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 labor 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. A simple shift in the evaluation mode, from SE to JE, may also change people's decision making in consumption and policy preferences (e.g., List 2002; Milkman et al. 2012).

The JE mode makes performance data more evaluable because it is compatible with people's information processing behavior, 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 centered 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 judgments on the organizational performance by comparing and benchmarking, and in turn, their evaluation will be more likely to be driven by the actual performance data.

### *Hypotheses*

Following the theoretical mechanism of the JE mode, this study tests the following hypotheses:

**H1** Organizational characteristics in the JE mode are less likely to determine people's perceived performance of the organization than in the SE mode.

**H2** Performance data in the JE mode is more likely to determine people's perceived performance of the organization than in the SE mode.

We test these hypotheses in the context of high school performance in the U.S. Examining debiasing strategies 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 behaviors. 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., for example, segregation and the competition between public vs. private schools, have established widely accepted stereotypes of school performance that may influence people's understanding of the actual performance of one particular school.

Following our discussion on the evaluation mode in terms of JE and SE, as Black students have long suffered from a negative stereotype about their group academic performance (Steele and Aronson 1995) and the unequal distribution of funding to Black public schools (Bifulco 2005), the negative perception of Black students and schools in general could influence the perceptions of the school's performance when they are evaluated in isolation (SE). Such a stereotyping process leads to a biased evaluation of a specific Black high school compared to its White counterpart. However, when the Black school's performance is evaluated jointly with its White counterpart (JE), negative prior belief on Black schools will be updated, and the two schools' performance will be evaluated as equivalent given the same performance level. Since people's school choices are associated with their perceived performance, we anticipate that their school choices decision will also follow the same hypothesis.

Schools' ownership, whether public or private, also relates to both actual and perceived differences in school performance. With customized curricula, smaller class sizes, as well as different student populations, private schools have theoretical advantages over public schools. Some statistics have also shown that, in college, graduates from private high schools outperform their counterparts from public high schools from different perspectives (Torres 2018). In turn, people may hold a strong belief that public high schools perform poorly even when they have same level of performance as private schools in the SE mode. However, as the JE mode provides reference points that enhance the ability to evaluate numeric performance information, people will treat the two schools in the same way given the same performance information.

The JE mode also amplifies the effect of performance data in determining people's perceived performance of the school and their school choice intention. Therefore, when in the JE mode where the data shows that one school out-performs the other, people will be more likely to rely on such comparison and make judgments and related choices. However, since the SE mode does not facilitate the comparison, people will be less likely to use the performance data to judge the school performance because it is difficult to evaluate.

## 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 and Study 2 tests H1, and Study 3 tests H2. All experiments are pre-registered at [anonymous for peer review] (pre-registration reports for three studies are available at: appendix B i, appendix C i, appendix D i).

We used the same measurements of performance evaluations in all studies: Perceived performance and behavioral intention. 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 behavioral intention. We included manipulation check and attention test questions after the survey instruments for the dependent variables to detect any treatment noncompliance or random choice behavior. 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 (appendix A).

### *Study 1: De-stereotyping Black vs. White*

Study 1 tests H1 by examining whether the negative stereotype of the Black high schools' 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, 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.

***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 behavioral intention (see appendix B ii for the full survey). Figure 1 shows the experimental procedure. The unit of analysis of the study is the school profile rather than the individual.

[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[1]. After randomization, 316 subjects were assigned to evaluate the performance of the school in which the racial majority of the students is Black (Black school), 331 subjects were assigned to evaluate the school performance in which the racial majority of the students is White (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 are in SE mode and 682 profiles are in JE mode. Figure 2 illustrates the results of random assignment, and appendix B iii: table B.1 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 6.48% 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 behavioral intention showed similar results. Subjects in the SE mode were 10.07% 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 3.99% 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 or White) and their interaction to predict subjects' perceived performance and behavioral intention regarding one school profile. 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: $(JE_A – JE_B) – (SE_A – SE_B) = 0$, in which stereotypical alternatives were denoted as A and B (in Study 1, the alternatives are Black and White students). In the following regression model, $\beta_3$ 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.

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

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 6.65% (S.E. = 1.67, p = 0.00) and that in behavioral intention by 6.08% (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 H1. It shows that in the SE mode, when given the same average SAT score, people downgraded the performance of the Black high school stereotypically, but the difference attributable to racial stereotyping was mitigated in the JE mode. This indicates that the JE mode nudged people to evaluate school performance based on performance data rather than on prior stereotypical beliefs on student racial majorities.

### *Study 2: De-stereotyping Public vs. Private*

The purpose of Study 2 was to test this SE-JE effect using the public/private ownership stereotype to examine the ability to generalize our theory. We expect that when given the same performance information (SAT score), people will stereotypically downgrade the public high school's performance in SE mode and rate public and private schools' performance more closely in the JE mode.

***Design***. Study 2 followed the similar experimental procedure of Study 1. First, we warmed up subjects with fact charts about public/private schools from academic reports. The aim of this step was to activate subjects' associations related to school ownership in the following school evaluation task (Baekgaard and Serritzlew 2016; James and Van Ryzin 2017). Adding a warm-up message would not affect causal inference because subjects in all experimental conditions have read the same message, thus any observed differences between profiles are due to experimental manipulations. Next, we randomly assigned subjects to three groups to evaluate school performance: Two SE groups with either a public or private school (randomly) and one JE group showing both public and private schools. We then collected data on perceived performance, behavioral intention, and demographics. Figure 3 shows the experimental procedure.

[Figure 3 is here]

***Results***. We recruited 804 subjects (51% female, 67% White, $M_{age} = 36$), 619 (77%) of whom passed both the manipulation check and attention test. After randomization, 262 subjects evaluated a private school, 276 a public school, and 266 both schools simultaneously. In total, subjects evaluated 1070 school profiles, in which 538 profiles were in the SE mode and 532 profiles were in the JE mode. Detailed information on our sample and the randomization check (F test) is reported in appendix C iii: Table C.1.

We employed the same analytical strategy as in Study 1. Figure 4 illustrates the t-test between ownership alternatives under both the SE and JE modes. In general, we replicated our findings in Study 1. The subjects rated the private school's performance 8.18% higher (C.I. = [5.09, 11.27], p = 0.00) than the public school's in the SE mode. This difference decreased to 2.63% (C.I. = [-0.33, 5.58], p = 0.08) in the JE mode. We also found similar effects in behavior intention. The subjects had 5.75% stronger intentions (C.I. = [1.79, 9.72], p = 0.00) to send their children to private rather than public schools in the SE mode, and this difference decreased to 3.68 (C.I. = [-0.43, 7.80], p = 0.07) in the JE mode.

[Figure 4 is here]

Models in table 2 without the covariate adjustment (Models 1 & 3) suggest that subjects decreased their difference attributable to their sector stereotype in the schools' performance evaluations by 10.81% (S.E. = 1.85, p = 0.00); they also decreased their degrees of difference in their sector stereotype in school choice by 9.44% (S.E. = 2.78, 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]

***Discussion***. Findings of Study 2 also support H1. It showed that the JE mode reduced the gaps in perceived performance and preferences for high schools attributable to the subjects' sector stereotype effectively. Thus, both Studies 1 and 2 documented the JE mode's effectiveness in reducing biases attributable to stereotyping in school performance evaluations.

### *Study 3: Conjoint Experiments*

Study 3 has two purposes. First, it aimed to test H2 systematically by manipulating the performance information with different levels. 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. Second, Study 3 also aimed to assess the robustness of the results in Studies 1 & 2 in a more complex, real-world information environment. In particular, the outcome measure in Study 1 might be suffered from social desirability bias, as people might hesitate to answer racially sensitive questions truthfully.

Study 3 employed a conjoint experimental design that has two major advantages. First, the design allows us to manipulate the stereotypical attributes and the performance attributes simultaneously and compare their effect sizes. If performance attributes outweighed stereotypical attributes in performance evaluations, it indicates that people increased their use of performance information rather than stereotypical information in the JE mode. Second, we combined multiple pieces of stereotypical and performance information in a complex multi-dimensional conjoint evaluation task. This task not only prevented the subjects from answering sensitive questions directly, but also made our survey design closer to real world situation where people are surrounded with different information (Hainmueller, Hangartner, and Yamamoto 2015; Horiuchi et al. 2020).

***Design***. The subjects were assigned randomly to the SE or JE condition. The SE condition contained a single conjoint experiment, in which subjects read a single conjoint school profile; the JE condition contained a pair conjoint experiment, in which subjects read a pair conjoint of school profiles. Consistent with Study 1 & 2, the unit of analysis of conjoint experiment is school profile, and each subject read one school profile in the SE condition and two profiles in the JE condition. To balance the group size of each mode, we assigned two thirds of the subjects randomly to the single conjoint condition and one third to the pair conjoint. Both single and pair conjoint designs are common in conjoint studies (see Bansak et al. 2019; Hemker and Rink 2017; Stokes and Warshaw 2017). Other than the evaluation mode, we held everything the same between single and pair conjoint, including a same set of attributes and their values (similar design as Hainmueller et al. 2015). Figure 5 shows the experimental procedure.

[Figure 5 is here]

We constructed the school profile by nesting two stereotypical attributes (students' race and school ownership) and two school performance attributes (SAT and learning environment). Table 3 lists our conjoint attributes and possible values. Every school profile contained these four attributes, and their values were presented randomly. We also randomized the order of attributes across subjects. Appendix D ii provides examples of the subject interfaces for both designs.

[Table 3 is here]

The main measure of interest in this study included the Average Marginal Component Effect (AMCE) (Equation 2) and difference-in-AMCEs (Equation 3). AMCE identifies the average change in the profile evaluation when one attribute value was alternated with the other (Hainmueller, Hopkins, and Yamamoto 2014). For example, it allows us to compare the marginal effect on school evaluation between "public ownership" and "private ownership", holding all other possible attribute values at average levels. In our fully randomized conjoint design, all attributes are strictly exogenous. Therefore, we regressed them as explanatory variables in an ordinary least squares regression and their coefficients are AMCEs. We ran two separate models for single and pair conjoint experiments to see whether they output different effects from attributes. With this strategy, we can examine how subjects weight performance information and stereotypical information when evaluating schools in each evaluation mode. We clustered standard errors by individuals to control non-independence in the pair conjoint group.

$$Y = \beta_1 Race + \beta_2 Ownership + \beta_3 SAT + \beta_4 Learning + C_i^{'}\gamma + \mu \qquad \text{(Equation 2)}$$

Moreover, we also regressed both single and pair conjoint groups in one integrated model to examine the Difference-in-AMCEs (Equation 3). Difference-in-AMCEs describes difference in preference between the groups, which captures group difference for a particular attribute's effect size (Leeper, Hobolt, and Tilley 2020). In our design, it is the coefficient of interaction term between each attribute and the evaluation mode ($\delta_1, \delta_2, \delta_3, \delta_4$). It demonstrates how evaluation mode change affects subjects' weight on a particular attribute in their school evaluation process. We also clustered standard errors by individuals in this model.

$$Y = \beta_1 Race + \beta_2 Ownership + \beta_3 SAT + \beta_4 Learning + \omega Mode + \delta_1 Race \times Mode+$$
$$+ \delta_2 Ownership \times Mode + \delta_3 SAT \times Mode + \delta_4 Learning \times Mode + C_i^{'}\gamma + \mu \quad \text{(Equation 3)}$$

***Results***. The final sample size in this study was 1502 (52% female, 67% White, $M_{age} = 35$), and 1342 (89%) subjects passed both the manipulation check and attention test. 961 subjects were assigned to the single conjoint group and 541 to the conjoint pair group. In total subjects evaluated 2043 school profiles, in which 961 school profiles in the SE mode and 1082 profiles in the JE mode. Appendix D iii: table D.1 provides the sample characteristics and t-test randomization check.

Figure 6 shows the SE and JE's AMCE in predicting both dependent variables. The benchmark models for this figure are in appendix D iv table D.2. In the upper panel of figure 6, the subjects did not base their performance evaluations on stereotypes in either evaluation mode, and instead, used the learning environment primarily to rate schools in the SE mode, and both learning environment and SAT scores in the JE mode. In the lower panel of figure 6, the subjects' school choice was based on students' race and learning environment in the SE mode. When the students' racial majority switched from White to Black, the subjects' school preference decreased by 4.19% (S.E. = 1.40, p = 0.00); when performance measured by the learning environment improved, the subjects' school preference increased by 6.89% (S.E. = 1.42, p = 0.00). In contrast, students' racial majority had no effect in the JE mode, where the subjects' decisions relied on the schools' merits. In addition, ownership had no effect in both conditions.

[Figure 6 is here]

The difference-in-AMCEs in figure 7 (benchmark models with interaction terms are in appendix D iv table D.4) shows that the evaluation mode change did not affect people's use of stereotypical information to evaluate school performance. However, people held a stronger preference for a high SAT score in the JE than the SE mode, but a weaker preference for a positive learning environment in the JE than the SE mode. Further, although subjects' significant racial stereotype on students regarding school choice in SE mode have become insignificant in JE mode, this change was statistically indistinguishable (difference = 2.46, S.E. = 1.90, p = 0.20).

[Figure 7 is here]

***Discussion***. The results of Study 3 are inconsistent with Study 1 and 2 and only partially support H2. In both single (SE) and pair conjoint (JE) experiments, the effect sizes of stereotypical information were much smaller than performance information regarding perceived performance. We observed that one of the performance attributes, the perceived learning environment, was consistently important in both evaluation modes. In contrast, the effect of SAT score was only amplified in the JE mode.

Findings of Study 3 regarding stereotypical attributes are inconsistent with Study 1 & 2 probably because the performance data on learning environment is the most evaluable attribute in the school profile. Similar to the mechanism of the JE mode, more evaluable information can update people's prior knowledge of the school and make decision accordingly. In our case, the learning environment is more evaluable than others probably for two reasons. First, it was presented in percentages, which, compared with absolute number for SAT, provides more information of relative performance. Second, people maintain a leftmost-digit-bias (Olsen 2013). The better SAT performance "1280" and the worse "1200" shared the same leftmost digits, so subjects could not detect a difference between them in the SE mode. In contrast, the perceived learning environment was presented as “80%” vs. “70%,” so the subjects may not have needed additional reference points to improve their ability to evaluate this attribute in the SE mode, since “80%” may serve as a natural threshold for good school performance.

However, we did find that the effect of SAT score was amplified in the JE mode in perceived performance and school preference, supporting H2. Although the effect of learning environment was reduced in the JE mode, its effect size was still significantly driving people’s evaluation and school choice. Therefore, people placed more balanced attention to both performance attributes in the JE mode than in the SE mode.

## Conclusion

Cognitive biases in evaluating and understanding public organizations’ performance ultimately undermine people’s wellbeing. With more biases such as stereotyping being reported, finding and testing debiasing strategies become more urgent. This study proposes that the JE mode can be considered as an effective nudge to amplify the effect of performance information in evaluation and discourage stereotyping, if any. Our experimental results support our hypotheses partially. Studies 1 and 2 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. Although Study 3 did not show stereotyping in the SE mode, it did indicate that the JE mode can amplify the effect of performance information which is relatively hard to evaluate in the SE mode, so that evaluators can consider performance indicators with different levels of evaluability more equally. In sum, this study has several implications for public management.

First, 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, and such findings were replicated with different stereotypes. People’s stereotypical knowledge of social groups and types of organization 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. The JE mode provides a starting point of ways to encourage evidence-based judgments.

Second, the JE mode can amplify the role of certain type of performance information in evaluation. Even when stereotyping does not drive people's judgment, we show that people pay more attention in the JE mode than in the SE mode to less evaluable performance information, such as the average SAT score of the high school. In our study, the effect of SAT score was balanced with the indicator of students' learning environment in the JE mode for both people's perceived performance and school choice decision.

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 these 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.

We acknowledge several limitations of our study. 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, 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 behavioral change, which is impossible to determine in survey experiments. 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 behaviors 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 behavior 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 problem, the JE 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 behaviors.

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. Unlike the non-elite subjects, public officials have stronger prior political knowledge and beliefs, which often renders political biases more resistant to behavior 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.

## Notes

[1] Analyses of the samples that excluded people who failed to pass the manipulation check and attention test showed robust results in all studies (see see appendix B iv, appendix C iv, and appendix D v).

# 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.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins, Teppei Yamamoto, JN Druckman, and DP Green. Conjoint survey experiments 2019.

Battaglio Jr, R Paul, Paolo Belardinelli, Nicola Bellé, 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.

Cantarelli, Paola, Nicola Bellé, 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, Étienne 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*.

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.

Hainmueller, Jens, Daniel J Hopkins, and Teppei Yamamoto. 2014. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis* **22** (1):1–30.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences* **112** (8):2395–2400.

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

Hemker, Johannes and Anselm Rink. 2017. Multiple dimensions of bureaucratic discrimination: Evidence from german welfare offices. *American Journal of Political Science* **61** (4):786–803.

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.

Horiuchi, Yusaku, Zachary D Markovich, and Teppei Yamamoto. 2020. Does conjoint analysis mitigate social desirability bias?

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.

James, Oliver, Asmus Leth Olsen, Donald Moynihan, and Gregg G Van Ryzin 2020. *Behavioral Public Performance: How People Make Sense of Government Metrics*. Cambridge University Press.

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.

Leeper, Thomas J, Sara B Hobolt, and James Tilley. 2020. Measuring subgroup preferences in conjoint experiments. *Political Analysis* **28** (2):207–221.

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.

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.

Olsen, Asmus Leth. 2013. Leftmost-digit-bias in an enumerated public sector? an experiment on citizens' judgment of performance information. *Judgment and Decision Making* **8** (3):365.

——. 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*.

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.

Stokes, Leah C and Christopher Warshaw. 2017. Renewable energy policy design and framing influence public support in the united states. *Nature Energy* **2** (8):1–6.

Torres, Amada. 2018. New NAIS-Gallup Report on Student Outcomes. *Independent School Magazine*. URL https://www.nais.org/magazine/independent-school/winter-2018/life-lessons/.

Weimer, David L. 2020. When are nudges desirable? benefit validity when preferences are not consistently revealed. *Public Administration Review* **80** (1):118–126.

Xu, Chengxin. 2020. The perceived differences: The sector stereotype of social service providers. *Nonprofit and Voluntary Sector Quarterly* page 0899764020925903.

# Figure and Table

## *Study 1*

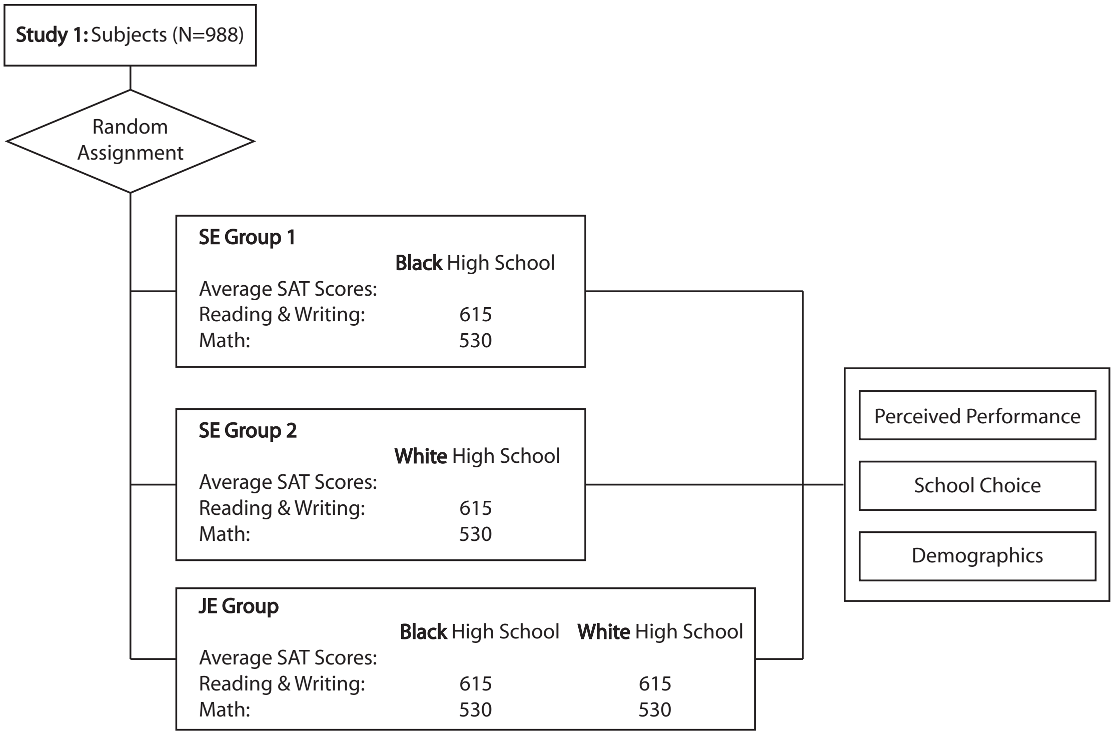


Figure 1: **Study 1 Experimental Procedure**

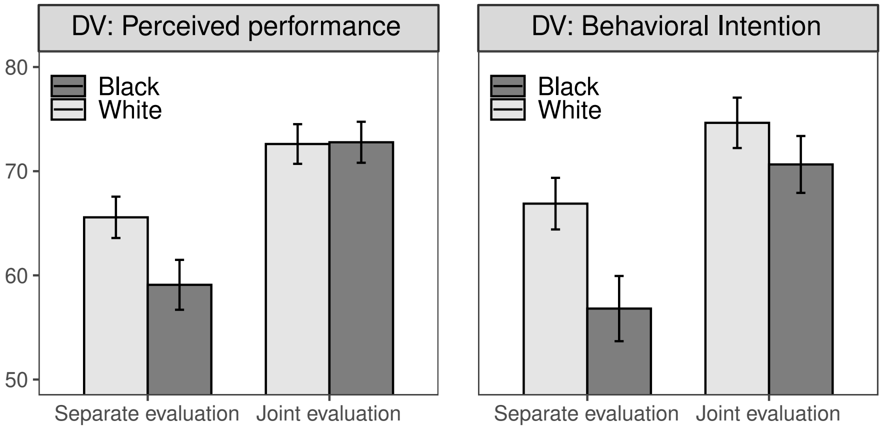


Figure 2: **Racial Stereotype in SE and JE**

*Note*: Bars are 95% confidence intervals.

Table 1: **Racial Stereotype in SE and JE**

| | Perceived Performance | | Behavioral Intention | |
|---|---|---|---|---|
| | (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 |
| State FE | No | Yes | No | Yes |
| Observation | 1,328 | 1,320 | 1,327 | 1,319 |
| Adjusted R$^2$ | 0.078 | 0.102 | 0.063 | 0.085 |

*Note*: OLS estimates. *Mode* (baseline: *SE*) and *Students' major race* (baseline: *White*) are dummies. Clustered standard errors are in brackets. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

### *Study 2*

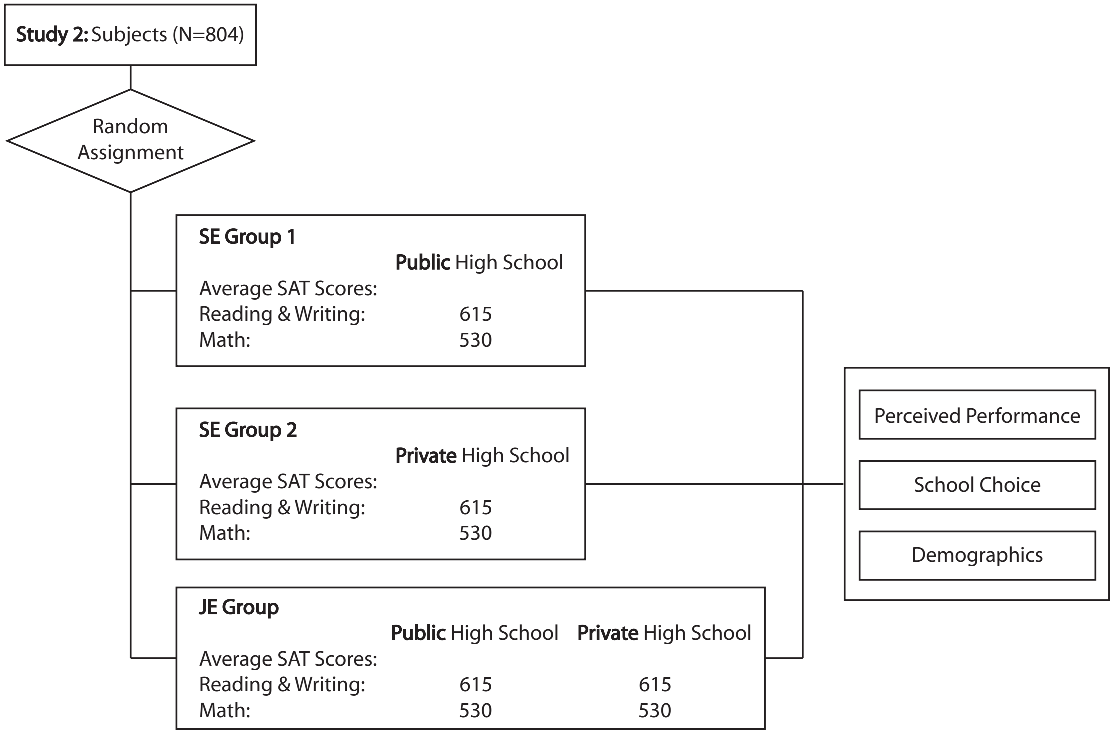


Figure 3: **Study 2 Experimental Procedure**

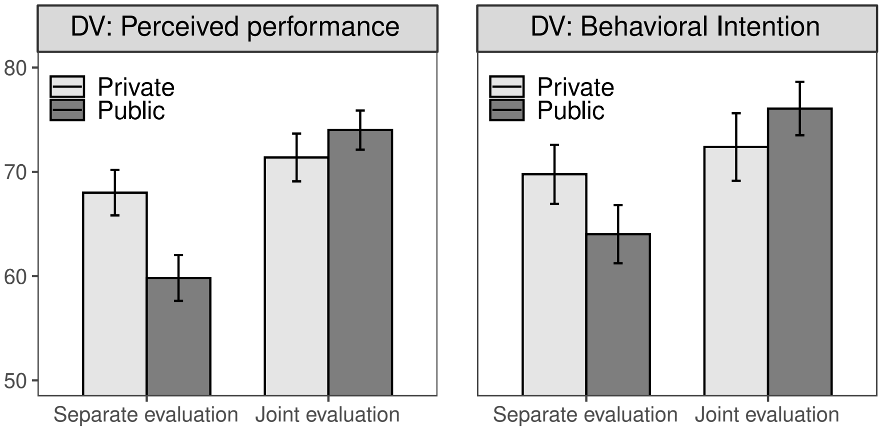


Figure 4: **Sector Stereotype in SE and JE**

*Note*: Bars are 95% confidence intervals.

Table 2: **Sector Stereotype in SE and JE**

| | Perceived Performance | | Behavioral Intention | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| JE×Public | $10.809^{***}$ | $10.171^{***}$ | $9.437^{***}$ | $9.277^{**}$ |
| | $(1.846)$ | $(1.902)$ | $(2.782)$ | $(2.911)$ |
| Mode: JE | $3.372^{*}$ | $4.134^{*}$ | $2.616$ | $3.240$ |
| | $(1.610)$ | $(1.623)$ | $(2.182)$ | $(2.263)$ |
| Sector: Public | $-8.181^{***}$ | $-7.543^{***}$ | $-5.752^{**}$ | $-5.593^{**}$ |
| | $(1.574)$ | $(1.637)$ | $(2.018)$ | $(2.117)$ |
| Constant | $68.004^{***}$ | $68.949^{***}$ | $69.763^{***}$ | $70.065^{***}$ |
| | $(1.111)$ | $(5.756)$ | $(1.437)$ | $(7.311)$ |
| Covariates | No | Yes | No | Yes |
| State FE | No | Yes | No | Yes |
| Observation | 1,070 | 1,063 | 1,070 | 1,063 |
| Adjusted $R^2$ | 0.081 | 0.111 | 0.031 | 0.019 |

*Note*: OLS estimates. *Mode* (baseline: *SE*) and *Sector* (baseline: *Private*) are dummies. Clustered standard errors are in brackets. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

***Study 3***

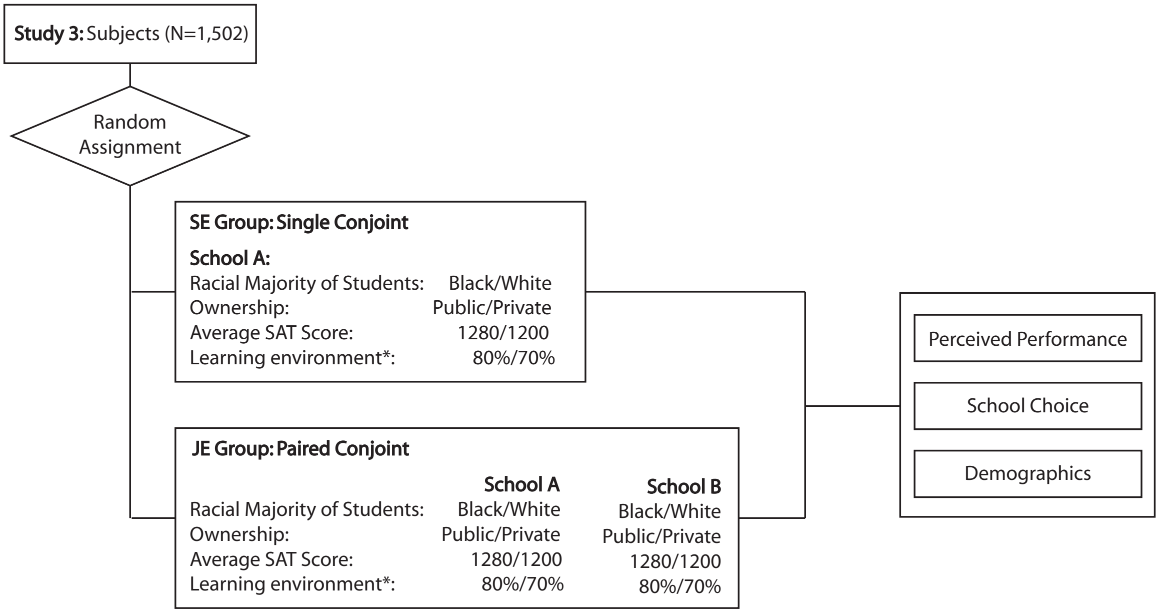

Study 3: Subjects (N=1,502)
Random Assignment
SE Group: Single Conjoint
School A:
Racial Majority of Students: Black/White
Ownership: Public/Private
Average SAT Score: 1280/1200
Learning environment*: 80%/70%
JE Group: Paired Conjoint
School A
School B
Racial Majority of Students: Black/White Black/White
Ownership: Public/Private Public/Private
Average SAT Score: 1280/1200 1280/1200
Learning environment*: 80%/70% 80%/70%
Perceived Performance
School Choice
Demographics

Figure 5: **Study 3 Experimental Procedure**

*Note*: Learning environment is measured by the percentage of students feeling supported to meet high expectations in learning.

Table 3: **Attributes for School Profile in Conjoint Task**

| Attributes | Values |
|---|---|
| **Stereotype attribute** | |
| Race majority of students | (1) Black |
| | (2) White |
| Ownership | (1) Public |
| | (2) Private |
| **Performance information** | |
| Students' average SAT score | (1) 1280 |
| | (2) 1200 |
| % of students feel supported to meet high expectations in learning | (1) 80% |
| | (2) 70% |

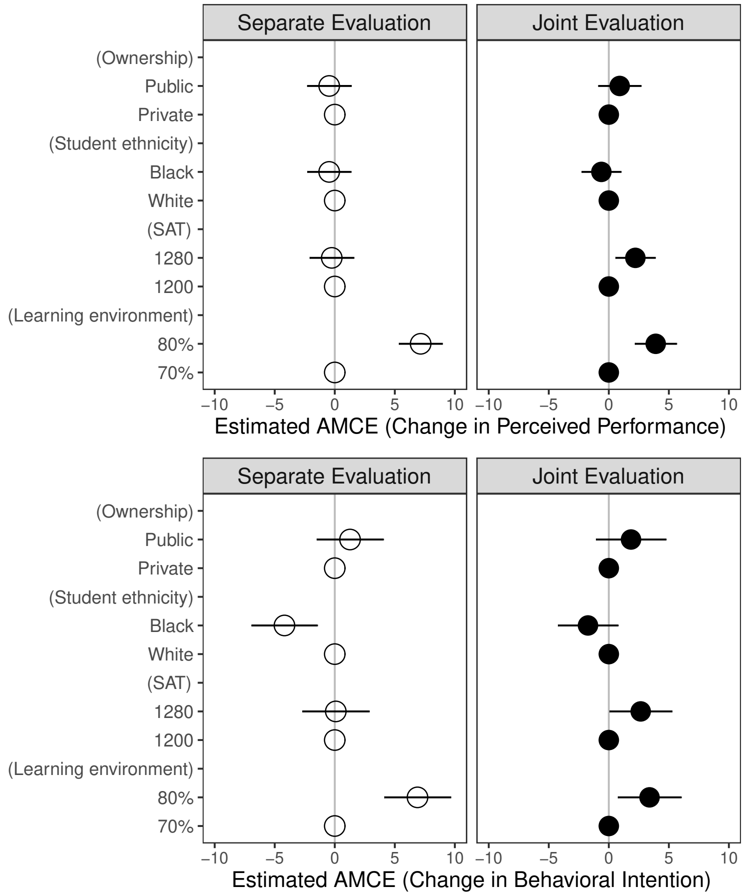


Figure 6: **SE and JE in Conjoint Environments**

*Note*: Bars are 95% confidence intervals.

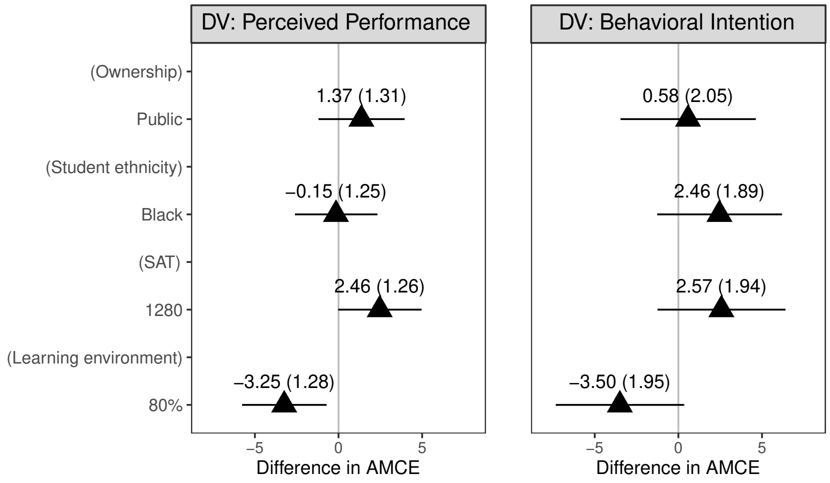


Figure 7: **JE-SE Treatment Effects in Conjoint Environments**

*Note*: Clustered standard errors are in brackets. Bars are 95% confidence intervals.

# Supplemental Information

## Contents

## Appendix A 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: $17 + 63 = ?$

## Appendix B Study 1

### *Appendix B i Pre-registration report*

**Have any data been collected for this study already?**
No, no data have been collected for this study yet.

**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?

**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)

**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.

**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.

**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.

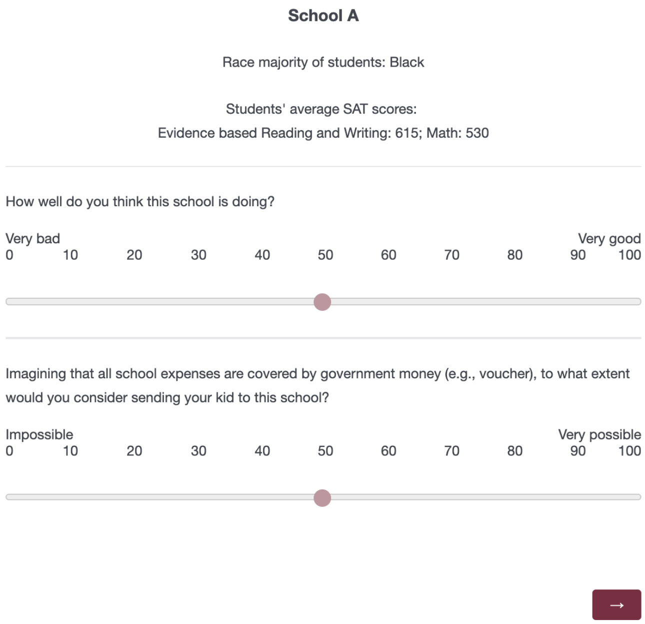
**How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.**
We will stop data collection once 1000 subjects have submitted a responses on MTurk. Deviations from this goal are entirely due to MTurk software and outside of our control.

**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 B ii Experimental intervention*

After a brief introduction to the American high school scenario, we randomly assigned subjects into one of the following three conditions.

**Separate Evaluation Condition: Black School**



**School A**

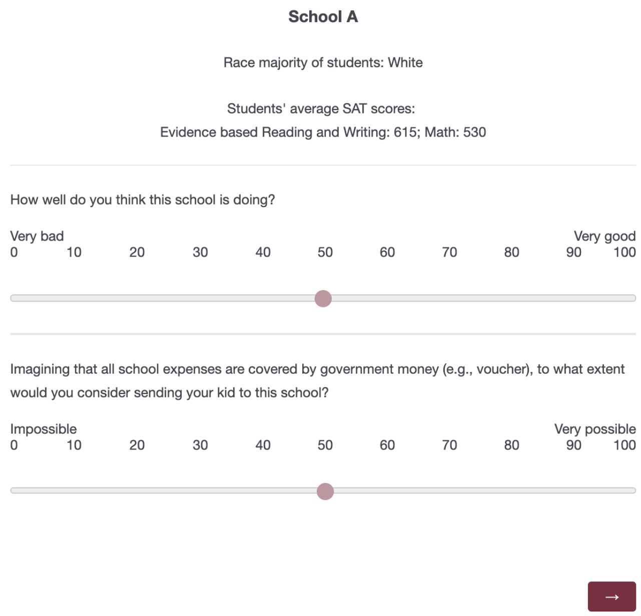
Race majority of students: Black

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?

**Separate Evaluation Condition: White School**



**School A**

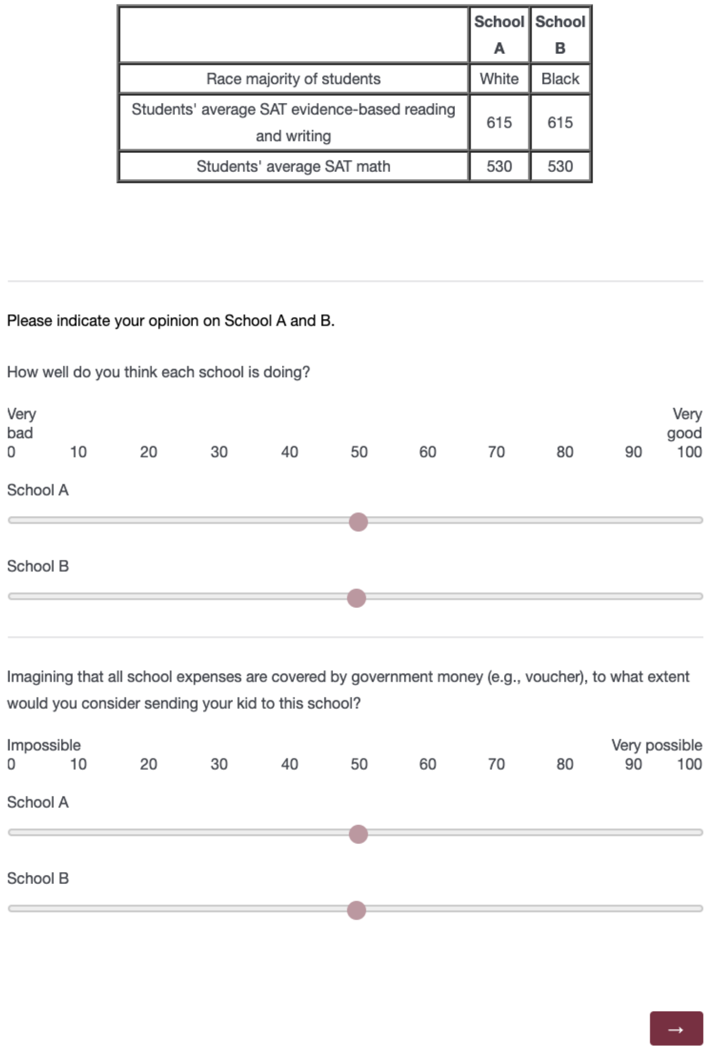
Race majority of students: White

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?

### 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 Very possible 90 100

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 B iii Characteristics of sample*

Table B.1: **Study 1 Sample**

| | Total Sample | | Separate Evaluation | | | | Joint Evaluation | | |
|---|---|---|---|---|---|---|---|---|---|
| | $(N = 988)$ | | Black $(N = 316)$ | | White $(N = 331)$ | | $(N = 341)$ | | |
| | 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 |
| Age: 30-49 | 494 | 50 | 162 | 51 | 168 | 51 | 164 | 48 | 0.68 |
| Age: $\geq$ 50 | 141 | 14 | 45 | 14 | 47 | 14 | 49 | 14 | 1.00 |
| Income: $<$ \$25k | 175 | 18 | 55 | 17 | 57 | 17 | 63 | 19 | 0.90 |
| Income: \$25k-75k | 503 | 51 | 164 | 52 | 168 | 51 | 171 | 50 | 0.92 |
| Income: $\geq$ \$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 |

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

### *Appendix B iv Manipulation check (MC) and attention test (AT)*

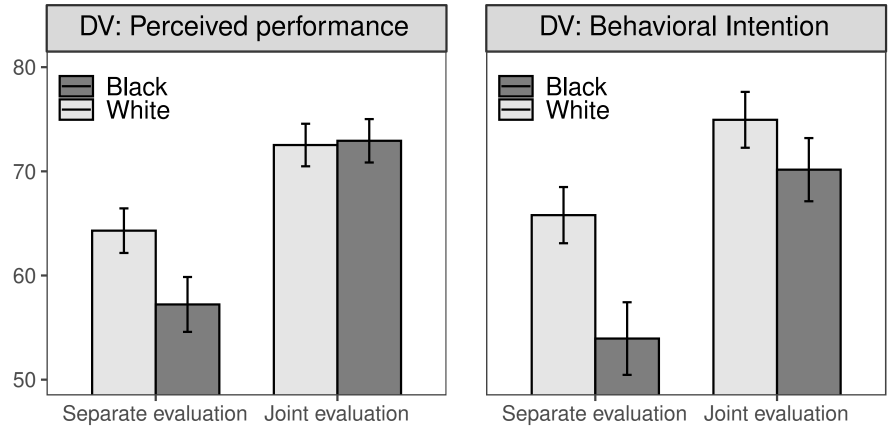


Figure B.1: **Racial Stereotype in SE and JE (MC & AT Pass Sample)**

*Note*: Bars are 95% confidence intervals.

Table B.2: **Racial Stereotype in SE and JE (MC & AT Pass Sample)**

| | Perceived performance | | Behavioral Intention | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| JE×Black | 7.489*** | 8.153*** | 7.064** | 6.995** |
| | (1.807) | (1.916) | (2.581) | (2.684) |
| Mode: JE | 8.222*** | 7.841*** | 9.151*** | 9.116*** |
| | (1.503) | (1.541) | (1.932) | (2.023) |
| Students' major Race: Black | −7.083*** | −7.746*** | −11.849*** | −11.793*** |
| | (1.724) | (1.831) | (2.239) | (2.340) |
| Constant | 64.304*** | 75.079*** | 65.795*** | 81.097*** |
| | (1.087) | (5.804) | (1.370) | (8.498) |
| Covariates | No | Yes | No | Yes |
| State FE | No | Yes | No | Yes |
| Observation | 1,117 | 1,113 | 1,117 | 1,113 |
| Adjusted $R^2$ | 0.102 | 0.131 | 0.082 | 0.104 |

*Note*: OLS estimates. *Mode* (baseline: *SE*) and *Students' major race* (baseline: *White*) are dummies. Clustered standard errors are in brackets. $^*p < .05$; $^{**}p < .01$; $^{***}p < .001$

## Appendix C Study 2

### *Appendix C i Pre-registration report*

**Have any data been collected for this study already?**
No, no data have been collected for this study yet.

**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?

**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)

**How many and which conditions will participants be assigned to?**
Before manipulation, all subjects will read a page of warm up information about how private schools outperform public schools. This information is from real news and reports. After that, we ask subjects: Please indicate that, in general, to what extent do you agree that private schools outperform public schools? (Moving a 0-100 scale bar: 0 = strongly disagree; 100 = strongly agree)

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 public high school or a private high school with the same SAT performance. The only difference between SE groups is the ownership information (public or private). In the JE group, subjects will see both public school and private school 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.

**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.

**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.

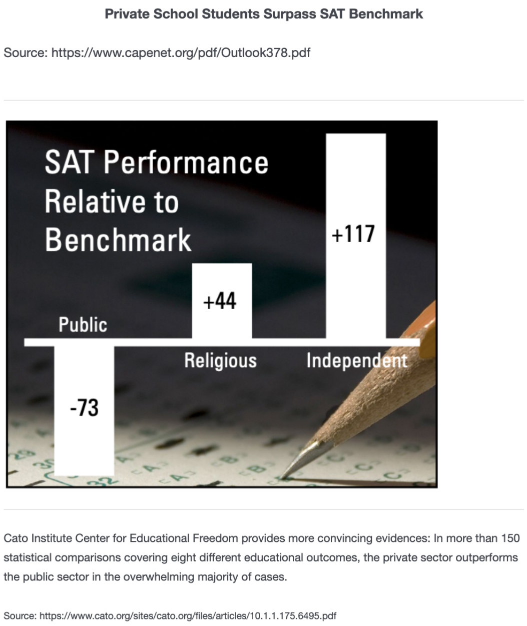
**How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.**
We will stop data collection once 800 subjects have submitted a responses on MTurk. Deviations from this goal are entirely due to MTurk software and outside of our control.

**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 C ii    Experimental intervention*

First, we asked all subjects to read the following information.

**Pro-Private Information Cue**



**Private School Students Surpass SAT Benchmark**

Source: https://www.capenet.org/pdf/Outlook378.pdf

Cato Institute Center for Educational Freedom provides more convincing evidences: In more than 150 statistical comparisons covering eight different educational outcomes, the private sector outperforms the public sector in the overwhelming majority of cases.

Source: https://www.cato.org/sites/cato.org/files/articles/10.1.1.175.6495.pdf

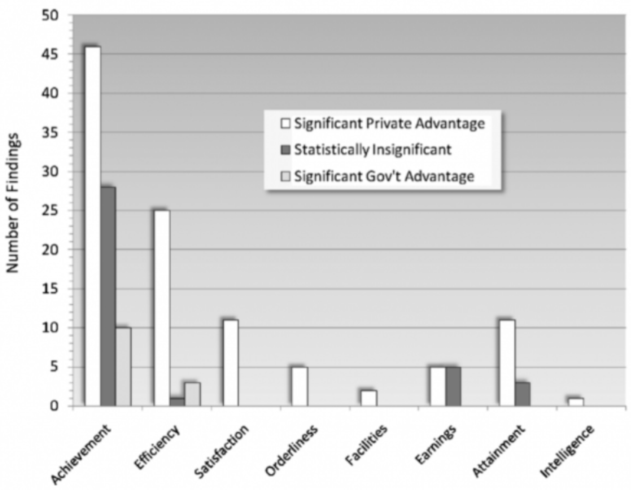
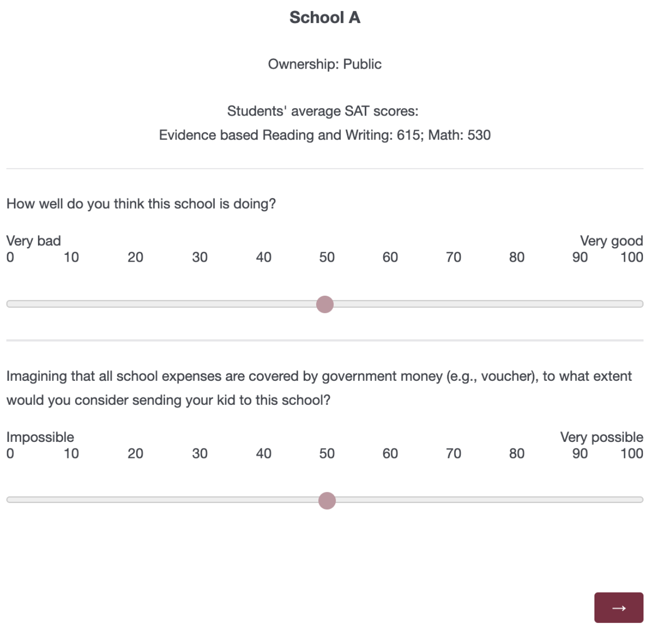


**FIGURE 1** Private school versus government school outcomes, number of significant and insignificant findings, worldwide

[public-private attitude] Please indicate that, in general, to what extent do you agree that private schools outperform public schools? 0 = Strongly disagree = Strongly agree (Please move the slide between 0 and 100)

Next, after a brief introduction to the American high school scenario, we randomly assigned subjects into one of the following three conditions.

### Separate Evaluation Condition: Public School



**School A**

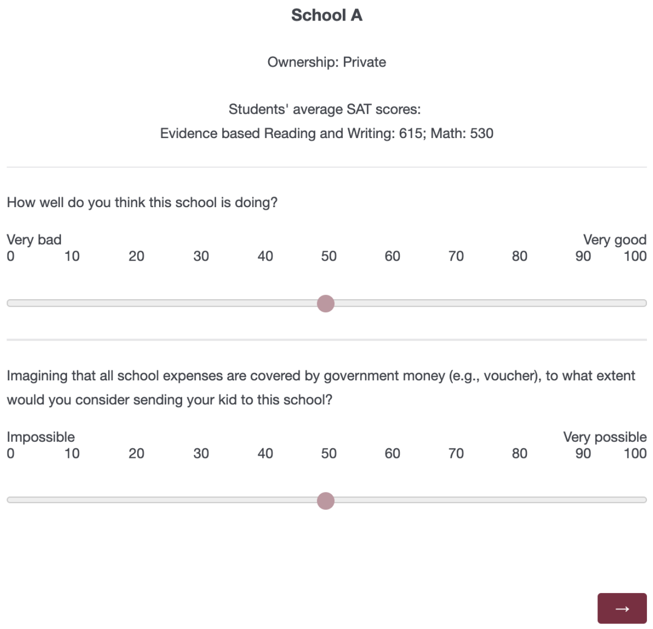
Ownership: Public

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?

### Separate Evaluation Condition: Private School



**School A**

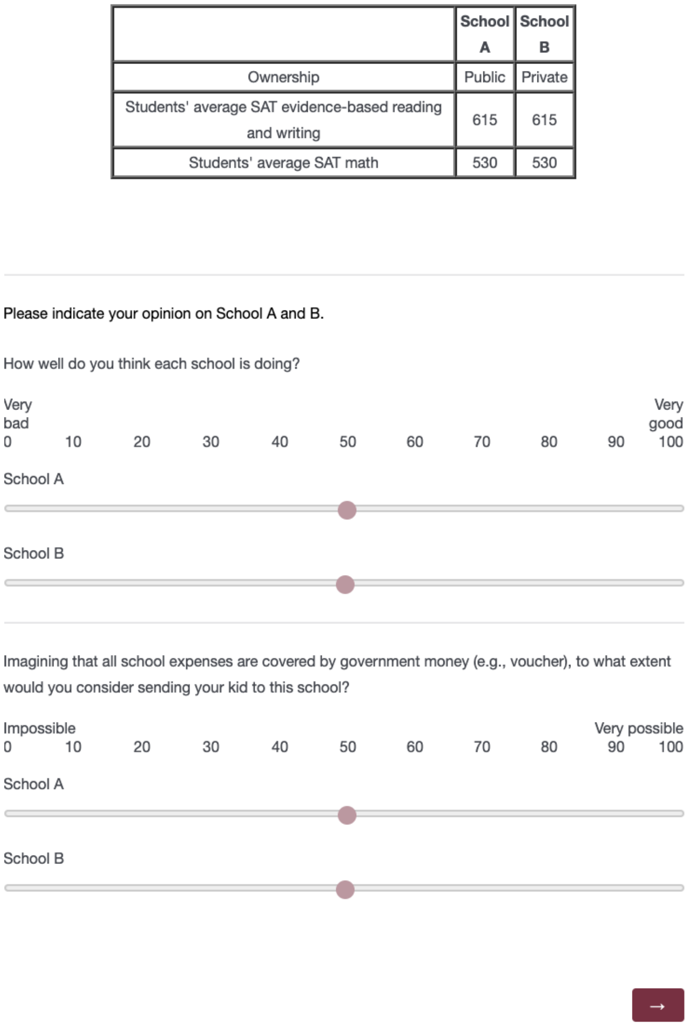
Ownership: Private

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?

## Joint Evaluation Condition



| | School A | School B |
|---|---|---|
| Ownership | Public | Private |
| 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 Very possible 90 100

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 it was a public school.
- I only saw School A, and it was a private school.
- I only saw School A, and it was a nonprofit charter school.
- I saw two schools. School A was public, and School B was private.

## *Appendix C iii   Characteristics of sample*

Table C.1: **Study 2**

| | Total Sample | | Separate Evaluation | | | | Joint Evaluation | | |
|---|---|---|---|---|---|---|---|---|---|
| | ($N = 804$) | | Private ($N = 262$) | | Public ($N = 276$) | | ($N = 266$) | | |
| | Frequency | % | Frequency | % | Frequency | % | Frequency | % | P-value |
| Female | 414 | 51 | 135 | 52 | 132 | 48 | 147 | 55 | 0.22 |
| Male | 390 | 49 | 127 | 48 | 144 | 52 | 119 | 45 | 0.22 |
| White | 575 | 72 | 188 | 72 | 202 | 73 | 185 | 70 | 0.64 |
| Black | 72 | 9 | 22 | 8 | 23 | 8 | 27 | 10 | 0.71 |
| Hispanic | 64 | 8 | 19 | 7 | 19 | 7 | 26 | 10 | 0.41 |
| Asian | 72 | 9 | 28 | 11 | 23 | 8 | 21 | 8 | 0.48 |
| Other | 21 | 3 | 5 | 2 | 9 | 3 | 7 | 3 | 0.62 |
| Age: 18-29 | 292 | 37 | 88 | 34 | 89 | 33 | 115 | 43 | 0.02 |
| Age: 30-49 | 389 | 49 | 139 | 53 | 143 | 52 | 107 | 40 | 0.00 |
| Age: $\geq$ 50 | 118 | 15 | 33 | 13 | 41 | 15 | 44 | 17 | 0.46 |
| Income: $<$ \$25k | 105 | 13 | 30 | 11 | 44 | 16 | 31 | 12 | 0.21 |
| Income: \$25k to \$75k | 407 | 51 | 146 | 56 | 124 | 45 | 137 | 52 | 0.05 |
| Income: $\geq$ \$75k | 291 | 36 | 86 | 33 | 107 | 39 | 98 | 37 | 0.33 |
| College degree | 502 | 63 | 172 | 66 | 169 | 61 | 161 | 61 | 0.43 |
| Conservative | 184 | 23 | 58 | 22 | 62 | 22 | 64 | 24 | 0.85 |
| Liberal | 375 | 47 | 130 | 50 | 121 | 44 | 124 | 47 | 0.41 |
| Moderate | 245 | 30 | 74 | 28 | 93 | 34 | 78 | 29 | 0.35 |
| Parenthood | 376 | 47 | 132 | 50 | 134 | 49 | 110 | 41 | 0.09 |

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

### *Appendix C iv   Manipulation check (MC) and attention test (AT)*

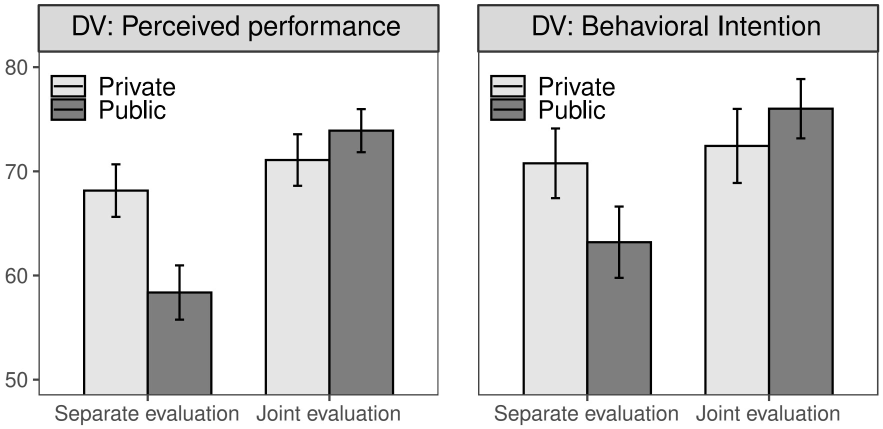


Figure C.1: **Sector Stereotype in SE and JE (MC & AT Pass Sample)**

*Note*: Bars are 95% confidence intervals.

Table C.2: **Sector Stereotype in SE and JE (MC & AT Pass Sample)**

| | Perceived Performance | | Behavioral Intention | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| JE×Public | 12.603*** | 11.447*** | 11.149*** | 10.799** |
| | (2.089) | (2.192) | (3.209) | (3.408) |
| Mode: JE | 2.936 | 3.577* | 1.668 | 2.285 |
| | (1.793) | (1.815) | (2.476) | (2.575) |
| Sector: Public | −9.783*** | −8.627*** | −7.574** | −7.224** |
| | (1.838) | (1.954) | (2.426) | (2.576) |
| Constant | 68.150*** | 68.812*** | 70.770*** | 74.153*** |
| | (1.279) | (8.248) | (1.695) | (10.379) |
| Covariates | No | Yes | No | Yes |
| State FE | No | Yes | No | Yes |
| Observation | 852 | 848 | 852 | 848 |
| Adjusted R$^2$ | 0.093 | 0.124 | 0.032 | 0.018 |

*Note*: OLS estimates. *Mode* (baseline: *SE*) and *Sector* (baseline: *Private*) are dummies. Clustered standard errors are in brackets. $^*p < .05$; $^{**}p < .01$; $^{***}p < .001$

# Appendix D Study 3

## *Appendix D i Pre-registration report*

**Have any data been collected for this study already?**
No, no data have been collected for this study yet.

**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?

**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)

**How many and which conditions will participants be assigned to?**
We will randomize participants in two groups: one single conjoint task group and one pair conjoint task group. All subjects will see a conjoint task to evaluate a school with four attribute information: school's ownership, students' race majority, students' average SAT score, and students' perceived learning environment.

All attributes, values, and their orders will be randomly appeared to subjects. In single conjoint group, subjects will only see one school profile, but in pair conjoint group they will see two school profiles. To balance the sample size between single and pair conjoint groups, the probability of a subject to see a single conjoint task is 2/3, and the probability of a subject to see a pair conjoint task is 1/3.

**Specify exactly which analyses you will conduct to examine the main question/hypothesis.**
Analyses will be based on the standard practices in the conjoint experimental design:

i Average Marginal Component Effect (AMCE).

ii Marginal Means (MM).

The treatment effect (difference between single and pair conjoint) will be analyzed by the difference-in-AMCE between two groups. The unit of analysis is profile, so we will cluster standard errors at individuals.

**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.

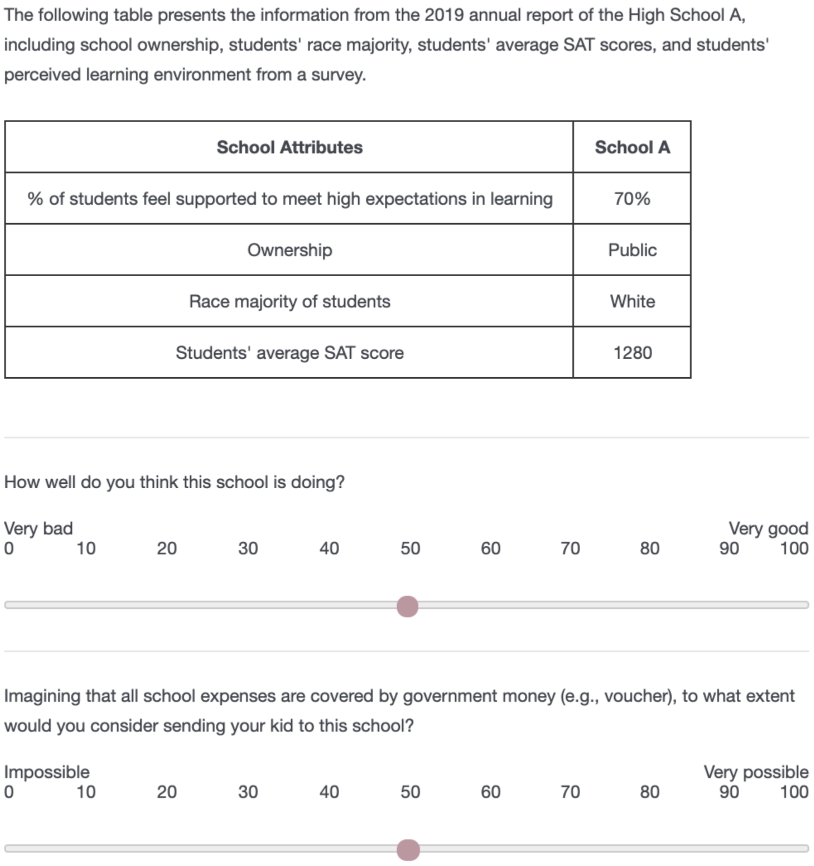
**How many observations will be collected or what will determine the sample size? No need to justify decision, but be precise about exactly how the number will be determined.**
We will stop data collection once 1500 subjects have submitted a responses on MTurk. Deviations from this goal are entirely due to MTurk software and outside of our control.

**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 D ii Experimental intervention*

The following figures are examples of single and pair conjoint. Every subject randomly saw different attribute values.

**Single Conjoint Condition**



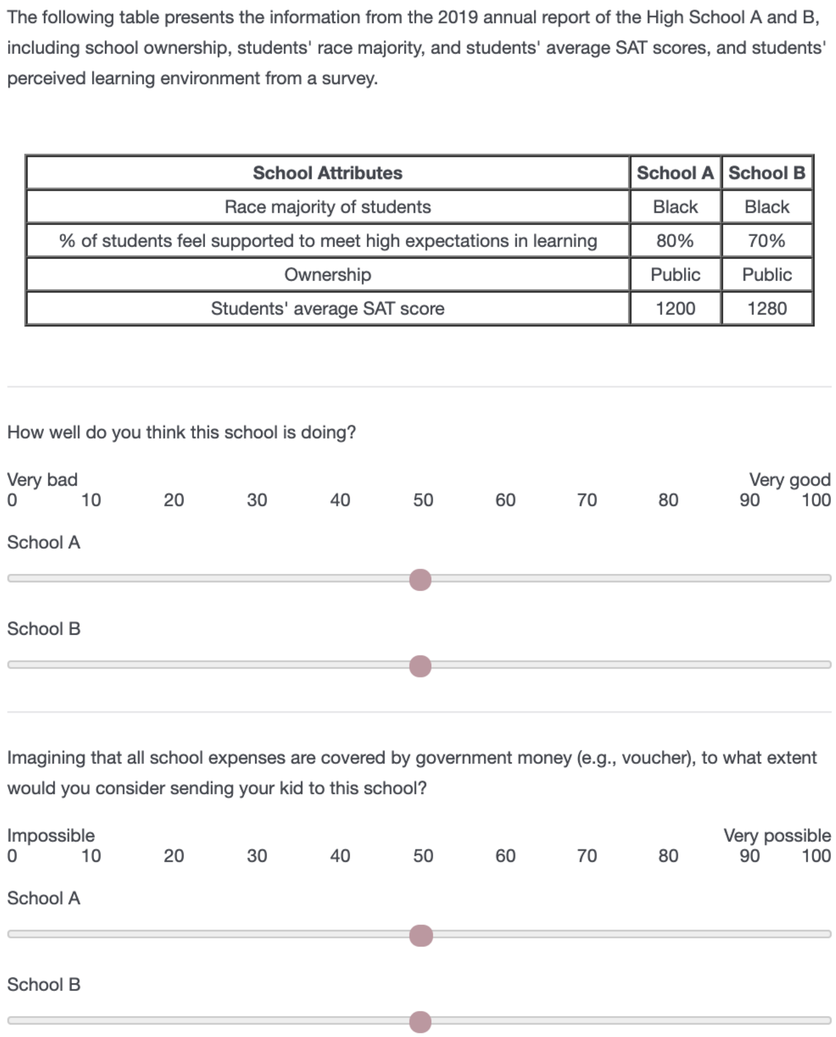
The following table presents the information from the 2019 annual report of the High School A, including school ownership, students' race majority, students' average SAT scores, and students' perceived learning environment from a survey.

| School Attributes | School A |
|---|---|
| % of students feel supported to meet high expectations in learning | 70% |
| Ownership | Public |
| Race majority of students | White |
| Students' average SAT score | 1280 |

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?

### Pair Conjoint Condition



The following table presents the information from the 2019 annual report of the High School A and B, including school ownership, students' race majority, and students' average SAT scores, and students' perceived learning environment from a survey.

| School Attributes | School A | School B |
|---|---|---|
| Race majority of students | Black | Black |
| % of students feel supported to meet high expectations in learning | 80% | 70% |
| Ownership | Public | Public |
| Students' average SAT score | 1200 | 1280 |

How well do you think this school is doing?

Very bad — Very good

0 10 20 30 40 50 60 70 80 90 100

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 — Very possible

0 10 20 30 40 50 60 70 80 90 100

School A

School B

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

- I only saw one school, School A.
- I only saw one school, School B.
- I saw two schools. School A and School B.

### *Appendix D iii    Characteristics of sample*

Table D.1: **Study 3 Sample**

| | Total Sample | | Separate Evaluation | | Joint Evaluation | | |
|---|---|---|---|---|---|---|---|
| | ($N = 1502$) | | ($N = 961$) | | ($N = 541$) | | |
| | Frequency | % | Frequency | % | Frequency | % | P-value |
| Female | 782 | 52 | 495 | 52 | 287 | 53 | 0.57 |
| Male | 720 | 48 | 466 | 48 | 254 | 47 | 0.57 |
| White | 1012 | 67 | 656 | 68 | 356 | 66 | 0.33 |
| Black | 155 | 10 | 103 | 11 | 52 | 10 | 0.49 |
| Hispanic | 132 | 9 | 84 | 9 | 48 | 9 | 0.93 |
| Asian | 161 | 11 | 96 | 10 | 65 | 12 | 0.23 |
| Other | 42 | 3 | 22 | 2 | 20 | 4 | 0.14 |
| Age: 18-29 | 605 | 40 | 391 | 41 | 214 | 40 | 0.73 |
| Age: 30-49 | 698 | 47 | 445 | 46 | 253 | 47 | 0.79 |
| Age: $\geq$ 50 | 196 | 13 | 125 | 13 | 71 | 13 | 0.92 |
| Income: $<$ \$25k | 239 | 16 | 163 | 17 | 76 | 14 | 0.13 |
| Income: \$25k to \$75k | 750 | 50 | 483 | 50 | 267 | 49 | 0.71 |
| Income: $\geq$ \$75k | 511 | 34 | 313 | 33 | 198 | 37 | 0.12 |
| College degree | 846 | 56 | 534 | 56 | 312 | 58 | 0.41 |
| Conservative | 311 | 21 | 199 | 21 | 112 | 21 | 1.00 |
| Liberal | 712 | 47 | 474 | 49 | 238 | 44 | 0.05 |
| Moderate | 477 | 32 | 287 | 30 | 190 | 35 | 0.04 |
| Parenthood | 633 | 42 | 429 | 45 | 204 | 38 | 0.01 |

*Note*: P-values are generated from two sample $t$-tests.

### *Appendix D iv Benchmark models*

Table D.2: **Conjoint Conditions (DV: Perceived Performance)**

| | Single Conjoint (SE) | | Pair Conjoint (JE) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Ownership: Public | −0.463 | −0.782 | 0.906 | 1.191 |
| | (0.939) | (0.979) | (0.911) | (0.877) |
| Students' major race: Black | −0.467 | −0.983 | −0.619 | −0.909 |
| | (0.931) | (0.957) | (0.838) | (0.834) |
| Higher SAT | −0.250 | −0.249 | 2.213$^{**}$ | 2.469$^{**}$ |
| | (0.941) | (0.978) | (0.847) | (0.824) |
| Better learning environment | 7.153$^{***}$ | 7.051$^{***}$ | 3.907$^{***}$ | 4.271$^{***}$ |
| | (0.925) | (0.941) | (0.888) | (0.842) |
| Constant | 72.843$^{***}$ | 68.038$^{***}$ | 74.051$^{***}$ | 77.171$^{***}$ |
| | (1.164) | (4.552) | (1.029) | (7.610) |
| Covariates | No | Yes | No | Yes |
| State FE | No | Yes | No | Yes |
| Observation | 960 | 957 | 1,082 | 1,072 |
| Adjusted $R^2$ | 0.055 | 0.078 | 0.024 | 0.083 |

*Note*: OLS estimates. *Ownership* (baseline: *Private*), *Students' major race* (baseline: *White*), *Higher SAT* (baseline: *1200*), and *Better learning environment* (baseline: *70%*) are dummies. Standard errors are in brackets (clustered by individuals). $^{*}$p $<$ .05; $^{**}$p $<$ .01; $^{***}$p $<$ .001

Table D.3: **Conjoint Conditions (DV: Behavioral Intention)**

| | Single Conjoint (SE) | | Pair Conjoint (JE) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Ownership: Public | 1.271 | 1.609 | 1.849 | 2.545 |
| | (1.420) | (1.501) | (1.491) | (1.479) |
| Students' major race: Black | −4.189$^{**}$ | −4.586$^{**}$ | −1.734 | −1.994 |
| | (1.402) | (1.467) | (1.281) | (1.309) |
| Higher SAT | 0.085 | 0.299 | 2.656$^{*}$ | 2.635$^{*}$ |
| | (1.422) | (1.494) | (1.331) | (1.323) |
| Better learning environment | 6.891$^{***}$ | 7.168$^{***}$ | 3.391$^{*}$ | 4.066$^{**}$ |
| | (1.416) | (1.428) | (1.348) | (1.336) |
| Constant | 72.198$^{***}$ | 87.829$^{***}$ | 73.157$^{***}$ | 78.447$^{***}$ |
| | (1.639) | (5.776) | (1.581) | (11.073) |
| Covariates | No | Yes | No | Yes |
| State FE | No | Yes | No | Yes |
| Observation | 960 | 957 | 1,082 | 1,072 |
| Adjusted $R^2$ | 0.032 | 0.056 | 0.009 | 0.062 |

*Note*: OLS estimates. *Ownership* (baseline: *Private*), *Students' major race* (baseline: *White*), *Higher SAT* (baseline: *1200*), and *Better learning environment* (baseline: *70%*) are dummies. Standard errors are in brackets (clustered by individuals). $^{*}$p $<$ .05; $^{**}$p $<$ .01; $^{***}$p $<$ .001

Table D.4: **JE-SE Treatment Effects in Conjoint Conditions**

| | DV: Perceived Performance | | DV: Behavioral Intention | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Public×JE | 1.369 | 1.365 | 0.577 | 0.436 |
| | (1.308) | (1.320) | (2.059) | (2.085) |
| Black×JE | −0.153 | −0.285 | 2.456 | 2.293 |
| | (1.253) | (1.265) | (1.899) | (1.952) |
| Higher SAT×JE | 2.463 | $2.718^{*}$ | 2.571 | 2.708 |
| | (1.266) | (1.287) | (1.947) | (1.954) |
| Better learning environment×JE | $-3.247^{*}$ | $-2.949^{*}$ | −3.499 | −3.373 |
| | (1.282) | (1.270) | (1.955) | (1.956) |
| Mode: JE | 1.208 | 1.242 | 0.959 | 1.159 |
| | (1.553) | (1.579) | (2.277) | (2.359) |
| Ownership: Public | −0.463 | −0.659 | 1.271 | 1.716 |
| | (0.939) | (0.962) | (1.420) | (1.461) |
| Students' major race: Black | −0.467 | −0.656 | $-4.189^{**}$ | $-4.376^{**}$ |
| | (0.931) | (0.955) | (1.402) | (1.455) |
| Higher SAT | −0.250 | −0.275 | 0.085 | 0.189 |
| | (0.941) | (0.966) | (1.422) | (1.462) |
| Better learning environment | $7.153^{***}$ | $7.090^{***}$ | $6.891^{***}$ | $7.157^{***}$ |
| | (0.925) | (0.936) | (1.415) | (1.429) |
| Constant | $72.843^{***}$ | $71.899^{***}$ | $72.198^{***}$ | $81.811^{***}$ |
| | (1.164) | (4.207) | (1.639) | (6.044) |
| Covariates | No | Yes | No | Yes |
| State FE | No | Yes | No | Yes |
| Observation | 2,042 | 2,029 | 2,042 | 2,029 |
| Adjusted $R^2$ | 0.042 | 0.062 | 0.023 | 0.045 |

*Note*: OLS estimates. JE-SE treatment effects are coefficients of interaction terms. *Ownership* (baseline: *Private*), *Students' major race* (baseline: *White*), *Higher SAT* (baseline: *1200*), *Better learning environment* (baseline: *70%*), and *Mode* (baseline: *SE*) are dummies. Standard errors are in brackets (clustered by individuals). $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

## *Appendix D v   Manipulation check (MC) and attention test (AT)*

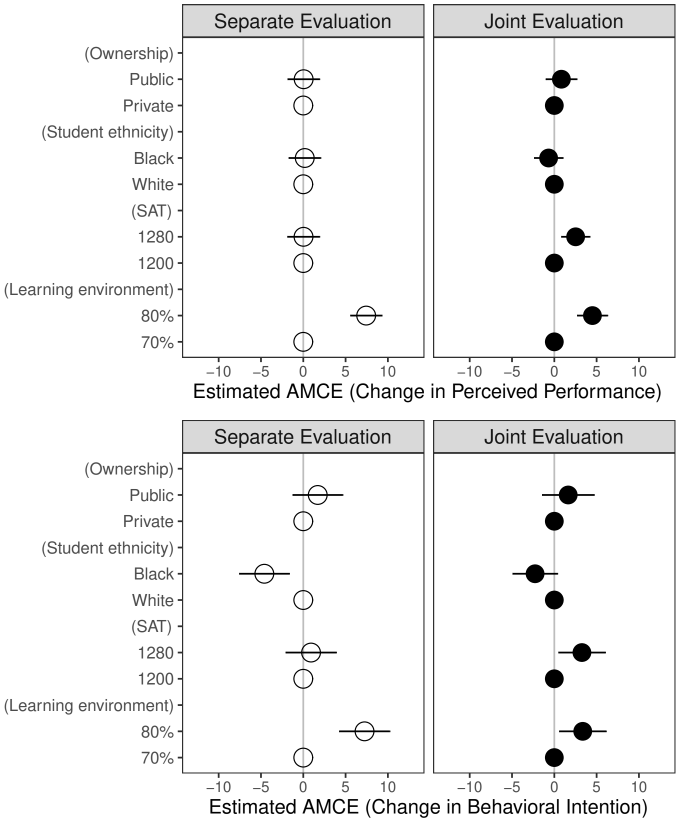


Figure D.1: **SE and JE in Conjoint Environments (MC & AT Pass Sample)**

*Note*: Bars are 95% confidence intervals.

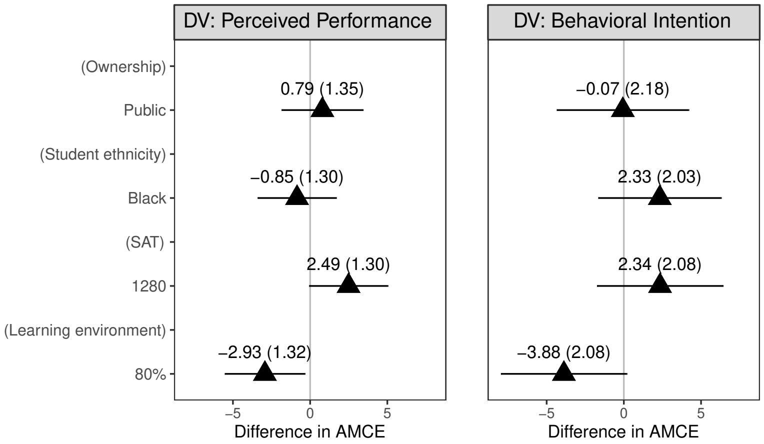


Figure D.2: **JE-SE Treatment Effects in Conjoint Environments (MC & AT Pass Sample)**

*Note*: Clustered standard errors are in brackets. Bars are 95% confidence intervals.