Received: 22 May 2022 | Revised: 18 June 2023 | Accepted: 20 June 2023

DOI: 10.1111/puar.13690

RESEARCH ARTICLE

PAR PUBLIC ADMINISTRATION REVIEW ASPA

Effectiveness, efficiency, and equity tradeoffs in public programs: A citizen experiment

Kenneth J. Meier[1,2,3] | **Jourdan Davis**[4] | **Xiaoyang Xu**[5]

[1]Department of Public Administration and Policy, School of Public Affairs, American University, Washington, DC, USA

[2]Cardiff School of Business, Cardiff University, Cardiff, UK

[3]Institute of Public Administration, Leiden University, The Hague, the Netherlands

[4]Department of Political Science and Public Administration, University of North Carolina-Charlotte, Charlotte, North Carolina, USA

[5]McCourt School of Public Policy, Georgetown University, Washington, DC, USA

Correspondence
Kenneth J. Meier, Department of Public Administration and Policy, School of Public Affairs, American University, Washington, DC 20016, USA.
Email: kmeier@american.edu

Abstract

Debates over public programs frequently focus on questions of effectiveness, equity, and efficiency and the tradeoff among these objectives. Missing from the literature is whether the general public cares about these tradeoffs, can perceive such differences, and will act on them. This article reports on two pre-registered vignette experiments where the effectiveness, equity, and efficiency are assessed relative to experimental treatments focused on U.S. K-12 education involving test scores, equality of test scores, and program costs. One experiment focuses on equity in race and the other on equity in income. The experiments show that the general public perceives differences in program effectiveness and equity, values both, and is unwilling to tradeoff one for the other. The public cares about program costs, but it lacks a sophisticated understanding of efficiency as a concept. Inequalities in income appear to influence equity concerns more than those involving race.

Practitioner points

- The general public can distinguish between effectiveness, equity, and efficiency in evaluating programs if given information on these dimensions.
- Effectiveness, equity, and efficiency are all comparative terms and some criteria for comparison needs to be available.
- Public judgments on effectiveness, equity, and efficiency are generally intuitive and direct and not subject to complex calculations or explicit tradeoffs.
- How information is provided to the public is likely to affect the public's ability to make informed judgments about program performance.
- Performance information should be designed to provide information on equity and efficiency as well as on effectiveness.

INTRODUCTION

Public programs tend to have multiple goals (Chun & Rainey, 2005), multiple stakeholders with different goals (Boyne, 2002, 2003), or heterogeneous impacts across different population groups (Baekgaard & Serritzlew, 2016); all three of these factors suggest that public programs involve balancing of values that might at times be in conflict (Frederickson, 2015; Hall, 2022; Okun, 1975). Arthur Okun (1975) argued that the central question of public policy was how much should government intervene in the market to provide for greater equity in contrast to market allocations of goods and services which he assumed to be efficient. Maximizing efficiency, he argued, will benefit those with competitive advantages and, thus, have consequences for those lacking education, resources, or luck. Efficiency, he felt, resulted in inequality. Within public administration, George Frederickson (2015) has advocated that equity be given a greater value and be at least equal to the classic pillars of economy and efficiency. Similar policy debates followed the publication of Thomas Piketty's *Capital in the Twenty-First Century* (Piketty, 2015) and a greater concern for the distributional consequences of public policy.

This research examines how one important stakeholder in public programs, the general public, assesses different values when evaluating public program outcomes based on effectiveness, efficiency, and equity. Although political debates frequently focus on these evaluative dimensions and some experimental evidence shows that civil servants consider these concepts (Fernández-Gutiérrez & Van de Walle, 2019), we were unable to find any studies of how individual citizens make specific assessments of effectiveness, equity, and efficiency when such factors vary (rather than as generic responses to the same set of circumstances, see Andrews & Van de Walle, 2013; Hvidman & Andersen, 2016; or where there is some information on implied inequalities, see Amirkhanyan et al., 2023; Walker et al., 2018). This article uses two preregistered internet vignette experiments focused on public K-12 education in the United States that manipulate performance indicators of effectiveness, equity, and cost to determine if citizens can distinguish among these concepts and whether or not they are likely to make any tradeoffs among the three values in their evaluations of government programs. The results indicate that the public can distinguish between effectiveness, equity, and efficiency in an intuitive manner although not in more sophisticated ways. A second experiment shows that the public is more sensitive to inequity based on socio-economic status than in terms of race. Both experiments show that individuals are more likely to act or intend to act based on effectiveness and equity but not on efficiency. They also show that there are group-specific heterogeneous responses to equity, effectiveness, and efficiency but evidence of motivated reasoning is less apparent.

The current study seeks to make contributions to both public administration research and practice. Although much scholarly literature emphasizes tradeoffs among effectiveness, efficiency, and equity; and policy makers clearly consider these factors, whether the public recognizes and responds to differences in effectiveness, equity, and efficiency as separate dimensions of evaluation remains an open question (see Brunner et al., 2022). It also probes whether the public tradesoff these values with each other or views them independently and whether these values influence their comfort with a public program and, hence, the willingness to participate. For the world of practice, the study illustrates the range of performance information that the public finds relevant and the need to stress equity as well as effectiveness (see Ruijer et al., 2023). Public administrators also frequently make decisions that tradeoff equity, efficiency, and effectiveness, and knowing public preferences permit more responsive public policy.

THE THEORETICAL AND EMPIRICAL TRADEOFF: EFFECTIVENESS, EQUITY, AND EFFICIENCY

A central tenet of democratic governance is that public policy and administration should be responsive to the general public (Redford, 1969). As a result, an extensive scholarly debate on conflict among of the values of effectiveness, equity, and efficiency (Frederickson, 2015; Okun, 1975; Piketty, 2015) is reflected in virtually all program evaluation and policy analysis texts (Dunn 2015; Jenkins-Smith 1990; Weimer and Vining 2017). Similar discussions exist in political science (Swank, 1998), economics (Gershberg & Schuermann, 2001; Okun, 1975) sociology (Daw, 2015), law (Viscusi & Zeckhauser, 2005), and other fields.

Perhaps representing the economics' ties to evaluation research, efficiency is universally accepted as an evaluative criterion for public programs (Andrews & Entwistle, 2013; Brunner et al., 2022; de Graaf & Paanakker, 2015; Hood, 1991; Wang, 2022), but clearly not as the sole value (Fernández-Gutiérrez & Van de Walle, 2019; Frederickson, 2015) and often a contested value (Le Grand, 1990). In fact, Le Grand (1990) argues that efficiency cannot stand on its own but must be considered as a secondary concern once the objectives of effectiveness and equity are attained.

The logic for different dimensions of performance starts with effectiveness; if a program does not achieve the goals established for it, it makes little sense to consider whether the program is efficient or if the distribution of program benefits is equitable. Given an effective program, theoretically individuals can vary in how much they value the program and thus how much they would be willing to invest in program outcomes, that is, in the relative efficiency of the program. Similarly, given an effective program, then theoretically the distribution of benefits across individuals and concerns about equity are likely to arise. Just as I.M.D. Little (2002) notes that perceptions of wealth are logically assessed in comparison to one's neighbors, the benefits of government programs fall unequally across individuals, and people are likely to respond in terms of how equitable they perceive the program is. At both theoretical and practical levels, these distinctions among effectiveness, equity, and efficiency require individuals to decide how to weigh each of these criteria depending on how much they favor equity, efficiency, and effectiveness relative to each other.

Despite the extensive literature debating these various dimensions of program performance and recent work that asks public administrators in general terms whether they consider efficiency or equity in making decisions (Fernández-Gutiérrez & Van de Walle, 2019), existing research until recently did not address how one important stakeholder, the public, values program outcomes in terms of effectiveness, equity, or efficiency. Belle and Cantarelli (2022) recently examine the willingness of the public to tradeoff economic benefits, individual restrictions, and lives lost using a conjoint experiment on COVID-19 restrictions in Italy. Although they are not dealing with the equity, efficiency, and effectiveness tradeoff directly, they demonstrate how a behavioral public administration approach can effectively manipulate various dimensions of public performance and determine how individuals

respond to these values and/or the tradeoff in these values.

Prior literature on behavioral public administration in this area focuses on overall performance (James et al., 2020) and is generally concerned with issues of framing performance information (Belardinelli et al., 2018; James & Van Ryzin, 2019; Olsen, 2015; Pedersen, 2017), questions of blame avoidance (Marvel & Girth, 2016; Piatak, Mohn and Leland 2017; Johnson et al., 2019), trust in the data (Schmidthuber et al., 2023), or how public officials use such information (James et al., 2020 and the citations therein).

Perhaps the closest approach is by Hvidman and Andersen (2016) in a study of sector bias in performance who recognized four dimensions including efficiency and effectiveness as outcome dimensions and developed scales to measure them in a vignette experiment in Denmark; similar work has been undertaken in the United States (Marvel, 2016; Meier et al., 2019; Meier et al., 2022). Those experiments, however, only measure individuals' responses on the various value dimensions in response to a single overall manipulation of performance. They do not independently manipulate separate treatments linked to effectiveness, efficiency, and equity. A recent eight-country study of COVID-19 government restrictions, similarly manipulated performance, whether governments restricted individual liberties, and the impact on lower income populations, but did not consider efficiency (Amirkhanyan et al., 2023; see Walker et al., 2018 on whether or not information on equity efforts rather than outcomes are provided). Favero and Rutherford (2020) use observational data from Korea to demonstrate that both parents and students put value on equity, but they do not assess responses to either effectiveness or efficiency. In a vignette experiment on U.S. education, Valant and Newark (2016) find a concern with equity; however, their experimental manipulation deals with equity in terms of income versus race rather than how they evaluate the different dimensions of performance.

Multiple program values implies that different individuals will view programs in terms of these values and alter their assessments based on how much they weight these value dimensions in their own mind. Belle and Cantarelli (2022) illustrate the value of an experimental approach to trading-off values with a vignette experiment where Italian respondents are asked their preferred government policy to address COVID-19 based on its impact on life (number of deaths), income, and freedom (length of lockdowns). Their conjoint experiment shows that respondents are more sensitive to changes in income than either deaths or length of lockdowns.

Belle and Cantarelli (2022) demonstrate that individuals are willing to make tradeoffs among program outcomes based on different values, including outcomes that involve taboo values (that is, valuing human lives). Their tradeoffs are based on the relative burdens policies place on citizens (deaths, loss of income, length of lockdowns) rather than the generic values of effectiveness, equity, and efficiency. The current research seeks to take the logic of tradeoffs to a more general level and directly assess program outcomes in terms of the three values, and then determine if support for those values changes as the other value dimensions change (that is, is a program judged more effective when equity increases or costs decrease?).

OPERATIONALIZING EFFECTIVENESS, EQUITY AND EFFICIENCY

Recognizing that program performance is multidimensional and that any assessment of performance is contested, and thus subjective, has both theoretical and measurement implications. Understanding how effectiveness, equity, and efficiency can be evaluated and how tradeoffs are made requires clear definitions of the concepts and how they relate to each other. We start by reducing our three concepts of concern – effectiveness, equity, and efficiency into their simplest elements to provide distinct definitions. The objective is to create a concept that reflects the ordinary language usage of the term and at the same time ensures that it is conceptually distinct from the other two concepts. Each is also a comparative concept; that is, there is no absolute level of effectiveness or efficiency, but those terms have meaning in comparison to a standard or to other organizations or programs on the same measure.

We further simplify the concerns to focus on program outcomes rather than process or access as alternative definitions. To illustrate, concerns with equity might focus on a wide variety of aspects (Frederickson, 2015). A person might value equality of access, that is a lack of barriers to seeking some end, but be willing to accept unequal outcomes that might still result that reflect differences in need, effort, or talent. Similarly, another person might value equity of process by which all individuals are treated equally but with outcomes that also reflect other differences among the individuals. Outcomes have the advantage of being comparable, whether concerned with effectiveness, equity, or efficiency, and can potentially be also compared across these three dimensions.

Following the logic of Fernández-Gutiérrez and Van de Walle (2019) concerning outcomes, we assume that there is an outcome measure X with some degree of reliability and validity. We will define effectiveness in relative terms and conclude that if program A produces more of X than program B, then program A is more effective on this specific indicator of performance. Equity also requires a comparison, but that comparison is between groups of clients. Program A can be considered more equitable than program B, if the gap between Group 1 and Group 2 in program A is less than the gap between the two groups in program B. Finally, efficiency involves a comparison of

costs, given two programs A and B that are equally effective, then A is more efficient than B, if the cost of program A is lower than the cost of program B.

Despite definitions that seek to separate these concepts, the three concepts are also likely to overlap in individuals' minds and at the theoretical level. Effectiveness is a broad concept and likely contains within it some assessment of both efficiency and equity. An ineffective program clearly cannot be an efficient program; programs that have unequal impacts on individuals might also be viewed as less effective (Amirkhanyan et al., 2023). The issues of overlap and ability to separate the concepts create methodological issues of measurement that need to be addressed (see below).

HYPOTHESES AND CONTEXT FOR THE EXPERIMENT

To assess the sensitivity of individuals to the performance dimensions of effectiveness, equity, and efficiency in a causal format, a research design needs to manipulate indicators of each concept and then determine if those treatments affect assessments of the latent variables (measures of effectiveness, equity, and efficiency). Tradeoffs between the values could then be assessed by determining if the treatment effects for any one concept (say effectiveness) are influenced by the existing levels of the other concepts (such as equity). The analysis would be facilitated with the inclusion of some end behavior or value such as willingness to use or comfort with using the program.

We opt for a between subjects $3 \times 3 \times 3$ factorial design rather than a conjoint experiment. The advantage of a conjoint experiment is that it forces the tradeoff among values to be reduced to a single preference scale for comparison purposes, and changes along this scale indicate willingness to tradeoff values. The alternative approach is to directly assess program outcomes on the individual dimensions of concern (here effectiveness, equity, and efficiency) and then examine how changes in various program dimensions affect these assessments. This alternative allows one to determine how much a respondent values each of the dimensions separately, and by including a measure of behavioral intention or willingness to use the service, the experiment also contains an overall evaluation on a single dimension.

In the preregistered vignette experiments, all respondents were given the same information in regard to the performance, equity, and cost of the average school district in the state. Three treatments were randomly manipulated for the hypothetical district, the overall test score (designated as the effectiveness treatment), the low income or African American test score (the equity treatment), and the cost per student (the efficiency treatment). These treatments will be used to examine how individuals evaluate the school district in using scales to measure effectiveness, equity, efficiency and red tape using a set of four regression equations that include all three treatments. Although we do not include a red tape treatment, we include this variable as a placebo to determine if individuals respond to the concept even though no information is specifically provided. The regression equations take the following form:

$$\text{Effectiveness} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e \quad (1)$$

$$\text{Equity} = \beta_4 X_1 + \beta_5 X_2 + \beta_6 X_3 + e \quad (2)$$

$$\text{Efficiency} = \beta_7 X_1 + \beta_8 X_2 + \beta_9 X_3 + e \quad (3)$$

$$\text{Red Tape} = \beta_{10} X_1 + \beta_{11} X_2 + \beta_{12} X_3 + e \quad (4)$$

Where X_1 is the overall test score treatment, X_2 is the low income test score treatment, and X_3 is the cost per student treatment, and e is an error term.

In the naïve view of tradeoffs, the base set of hypotheses predict a positive relationship with the treatment and the corresponding measure of performance, that is, overall test scores should be positively associated with effectiveness (H1), low income or African American test scores should be positively correlated with equity (H2), and costs per student should be negatively associated with efficiency (H3).

Or

Hypothesis 1. $\beta_1 > 0$.

Hypothesis 2. $\beta_5 > 0$.

Hypothesis 3. $\beta_9 < 0$.

Based on past research (Hvidman & Andersen, 2016), we hypothesize that there might also be halo effects when improved performance on one dimension bleeds over into another. That is, an increase in overall test scores might influence perceptions of equity or efficiency. These halo effects will be reflected in significant coefficients for the other regression variables in the model (β2, β3, β4, β6, β7, and β8), but these coefficients should be smaller in magnitude than the respective other coefficient in the equation. Thus, overall test score performance should have the largest influence on effectiveness (H4), low income or African American test scores should have the largest influence on equity (H5), and costs per student should be most negatively associated with efficiency (H6).

Or

Hypothesis 4. $= \beta_1 > \beta_2, \beta_3$.

Hypothesis 5. $= \beta_5 > \beta_4, \beta_6$.

Hypothesis 6. $= \beta_9 < \beta_7, \beta_8$.

Since there is no manipulation of any factors related to red tape or administrative procedures, the red tape

question is included to further check for halo effects. This is essentially a null hypothesis that suggests there will be no relationship between red tape and any of the experimental treatments.

The hypotheses 2 and 3 (and by analogy hypotheses 5 and 6) are referred to as naïve hypotheses because they do not fully capture the concepts of equity and efficiency. Low income test scores (or African American scores) could increase but if overall test scores go up more, then the results could be less equitable. Similarly, a school district might operate with lower expenditures but get less out of the money it spent than a district that spent more. A better measure of equity would be reflected in the ratio of low income test scores to overall test scores (or X_2/X_1). A more sophisticated measure of efficiency would be a measure that indicated how much each increase in test score cost in terms of expenditures of (X_3/X_1, see Brunner et al., 2022 for a similar measure). This measure could be reversed to measure efficiency or simply be kept as is for a measure of inefficiency. This suggests re-estimating equations 2 and 3 as follows:

$$\text{Equity} = \beta_4 X_1 + \beta_5 X_2 + \beta_6 X_3 + \beta_{13} X_2/X_1 + e \quad (5)$$

$$\text{Efficiency} = \beta_{10} X_1 + \beta_{11} X_2 + \beta_{12} X_3 + \beta_{14} X_3/X_1 + e \quad (6)$$

and retesting the hypotheses to determine if these more sophisticated measures of equity and efficiency add any additional explanation over and above the naïve measures.

The above hypotheses examine whether individuals are sensitive to differences in overall performance (effectiveness), low income or African American performance (equity), and costs (efficiency), but do not incorporate whether such values influence their overall assessment or willingness to use a public service. Meier et al. (2022) demonstrated that perceived willingness to use a public service could be assessed with a five-point Likert scale asking how comfortable the person would be in using the service. The specific question asks if the person would be comfortable placing their child in the school district in question. Three hypotheses will be examined for effectiveness, equity, and efficiency:

Hypothesis 7. Increased overall test scores (effectiveness) will be positively associated with comfort with the school district.

Hypothesis 8. Increased low income/African American test scores (equity) will be positively associated with comfort with the school district.

Hypothesis 9. Increased expenditures per student (efficiency) will be negatively associated with comfort with the school district.

Heterogeneous effects

An extensive literature suggests that program performance information is subject to motivated reasoning whereby information consistent with pre-existing values is given greater credence than information inconsistent with pre-existing values (Baekgaard & Serritzlew, 2016; Bolsen et al., 2014). The willingness to tradeoff values like equity and efficiency after all reflects traditional U.S. partisan conflicts regarding the size of the public sector. The most likely variable that might tap motivated reasoning is partisanship given the different party orientations to issues of equity in terms of either income or race (Bartels, 2018; Einstein & Glick, 2018; Fernández-Gutiérrez & Van de Walle, 2019; Kelly, 2009; Meier & Rutherford, 2017; Westwood & Peterson, 2020). The study will use the traditional U.S. measures of partisanship to investigate whether motivated reasoning affects these relationships.

Hypothesis 10. Democratic respondents will be more sensitive to equity concerns than Republicans.

Hypothesis 11. Republican respondents will be more sensitive to efficiency concerns than Democrats.

The willingness to tradeoff equity for efficiency or the relative preference for equity relative to overall effectiveness is also likely to vary by the interests of the individual. While partisanship and anti-public sector attitudes are likely candidates, so too are some demographic factors. Based on the existing literature, we expect that women will be more likely to respond to equity concerns, and men more likely to respond to efficiency (Barnes & Cassese, 2017; Box-Steffensmeier et al., 2004; Norrander & Wilcox, 2008; Poggione, 2004).

Hypothesis 12. Equity will matter more for Women.

Hypothesis 13. Efficiency will matter more for Men.

METHODS

Research setting

To create mundane realism in the experiment, the experiments focused on a specific policy area rather than just general preferences so that subjects will need to respond not as a general philosophy of what they think is important, but rather what they think is important in a specific policy area (in contrast to Fernández-Gutiérrez & Van de

Walle, 2019). The test case will rely on public education in the United States, an area where there are longstanding performance measures and most citizens would be familiar with the service, be aware that such programs are evaluated, and perhaps have seen specific evaluations in the past. Elementary and secondary education in the United States is generally provided by independent local school districts with their own elected board and tax capacity. These districts operate under regulations and with some financial support from state governments and to a lesser degree the federal government. The area of education also has a long history of providing accessible performance data to the public which is not the case in many other public service areas (see Brunner et al., 2022).

The experiment uses a hypothetical U.S. public school district. Respondents were given a vignette describing a public school district (see Appendix S1) and then three items of information about the school district and the statewide average for all school districts in the state. To increase the external validity of the experiment (Gaines et al., 2007), we examined U.S. state education websites for how school performance data were reported. Because concepts such as effectiveness, equity, and efficiency are inherently comparative in nature, the statewide performance data were included as a comparison figure for the subjects to evaluate the hypothetical school district (see Olsen, 2017). All respondents were told that the statewide average for all school districts test scores was an overall pass rate of 80 percent, a low income pass rate of 70 percent, and per student expenditures of $9950 (see Appendix S1 for exact wording and all experiment materials). Both expenditures per student and test scores vary a great deal from state to state in the United States; in the latter case because states set their own testing standards and these vary significantly in difficulty (Linn et al., 2002). To make the experiments reflect actual circumstances as closely as possible, these figures were selected to place the hypothetical state at approximately the average of U.S. states in terms of expenditures per student; the test scores were selected also using similar criteria.

The $3 \times 3 \times 3$ factor experiment then randomly assigned subjects to the hypothetical school district with an overall test score for all students (factor 1), a low income students' test score (factor 2), and a figure for per student spending (factor 3). The overall test score values were 85, 80, or 75 for the hypothetical school district, representing performance either at the state average (80) or about 6 percent above (85) or below (75) the average (relying on whole numbers as generally reflected in exam reporting). The variable was labeled as "all test scores." The low income test score values (labeled as "low income tests") were 75, 70, and 65 to maintain comparability with the all-test scores, and the expenditures per student (labeled as "cost") took values of $10,572, $9950, and $9328. The randomly assigned variables were selected to be approximately 6 percent above or below the state average to allow for comparisons both within the indicators and across the indicators. The data were kept in this range rather than using more extreme differences to reflect the mundane realism of the comparisons most individuals would be making (Gaines et al., 2007).

The second experiment was identical to the first experiment but rather than providing test scores for low income students, subjects were given data on African American students. The same set of values were used for all test scores, African American test scores, and educational costs as in the first experiment. For both experiments, the vignette about the hypothetical school district immediately followed the consent form and screening process. Respondents then filled out the questions measuring the dependent variables followed by information on demographics and three manipulation checks. Both experiments were approved by the American University's Institutional Review Board (IRB 2021–36, IRB 2021–144).

Data collection

Subjects for the experiments were adult U.S. residents recruited using Mechanical Turk (MTurk), and we used several checks to enhance data quality (see Stritch et al., 2017). First, to ensure that respondents were actually residing in the United States rather than using Virtual Private Server (VPS), Virtual Private Networks (VPN), or a proxy to hide their country location, we required participants to deactivate any software on their machines that met the listed criteria before being allowed to participate in the study. We also used the Winter et al. (2019) protocol to screen out respondents who do not currently reside in the United States (using IPhub) or those who were using VPNs to hide their location. This tool was used in addition to imposing country residence criteria in MTurk. Second, we restricted access to the survey by allowing only one respondent from a single IPM to prevent individuals from taking the survey more than once and included a reCaptcha question at the start of the survey to prevent bots from taking the survey. Respondents were compensated $0.80 for completion of the experiment. All data and documentation to replicate these experiments can be found at 10.7910/DVN/0I991V.

The initial experiment was designed as a $3 \times 3 \times 3$ experiment with 27 total groups. Power analysis for the first experiment for an effect size of 0.1 [one tenth of a standard deviation] with an alpha of $p = .05$, df 8, and 27 groups indicated that 1650 respondents would need to be recruited. The experiment was designed, however, to be symmetrical with positive treatments and negative treatments at equal distance from the middle treatment. Preliminary analysis as a $3 \times 3 \times 3$ experiment (see the Appendix Tables A3 and A4) showed that the treatment effects were symmetrical (that is, the magnitude of the positive and negative effects were not statistically different from each other) and could be estimated as a single

interval variable. In addition, to this simplifying the presentation of results with little loss of information, a post hoc power analysis (effect size 0.27, alpha 0.05, sample size 1594, numerator df 2, number of groups 3) produced a power estimate of (1 – alpha) of 1.000 for this experiment, well above the traditional threshold of 0.80. Using this information, the second experiment used a sample of 900; the post hoc power analysis of that experiment (effect size 0.15, alpha 0.05, sample size 864, numerator df 2, number of groups 3) produced a power estimate of (1-alpha) of 0.9824, indicating the experiment was adequately powered.

MTurk and other internet samples are samples of convenience and are not fully representative of the adult population of the United States (even if artificially weighted). The current study is similar to most internet studies where the sample overrepresents white respondents, better educated respondents, and higher income respondents (See Appendix Table A2 for the composition of the samples). It underrepresents Latinos but appears to be slightly over representative of the African American population, particularly in the second experiment. The samples are also more balanced in terms of partisanship than usual internet samples that skew more Democratic. Given the randomization of the treatments, demographic controls did not affect the relationship between the treatments and dependent variables (results available from the authors).

Before analyzing the data, we conducted several checks on the experimental data. Manipulation checks after the dependent variables were run for all three treatments in both experiments comparing individuals who correctly perceived the treatment effect to those who did not generating a set of three-by-three tables; all were highly significant indicating that the respondents recognized the specific treatments that they received (Mutz & Pemantle, 2015). For the low income experiment, the chi-square statistics (4 df) were 1257.69 ($p < .00001$) for cost, 1129.30 ($p < .0001$) for all test scores, and 1021.52 ($p < .00001$) for low income test scores. For the racial equity experiment, the chi-squares were lower owing to the smaller sample, but still highly significant: cost = 405.37 ($p < .00001$), all test scores = 334.99 ($p < .00001$), and African American test scores = 136.28 ($p < .00001$).

Balance tests were performed for both experiments using age, education, income, partisanship, gender, and race for each of the three treatments. This total of 18 f-tests for each experiment. None of the 36 tests were significant at the 0.05 level.

MEASUREMENT: THE DEPENDENT VARIABLES

Five dependent variables are used in the analysis: effectiveness, equity, efficiency, red tape, and comfort with using the school district. The measures were based on existing measures used in Denmark (Hvidman & Andersen, 2016) and the United States (Meier et al., 2019), as adjusted by Meier et al. (2022) to improve the reliability of the measures. Because the concepts of effectiveness, equity, and efficiency are theoretically related to each other, and in the cases of equity and efficiency assume that a program has some degree of effectiveness, the measurement approach was to directly measure each concept and not force them to be uncorrelated with each other. In every case, the measures provided consistent reliability across the two experiments with generally similar loadings for each of the indicators.[1] Effectiveness is a factor score created using five indicators (see Appendix Table A1 for all measures and factor analysis results) with a Cronbach's alpha of 0.90 in both experiments. Equity is a factor score using three indicators with a Cronbach's alpha of 0.86 in both experiments. Efficiency used four indicators to create a factor score with a Cronbach's alpha of 0.89 in the low income experiment and 0.87 in the racial experiment. The two indicator factor score for red tape had Cronbach's alphas of 0.79 and 0.81 in the low income and race experiments respectively. The respondents were presented with the items for these scales randomly to avoid response sets that might correlate with the individual measures. The five-point Likert scale question on comfort with using the school district had a mean of 3.81 and standard deviation 1.05 in experiment 1 and a mean of 4.02 and a standard deviation of 0.98 in experiment 2.

Experiment 1 results: low income test scores

Table 1 presents the experimental treatment results for all test scores, low income test scores, and costs per student; the experiments are coded to reflect the interval treatments for all test scores (1 = 75, 2 = 80, 3 = 85), low income test scores (1 = 65, 2 = 70, 3 = 75), and costs (1 = 9328, 2 = 9950, 3 = 10,572) for two reasons. Using the actual raw scores results in small coefficients that then have to be reinterpreted to deal with the different metrics for each of the treatments and then standardized for comparison purposes. Because the treatments were created to compare approximately 6 percent changes in the variables relative to the mean value, using the interval categories also allows comparisons both within and across dependent variables.

Both increases in the overall test score and the test score for low income students are positively associated with increases in the effectiveness measure, and the relative impact is approximately equal. A 6 percent increase in overall test scores from the mean is associated with an increase of about one-fourth of a standard deviation in effectiveness (0.274); for low income test scores the increase is slightly less (0.255). The difference in coefficient sizes is not statistically significant, suggesting that respondents give equal weight to both overall performance and the performance of subgroups of students

TABLE 1 The influence of test scores, low income scores and costs on perceived performance: main effects with standardized manipulations (6.25%).

	(1) Effectiveness	(2) Equity	(3) Efficiency	(4) Red tape	(5) Comfort
All test scores	0.274**	0.047	0.212**	−0.035	0.321***
	(0.030)	(0.029)	(0.029)	(0.030)	(0.0311)
Low income tests	0.255**	0.289**	0.271**	−0.053	0.208***
	(0.029)	(0.030)	(0.029)	(0.030)	(0.0306)
Cost per student	−0.040	−0.021	−0.140**	0.052	0.0280
	(0.029)	(0.029)	(0.030)	(0.030)	(0.0306)
_cons	−0.976**	−0.631**	−0.687**	0.074	2.696***
	(0.105)	(0.103)	(0.103)	(0.104)	(0.113)
Obs.	1594	1630	1622	1637	1643
R^2	0.101	0.059	0.098	0.005	0.095

Note: Standard errors are in parenthesis. $^{**}p < .01$, $^{*}p < .05$.

when assessing overall effectiveness. In contrast, the cost per student of education has no influence on the assessment of effectiveness; it is appropriately signed (lower costs suggesting greater effectiveness) but statistically insignificant. Hypothesis 1 is supported (higher test scores are associated with greater effectiveness), but the evidence for Hypothesis 4 (overall test scores have the highest impact on effectiveness) is rejected given that low income test scores have essentially an equal influence on effectiveness (although costs do not).

The equity equation shows the clearest result in terms of the treatment effects. Only the low income test score affects the equity measure with a 6 percent increase in low income test scores from the mean resulting in slightly less than a three-tenths of the standard deviation change in perceived equity. Neither the overall test score nor the cost per student have a statistically significant impact on the equity measure. These results support hypotheses 2 (low income test scores are positively associated with equity) and Hypothesis 5 (low income test scores have the strongest influence on assessments of equity).

The results for efficiency are the most problematic in that all three manipulations affect the rating for efficiency in the correct direction (positive for both test scores, negative for cost). Respondents appear to incorporate a variety of factors in assessing efficiency including the level of performance, the performance of low income (disadvantaged) students, and the cost of education. At one level this might appear to be a sophisticated approach to efficiency where costs are considered relative to the level of performance (but see below on alternative operationalizations of efficiency). Alternatively, it could reflect a lack of understanding of the concept of efficiency (confusing it with equity or effectiveness). One possible indicator of the lack of understanding in this table is that the relative size of the cost per student variable, something that should play a larger role in assessing efficiency, is only about one-half the size of the two test score variables. The results support Hypothesis 3 linking costs to efficiency but reject Hypothesis 6 that costs would have the largest influence on assessments of efficiency.

The red tape questions were included to test for halo effects since there were no treatments related to red tape in the experiment. Only the statement "As a public school system, it must comply with all state laws and regulations on curriculum, testing, and teacher qualifications" was relevant to any aspect of red tape and that was included solely so that the questions on red tape would have some referent.[2] None of the relationships for the red tape measure are statistically significant; given the large sample size (1600+), this suggests that the experiment does not have a halo effect across performance measures and has implications for the efficiency findings. Absent halo effects, the efficiency results might indicate that individuals have a multidimensional concept of efficiency that covers more than costs or that links costs to performance in an assessment of cost-effectiveness.

Table 2 investigates more sophisticated assessments of efficiency and equity. For efficiency, a measure that combines both outcomes and costs is used by dividing the total cost per student by the overall test score. This creates essentially a cost per test score point measure of efficiency similar to Brunner et al. (2022). Because this item incorporates both costs and overall test scores, it will generate some collinearity in the equation. While the results in column 1 show that this new measure of efficiency has a negative association with respondent perceptions of efficiency, the impact is modest (0.148 standard deviations), fails to attain statistical significance at conventional levels ($p < .05$), and adds little additional explanation to the model. The likely conclusion is that there might be some subset of respondents who respond in this cost effectiveness manner, but that it is not a widespread phenomenon. An equally likely explanation is that the public does not have a clear concept of efficiency in public programs and conflates the term with other objectives for government programs.

TABLE 2 Sophisticated assessments of equity and efficiency.

	(1) Equity	(2) Efficiency
All test scores	0.054	0.113
	(0.067)	(0.061)
Low income tests	0.283**	0.271**
	(0.063)	(0.029)
Cost per student	−0.021	−0.049
	(0.029)	(0.059)
Cost/test scores		−0.148
		(0.089)
Relative equity	0.010	
	(0.085)	
_cons	−0.643**	−0.492**
	(0.153)	(0.155)
Obs.	1630	1622
R^2	0.059	0.100

Note: Standard errors are in parenthesis. **$p < .01$, *$p < .05$.

The second column of Table 2 introduces a more sophisticated version of equity by dividing the low income test scores by the all-students test scores, essentially setting up a proportional measure of equity. This measure clearly introduces excessive collinearity into the model and does not result in any additional explanation. The tentative conclusion is that assessments of equity are made in a simple, direct way rather than explicitly in comparison to the relationship between the disadvantaged group and all others.

In theory, changes in indicators of effectiveness, equity, and efficiency should not only be related to the latent variables measuring effectiveness, equity, and efficiency, but also an overall evaluation of the program. Using the comfort question as a general indicator of potential behavior, the last column of Table 1 presents the impact of the treatments on this variable. Both all test scores (supporting Hypothesis 7 on effectiveness) and low income test scores (supporting Hypothesis 8 on equity) are positively associated with assessments of comfort with using the school district with over all test scores a slightly larger influence (about one-third of a standard deviation) for a 6 percent change from the mean. Cost is an outlier and is unrelated to comfort with the program, another indicator that the public puts a lower priority on efficiency; Hypothesis 9 is rejected.

Experiment 2 results: African American test scores

The racial equality experiment was identical to the income inequality experiment in all aspects but one. Rather than the equality comparison group being low income students, the survey experiment used test scores for African American students. The main effects of the three treatments are given in Table 3. Overall, the results look very similar to those for low income students although the effect sizes are generally smaller. Respondents rate the school district higher in terms of effectiveness when it reports higher overall test scores and higher African American test scores with the coefficients approximately equal in size. Similar to the situation for the low income test scores' experiment, the cost per student has no influence on the assessments of effectiveness. These results support Hypothesis 1 on the influence of all test scores but do not support Hypothesis 4 that all test scores would have the highest impact on perceptions.

The results for perceptions of equity in Table 3 are very similar to those from Table 1; only African American test scores matter for perceptions of equity, just as only low income scores were the only significant factor in the first experiment. Equity again appears to be the criterion that is most consistently linked to the designed treatment (African American test scores) and not the other treatments that were not focused on equity. Hypotheses 2 and 5 are both supported by these results.

The results for efficiency are partially consistent with those in Table 1, but the deviation from those results is troublesome. Costs in this experiment are unrelated to assessments of efficiency, but all test scores and African American test scores are positively correlated. These results reject hypotheses 3 and 6 that predict a negative relationship between costs and assessments of efficiency and also that the cost relationship would be the strongest relative influence.

The placebo test for red tape shows a single anomalous finding. Perceptions of red tape are negatively (although very modestly, less than one tenth of a standard deviation) associated with high overall test scores. Although one might infer that a school district that attained high test scores had less red tape, this result is likely just an artifact. The estimated coefficient, on further examination, is well within the sampling range based on the first experiment.

Table 4 uses the more sophisticated definitions of equity (comparing African American test scores to overall test scores) and efficiency (dividing costs by overall test scores) to predict equity and efficiency. Neither of the two more sophisticated measures is statistically significant.

The last column of Table 3 presents the racial experiment's assessment of comfort with using the school district, the behavioral intention measure. Again, both Hypothesis 7, a positive relationship between increases in overall test scores and comfort with using the school district, and Hypothesis 8, a positive relationship with greater equity and comfort, are supported. Similar to the previous experiment, there is no relationship between the efficiency treatment and comfort with using the school district (rejecting Hypothesis 9).

TABLE 3 The racial experiment: the impact of all test scores, African American test scores and costs on perceptions of performance (6.25% changes).

	(1) Effectiveness	(2) Equity	(3) Efficiency	(4) Red Tape	(5) Comfort
All test scores	0.149**	−0.022	0.149**	−0.093*	0.222***
	(0.042)	(0.041)	(0.042)	(0.040)	(0.0410)
African American test scores	0.173**	0.213**	0.153**	−0.050	0.125***
	(0.041)	(0.042)	(0.043)	(0.041)	(0.0406)
Cost per student	−0.008	0.027	−0.053	0.048	0.0101
	(0.042)	(0.042)	(0.042)	(0.042)	(0.0398)
_cons	−0.632**	−0.438**	−0.504**	0.190	3.298***
	(0.158)	(0.153)	(0.164)	(0.148)	(0.154)
Obs.	864	872	869	875	881
R^2	0.035	0.031	0.033	0.009	0.046

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

TABLE 4 The race experiment: more sophisticated model assessments of efficiency and equity.

	(1) Equity	(2) Efficiency
All test scores	−0.113	0.037
	(0.094)	(0.090)
African American test scores	0.298**	0.154**
	(0.083)	(0.043)
Cost per student	0.027	0.049
	(0.042)	(0.078)
Cost/test scores		−0.170
		(0.129)
Relative equity	−0.137	
	(0.118)	
_cons	−0.259	−0.277
	(0.226)	(0.239)
Obs.	872	869
R^2	0.032	0.035

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

The differences between the experiments for low income students and African American students merit discussion. Racial inequities are more salient and perhaps more controversial in the United States than income inequities (Valant & Newark, 2016). One relatively consistent difference in the experimental results was that the significant findings in the race experiment were smaller in substantive magnitude than those in the low income experiment. Low income test scores generated larger regression coefficients than did the corresponding African American test scores in all three substantive cases (effectiveness, equity, and efficiency); in the cases of effectiveness and efficiency, these differences were statistically significant. Similarly, the two significant coefficients for all test scores (effectiveness and efficiency) were higher in the low income experiment, the former statistically so. And cost only mattered in the low income case and that just for efficiency. Whether these differences reflect the relatively unwillingness to consider racial differences compared with income differences or some other factor cannot be determined from the existing data.

Trading off effectiveness, equity, and efficiency

The previous results were presented in regressions that estimated the impact of each treatment controlling for the other two treatments. This means that the influence of low income test scores on equity is estimated while holding constant overall test scores and costs. This estimation would dampen any tradeoff effects between the treatments. Rather a tradeoff effect would require that the impact of low income test scores would change as the values of all test scores (or costs) would change. This logic implies that the treatments would interact with each other, and the influence of any individual treatment would depend in part on the values of the other treatments. Tradeoffs between treatments of effectiveness and equity specifically would suggest that as all test scores increased, the value of low income test scores would decline in influence (that is, the interaction effect would be negative).[3]

Table 5 presents three regressions with all three of the treatment effects interacted with each other. None of the nine coefficients for the interactions is statistically significant, suggesting that the hypothesis that respondents make tradeoffs between effectiveness, equity, and efficiency should be rejected in this case. Table 6 presents a similar set of regressions for the racial equity experiment. Again, none of the interactions reach the standard 0.05 level of statistical significance, indicating that the tradeoffs hypothesis should be rejected.

TABLE 5 Examining tradeoff effects: interactions among the treatments for low income experiment.

	(1) Effectiveness	(2) Equity	(3) Efficiency
All test scores	0.318**	0.058	0.145
	(0.105)	(0.107)	(0.106)
Low income tests	0.208*	0.270**	0.177
	(0.104)	(0.102)	(0.102)
Cost per student	−0.181	−0.129	−0.306**
	(0.111)	(0.107)	(0.112)
All test scores × Low income scores	−0.034	−0.025	−0.001
	(0.035)	(0.035)	(0.035)
All test scores × Cost per student	0.012	0.020	0.034
	(0.035)	(0.034)	(0.036)
Low income scores × Cost per student	0.057	0.034	0.048
	(0.035)	(0.037)	(0.037)
_cons	−0.828**	−0.512	−0.360
	(0.275)	(0.269)	(0.275)
Obs.	1594	1630	1622
R^2	0.103	0.060	0.100

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

TABLE 6 Examining tradeoff effects: the interaction of treatments for the racial equity experiment.

	(1) Effectiveness	(2) Equity	(3) Efficiency
All test scores	0.336*	−0.023	0.242
	(0.147)	(0.152)	(0.148)
Low income tests	0.369*	0.315*	0.250
	(0.159)	(0.155)	(0.166)
Cost per student	0.039	0.116	−0.096
	(0.174)	(0.162)	(0.187)
All test scores × Low income scores	−0.085	−0.003	−0.058
	(0.050)	(0.050)	(0.052)
All test scores × Cost per student	−0.009	0.004	0.012
	(0.052)	(0.050)	(0.054)
Low income scores × Cost	−0.014	−0.047	0.009
	(0.053)	(0.053)	(0.055)
_cons	−1.060*	−0.632	−0.647
	(0.429)	(0.405)	(0.449)
Obs.	864	872	869
R^2	0.039	0.032	0.035

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

Heterogeneous responses: partisanship

Our preregistration of the experiment hypothesized two cases where heterogeneous responses were expected—partisanship and gender. The general partisan orientation of U.S. politics suggested that Democrats would be less concerned with efficiency (Hypothesis 11) and more concerned with equity (Hypothesis 10) than Republicans and that motivated reasoning would result in discounting of the performance information presented. Table 7 examines the low income experiment and restricts the analysis to partisans (omitting Independent identifiers) and interacts the three treatment effects by a dummy variable indicating a Republican Identifier. In all four equations, the joint f-tests reject the null hypothesis that there is no difference between Democrats and Republicans in terms of the relationships between the treatments and the performance assessments. Examining the individual coefficients, however, indicates that these results are generated primarily by the equity concerns as represented by the low income test score treatment. Democrats give more credit for higher low income test scores than Republicans do in terms of not just equity, but also efficiency and effectiveness (supporting Hypothesis 10, but rejecting Hypothesis 11). The red tape assessment is revealing in this case, since there was no information provided on red tape, but Democrats perceived less red tape when low income test scores were higher than did Republicans. The findings of this table should not be taken as evidence of motivated reasoning; both groups of individuals respond in the same way to all treatments,

TABLE 7 Heterogeneous responses by partisanship: low income experiment (democrats as excluded category).

	(1) Effectiveness	(2) Equity	(3) Efficiency
All test scores	0.275**	0.037	0.225**
	(0.044)	(0.043)	(0.045)
Low income tests	0.290**	0.346**	0.312**
	(0.046)	(0.048)	(0.045)
Cost per student	−0.070	−0.050	−0.182**
	(0.044)	(0.044)	(0.045)
Republican	0.590**	0.678**	0.550*
	(0.238)	(0.225)	(0.229)
Republican × All test scores	−0.042	0.058	−0.049
	(0.070)	(0.065)	(0.068)
Republican × Low income tests	−0.134*	−0.229**	−0.146*
	(0.069)	(0.069)	(0.068)
Republican × Cost per student	0.012	0.023	0.086
	(0.066)	(0.064)	(0.067)
_cons	−1.051**	−0.784**	−0.798**
	(0.158)	(0.154)	(0.154)
Obs.	1148	1178	1172
R^2	0.109	0.090	0.117

Joint Significance: (1) Republican = 0; (2) Republican × All test scores = 0; (3) Republican × Low income tests = 0; (4) Republican × Cost = 0

	Effectiveness	Equity	Efficiency
F-stat	6.18	14.61	9.85
Prob > *F*	0.0001	0.0000	0.0000

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

TABLE 8 Heterogeneous responses by partisanship: racial equity experiment (democrats as excluded category).

	(1) Effectiveness	(2) Equity	(3) Efficiency
All test scores	0.339**	0.097	0.354**
	(0.077)	(0.079)	(0.080)
African American test scores	0.205**	0.269**	0.204**
	(0.073)	(0.080)	(0.076)
Cost per student	−0.050	0.058	−0.130*
	(0.078)	(0.083)	(0.077)
Republican	1.086**	1.127**	1.158**
	(0.335)	(0.352)	(0.349)
Republican × All test scores	−0.339**	−0.171	−0.376**
	(0.092)	(0.094)	(0.094)
Republican × African American test scores	−0.098	−0.147	−0.151
	(0.090)	(0.096)	(0.092)
Republican × Cost per student	0.079	−0.031	0.137
	(0.094)	(0.098)	(0.092)
_cons	−1.140**	−1.033**	−1.000**
	(0.271)	(0.298)	(0.294)
Obs.	698	705	701
R^2	0.083	0.074	0.090

Joint Significance: (1) Republican = 0; (2) Republican × All test scores = 0; (3) Republican × Low income tests = 0; (4) Republican × Cost = 0

	Effectiveness	Equity	Efficiency
F-stat	8.17	8.28	8.49
Prob > *F*	0.0000	0.0000	0.0000

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

TABLE 9 Heterogeneous responses by gender low income experiment (males as excluded category).

	(1) Effectiveness	(2) Equity	(3) Efficiency
All test scores	0.344**	0.119**	0.260**
	(0.042)	(0.041)	(0.041)
Low income tests	0.249**	0.236**	0.250**
	(0.041)	(0.044)	(0.041)
Cost per student	−0.003	−0.019	−0.119**
	(0.041)	(0.041)	(0.042)
Woman	0.326	−0.026	0.148
	(0.209)	(0.206)	(0.206)
Woman × All test scores	−0.137*	−0.145*	−0.097
	(0.059)	(0.058)	(0.059)
Woman × Low income tests	0.015	0.105	0.045
	(0.058)	(0.061)	(0.058)
Woman × Cost per student	−0.067	0.001	−0.038
	(0.058)	(0.058)	(0.059)
_cons	−1.151**	−0.619**	−0.767**
	(0.149)	(0.144)	(0.143)
Obs.	1584	1620	1613
R^2	0.105	0.066	0.099

Joint Significance: (1) Woman = 0; (2) Woman × All test scores = 0; (3) Woman × Low income tests = 0; (4) Woman × Cost = 0

	Effectiveness	Equity	Efficiency
F-stat	2.21	3.00	1.06
Prob > *F*	0.0657	0.0177	0.3761

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

Democrats, however, simply place a higher relative value on low income test scores than do Republicans.

The racial equity experiment showed even stronger partisan responses to the equity, effectiveness, and efficiency assessments (Table 8). The joint f-tests are all highly significant indicating that Republicans respond differently to these treatments than Democrats did. Unlike the low income experiment, where the divergent responses were solely on the low income scores, in the racial equity experiment both the all-test score and the African American test score coefficients were significantly different from each other (confirming Hypothesis 10 on equity but rejecting Hypothesis 11 on efficiency). In essence, most of the test score coefficients for Republicans are negative so that they counterbalance out the positive Democratic assignments which means that the Republican respondents appear to give little to no credit to test scores when assessing effectiveness, equity, or efficiency. These findings suggest that race is the catalytic factor in generating motivated reasoning in regard to test scores and partisanship but that income levels do not. Republican respondents consistently discount positive performance information when linked to race.

Heterogenous effects: gender

Existing literature as well as the partisan gender gap in U.S. politics suggests that women might be more sensitive to equity concerns than men. Table 9 presents the interactions between the three treatments and gender for the four dependent variables. The joint f-tests indicate that women do respondent differently to the treatments in terms of equity and red tape, with border line results in terms of effectiveness; the efficiency tests are insignificant. The response pattern for women generally shows a lower sensitivity to the all-test scores and slightly more sensitivity to the low income test scores. Running the regression only for women (results not shown) shows a higher regression coefficient for low income scores than for all test scores in terms of effectiveness (0.264 vs. 0.207), equity (0.341 vs. −0.26), efficiency (0.294 vs. 0.163); the coefficients for equity and efficiency are statistically different from each other (confirming Hypothesis 12 on equity but rejecting Hypothesis 13 on efficiency).

The racial equity experiment reveals more consistent responses in regard to gender with one strong exception

TABLE 10 Heterogeneous responses by gender in racial equity experiment (males as excluded category).

	(1) Effectiveness	(2) Equity	(3) Efficiency
All test scores	0.111	−0.047	0.111
	(0.056)	(0.052)	(0.057)
African American test scores	0.124*	0.142**	0.124*
	(0.054)	(0.054)	(0.056)
Cost per student	0.002	0.027	−0.030
	(0.056)	(0.053)	(0.056)
Woman	−0.392	−0.589	−0.290
	(0.317)	(0.306)	(0.327)
Woman × All test scores	0.077	0.054	0.084
	(0.084)	(0.082)	(0.086)
Woman × African American test scores	0.114	0.160	0.065
	(0.084)	(0.085)	(0.087)
Woman × Cost per student	−0.033	−0.017	−0.063
	(0.086)	(0.085)	(0.086)
_cons	−0.444*	−0.160	−0.363
	(0.224)	(0.203)	(0.232)
Obs.	862	869	866
R^2	0.040	0.044	0.038

Joint Significance: (1) Woman = 0; (2) Woman × All test scores = 0; (3) Woman × Low income tests = 0; (4) Woman × Cost = 0

	Effectiveness	Equity	Efficiency
F-stat	0.83	2.63	0.95
Prob > *F*	0.5067	0.0330	0.4365

Note: Standard errors are in parentheses. **$p < .01$, *$p < .05$.

(Table 10). There are no differences in responses by gender for either effectiveness or efficiency as indicated by the insignificant joint f-tests. In terms of equity, however, women show greater sensitivity to African American tests scores than men (0.302 vs. 0.142) although the difference falls just below the threshold of statistical significance (rejecting Hypotheses 12 and 13). At each level of African American test scores, women rate the school district lower than men, indicating that they are more likely to penalize school districts for low African American test scores when assessing equity (if African American test score = 65, men −0.018, women −0.447; at 70, men 0.124, women −0.145; at 75, men +0.266, women +0.157).

DISCUSSION

Programs frequently have multiple outcomes and serve many stakeholders. This multidimensional program space logically implies that public evaluation of programs will vary based on how such programs attain different values or how individuals feel about the underlying values in question. Using two vignette experiments, this study sought to determine if the general public was able to distinguish among programs that varied in terms of effectiveness, equity, and efficiency. Overall in both experiments, we found evidence that the public could distinguish among effectiveness, equity, and efficiency as evaluative criteria. Although the assessments of equity and efficiency were not sophisticated ones, they were generally consistent with the use of these values to assess program outcomes. The values were treated by the public as standalone values, that is, they did not interact. While those values did vary by partisanship and gender, there was only modest evidence of motivated reasoning that discounted the information provided and that only for race and partisanship. For both effectiveness and equity, improvements were also associated with increased comfort in using the public service (efficiency had no impact).

In terms of theoretical contributions, these findings are supportive of the underlying philosophy of the performance management movement which implies that the public can hold governments responsible for meeting their expectations and that they can use more than a simple one-dimensional assessment in this evaluation. The public's ability to distinguish differences in equity was particularly acute as they responded to the equity measure but generally ignored overall performance and costs when asked about equity. While the response on efficiency to costs was also present, it was muddied by responses to other indicators that might not have been directly related to the concept of efficiency. Understanding what the public interprets as "efficiency" remains an important empirical and theoretical topic of interest (for similar results when efficiency is separated from effectiveness see Brunner et al., 2022).

The asymmetrical relationship between effectiveness and equity may have implications for what individuals expect from government programs. Overall assessments of effectiveness were influenced both by overall performance but also by subgroup performance (that is, by low income and African American test scores), but only subgroup performance affected equity evaluations. This finding suggests that individuals include equity when they are thinking about how government programs should work and may have some concept of fairness embedded in their assessments of effectiveness (see An et al., 2023; Entress et al., 2022). Future research should probe these relationships.

Given the dominant role that efficiency plays in the literature on program evaluation and the various discussions of policy scholars, the current experiments raise questions about whether or not this is a major concern of the public. While there was some modest sensitivity to costs, it had less impact on assessments of efficiency than either overall performance or subgroup performance. In addition, a more sophisticated measure of efficiency that was relative to the level of performance did not substantially improve evaluations. While these findings are consistent with Le Grand's argument that efficiency can only

be a secondary value, further analysis is clearly warranted in this regard.

The experimental results have practical implications for public managers. Given the public's ability to respond to questions of both equity and effectiveness, managers should increase the available information on how their programs are equitable as well as their effectiveness (see also Ruijer et al., 2023). The mixed findings on efficiency might also reflect how well public managers communicate in regard to efficiency. While figures on costs are generally available, actual information comparing costs to benefits (as in this case costs to test scores) is usually not reported. Public managers should investigate whether there are effective ways to communicate performance data that relate to efficiency.

This pair of experiments was the first effort to determine if the public could respond to differences in program outcomes based on effectiveness, equity, and efficiency. As an initial experiment, we should be aware of the potential limitations. First, we attempted to design this experiment to present distinct indicators of a single dimension of performance (test scores) with clear comparison criteria (state averages). We specifically tried to avoid information that might be considered ambiguous or framed in any way to bias perceptions. Not all performance information is presented in this way, and variations in presentation can result in framing effects or bias via motivated reasoning. The results could well differ if information is not presented in a concise and easy to interpret manner.

Second, experiments are focused on the internal validity for causal inference; as such they often use convenience samples including internet samples of subjects that are not representative of some larger population. It is also possible although unlikely that individuals may have participated in both experiments (fielded 2 months apart). Data quality can also be enhanced by following the protocols for screening respondents and checking for quality responses as recommended by Stritch et al. (2017) as this research did (see also the psychometric study by Chmielewski & Kucker, 2022). An extensive line of research beginning with Berinsky et al. (2012) indicates that MTurk samples generate results comparable to quality surveys that are designed to be more representative (see for example Chmielewski & Kucker, 2022; Zhang & Gearhart, 2020). At the same time, the generalizability of experiments on any sample needs to be replicated by similar experiments based on different selection methods.

Third, we only assessed equity in terms of outcomes and only for two subgroups (low income and African American students). Equity has many dimensions and assessments on equality of opportunity, equality of process, or other dimensions of equity would be valuable additions to our knowledge. Similarly, we did find more concern with equity in terms of income than race suggesting that the subgroup involved matters. Additional work assessing differences based on ethnicity, age, sexual orientation, disability, or other factors should be explored. The literature on social construction, in fact, suggests that public evaluations are likely to vary based on the perceived deservingness of the group under consideration.

Fourth, the experiment only dealt with education policy, an area with a set of factors that conditions any evaluation of performance information. Performance indicators, including test scores, are widely used and frequently publicized as measures of performance even if contested in the United States. This means that individuals will be more familiar with this case and possibly more informed than if the program involved job training, hunger programs, or other less visible and less salient programs. Although theoretically the findings should be generalizable to other areas of public policy and other contexts with clearly communicated performance items that are commonly presented to the public, only replications in other policy areas and other countries can determine how general the findings here are.

Finally, we examined tradeoffs only in the sense of whether each of the assessment of effectiveness, equity, and efficiency varied as the other values increased or decreased. We did not examine the respondents' willingness to pay for improvements in either effectiveness or equity. This topic merits additional experimental investigation.

CONCLUSION

This study used two preregistered vignette experiments to determine whether individuals could judge program performance on different dimensions. Both experiments showed that individuals were generally able to distinguish between program effectiveness, equity, and efficiency relative to changes in program outcomes. The evaluations were most focused in terms of equity and least focused in terms of efficiency, and there did not appear to be any tradeoffs among the values in the responses of the individuals. Although such preferences did vary among groups of individuals based on partisanship and gender, these differences generally reflected how much groups valued equity rather than discounting the information that was presented.

ENDNOTES

[1] In additional analysis, we forced efficiency to be unrelated to effectiveness and found relatively similar findings for the efficiency results in Tables 1 and 3. The reliability of this orthogonal measure of efficiency however was relatively low, alpha = 0.5. This low reliability suggests that permitting overlap in the measures more likely reflects how the concepts are linked both theoretically and in public perceptions.

[2] Some studies ask about red tape but do not provide any information on it per se (see Hvidman & Andersen, 2016; Meier et al., 2019). Follow ups with individuals who pretested the experiment noted that they did not know how to evaluate an organization if no relevant information was provided so this phrase was added.

[3] Positive interaction coefficients would likely reveal halo effects as increases in a treatment designed for one dependent variable created

a halo effect for a different dependent variable. We do not expect such effects given the results of the red tape placebo results presented previously.

REFERENCES

Amirkhanyan, Anna A., Kenneth J. Meier, Miyeon Song, Fei W. Roberts, Joohyung Park, Dominik Vogel, Nicola Bellé, Angel Molina, and Thorbjørn Sejr Guul. 2023. "Liberté, Égalité, Crédibilité1: An Experimental Study of Citizens' Perceptions of Government Responses to COVID-19 in Eight Countries." *Public Administration Review* 83(2): 401–18.

An, Brian Y., Simon Porcher, Shui-Yan Tang, and Oriane Maille-Lefranc. 2023. "COVID-19 Emergency Policies, Financial Security, and Social Equity: Worldwide Evidence." *Public Administration Review*. https://doi.org/10.1111/puar.13652.

Andrews, Rhys, and Tom Entwistle. 2013. "Four Faces of Public Service Efficiency." *Public Management Review* 15(2): 246–64.

Andrews, Rhys, and Steven Van de Walle. 2013. "New Public Management and Citizens' Perceptions of Local Service Efficiency, Responsiveness, Equity and Effectiveness." *Public Management Review* 15(5): 762–83.

Baekgaard, Martin, and Søren Serritzlew. 2016. "Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension." *Public Administration Review* 76(1): 73–82.

Barnes, Tiffany D., and Erin C. Cassese. 2017. "American Party Women: A Look at the Gender Gap within Parties." *Political Research Quarterly* 70(1): 127–41.

Bartels, Larry M. 2018. *Unequal Democracy: The Political Economy of the New Gilded Age*. Princeton, NJ: Princeton University Press.

Belardinelli, Paolo, Nicola Bellé, Mariafrancesca Sicilia, and Ileana Steccolini. 2018. "Framing Effects under Different Uses of Performance Information." *Public Administration Review* 78(6): 841–51.

Belle, Nicola, and Paola Cantarelli. 2022. "Your Money, your Life, or your Freedom? A Discrete-Choice Experiment on Trade-Offs during a Public Health Crisis." *Public Administration Review* 82(1): 59–68.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(2): 351–68.

Bolsen, Toby, James N. Druckman, and Fay Lomax Cook. 2014. "The Influence of Partisan Motivated Reasoning on Public Opinion." *Political Behavior* 36(2): 235–62.

Box-Steffensmeier, Janet M., Suzanna De Boef, and Tse-Min Lin. 2004. "The Dynamics of the Partisan Gender Fap." *American Political Science Review* 98(3): 515–28.

Boyne, George A. 2002. "Public and Private Management: What is the Difference?" *Journal of Management Studies* 39(1): 97–122.

Boyne, George A. 2003. "Sources of Public Service Improvement: A Critical Review and Research Agenda." *Journal of Public Administration Research and Theory* 13(3): 367–94.

Brunner, Eric J., Mark D. Robbins, and Bill Simonsen. 2022. "Citizen Perceptions of Public School Efficiency: Evidence from the US." *Public Management Review*. https://doi.org/10.1080/14719037.2022.2158211.

Chmielewski, Michael, and Sarah C. Kucker. 2022. "An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results." *Social Psychological and Personality Science* 11(4): 464–73.

Chun, Young Han, and Hal G. Rainey. 2005. "Goal Ambiguity in US Federal Agencies." *Journal of Public Administration Research and Theory* 15(1): 1–30.

Daw, Jonathan. 2015. "Explaining the Persistence of Health Disparities." *American Journal of Sociology* 120(6): 1595–640.

de Graaf, Gjalt, and Hester Paanakker. 2015. "Good Governance: Performance Values and Procedural Values in Conflict." *American Review of Public Administration* 45(6): 635–52.

Dunn, William N. 2015. *Public Policy Analysis*. New York: Routledge.

Einstein, Katherine L., and David M. Glick. 2018. "Mayors, Partisanship, and Redistribution: Evidence Directly from US Mayors." *Urban Affairs Review* 54(1): 74–106.

Entress, Rebecca M., Jenna Tyler, and Abdul-Akeem Sadiq. 2022. "Inequity after Death: Exploring the Equitable Utilization of FEMA's COVID-19 Funeral Assistance Funds." *Public Administration Review*. https://doi.org/10.1111/puar.13572.

Favero, Nathan, and Amanda Rutherford. 2020. "Will the Tide Lift all Boats? Examining the Equity Effects of Performance Funding Policies in US Higher Education." *Research in Higher Education* 61(1): 1–25.

Fernández-Gutiérrez, Marcos, and Steven Van de Walle. 2019. "Equity or Efficiency? Explaining Public Officials' Values." *Public Administration Review* 79(1): 25–34.

Frederickson, H. George. 2015. *Social Equity and Public Administration: Origins, Developments, and Applications: Origins, Developments, and Applications*. New York: Routledge.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(1): 1–20.

Gershberg, Alec Ian, and Til Schuermann. 2001. "The Efficiency–Equity Trade-off of Schooling Outcomes." *Economics of Education Review* 20(1): 27–40.

Hall, Jeremy L. 2022. "In Search of Social Equity in Public Administration: Race, Gender, and some "Class"-Ey New Ideas." *Public Administration Review* 82(3): 381–5.

Hood, Christopher. 1991. "A Public Management for all Seasons?" *Public Administration* 69(1): 3–19.

Hvidman, Ulrik, and Simon Calmar Andersen. 2016. "Perceptions of Public and Private Performance: Evidence from a Survey Experiment." *Public Administration Review* 76(1): 111–20.

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

James, Oliver, and Gregg G. Van Ryzin. 2019. "Rates and the Judgment of Government Performance." *Journal of Behavioral Public Administration* 2(2). https://doi.org/10.30636/jbpa.22.41.

Jenkins-Smith, Hank. 1990. *Democratic Politics and Policy Analysis*. Monterry, CA: Brooks-Cole.

Johnson, Austin P., Nehemia Geva, and Kenneth J. Meier. 2019. "Can Hierarchy Dodge Bullets? Examining Blame Attribution in Military Contracting." *Journal of Conflict Resolution* 63(8): 1965–85.

Kelly, Nathan J. 2009. *The Politics of Income Inequality in the United States*. New York: Cambridge University Press.

Le Grand, Julian. 1990. "Equity Versus Efficiency: The Elusive Trade-off." *Ethics* 100(3): 554–68.

Linn, Robert L., Eva L. Baker, and Damian W. Betebenner. 2002. "Accountability Systems: Implications of Requirements of the No Child Left behind Act of 2001." *Educational Researcher* 31(6): 3–16.

Little, Ian Malcolm David. 2002. *A Critique of Welfare Economics*. New York: Oxford University Press.

Marvel, John D. 2016. "Unconscious Bias in Citizens' Evaluations of Public Sector Performance." *Journal of Public Administration Research and Theory* 26(1): 143–58.

Marvel, John D., and Amanda M. Girth. 2016. "Citizen Attributions of Blame in Third-Party Governance." *Public Administration Review* 76(1): 96–108.

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): 820–8.

Meier, Kenneth J., and Amanda Rutherford. 2017. *The Politics of African-American Education: Representation, Partisanship, and Educational Equity*. New York: Cambridge University Press.

Meier, Kenneth J., Miyeon Song, Jourdan A. Davis, and Anna A. Amirkhanyan. 2022. "Sector Bias and the Credibility of Performance Information: An Experimental Study of Elder Care Provision." *Public Administration Review* 82(1): 69–82.

Mutz, Diana C., and Robin Pemantle. 2015. "Standards for Experimental Research: Encouraging a Better Understanding of Experimental Methods." *Journal of Experimental Political Science* 2(2): 192–215.

Norrander, Barbara, and Clyde Wilcox. 2008. "The Gender Gap in Ideology." *Political Behavior* 30(4): 503–23.

Okun, Arthur M. 1975. *Equality and Efficiency: The Big Tradeoff*. Washington DC: Brookings Institution Press.

Olsen, Asmus Leth. 2015. "Citizen (Dis) Satisfaction: An Experimental Equivalence Framing Study." *Public Administration Review* 75(3): 469–78.

Olsen, Asmus Leth. 2017. "Compared to What? How Social and Historical Reference Points Affect Citizens' Performance Evaluations." *Journal of Public Administration Research and Theory* 27(4): 562–80.

Pedersen, Rasmus T. 2017. "Ratio Bias and Policy Preferences: How Equivalency Framing of Numbers Can Affect Attitudes." *Political Psychology* 38(6): 1103–20.

Piatak, Jaclyn, Zachary Mohr, and Suzanne Leland. 2017. "Bureacratic Accountability in Third Party Governance." *Public Administration* 95 (4): 976–89.

Piketty, Thomas. 2015. *Capital in the Twenty-First Century*. Cambridge, MA: Belnap Press of Harvard University Press.

Poggione, Sarah. 2004. "Exploring Gender Differences in State legislators' Policy Preferences." *Political Research Quarterly* 57(2): 305–14.

Redford, Emmette S. 1969. *Democracy in the Administrative State*. New York: Oxford University Press.

Ruijer, Erna, Gregory Porumbescu, Rebecca Porter, and Suzanne Piotrowski. 2023. "Social Equity in the Data Era: A Systematic Literature Review of Data-Driven Public Service Research." *Public Administration Review* 83(2): 316–32.

Schmidthuber, Lisa, Jurgen Willems, and Bernhard Krabina. 2023. "Trust in Public Performance Information: The Effect of Data Accessibility and Data Source." *Public Administration Review* 83(2): 279–95.

Stritch, Justin M., Mogens Jin Pedersen, and Gabel Taggart. 2017. "The Opportunities and Limitations of Using Mechanical Turk (Mturk) in Public Administration and Management Scholarship." *International Public Management Journal* 20(3): 489–511.

Swank, Duane. 1998. "Funding the Welfare State: Globalization and the Taxation of Business in Advanced Market Economies." *Political Studies* 46(4): 671–92.

Valant, Jon, and Daniel A. Newark. 2016. "The Politics of Achievement Gaps: US Public Opinion on Race-Based and Wealth-Based Differences in Test Scores." *Educational Researcher* 45(6): 331–46.

Viscusi, W. Kip, and Richard J. Zeckhauser. 2005. "Recollection Bias and the Combat of Terrorism." *The Journal of Legal Studies* 34(1): 27–55.

Walker, Richard M., M. Jin Lee, Oliver James, and Samuel M. Y. Ho. 2018. "Analyzing the Complexity of Performance Information Use." *Public Administration Review* 78(6): 852–63.

Wang, Weijie. 2022. "How Does Performance Management Affect Social Equity? Evidence from New York City Public Schools." *Public Administration Review*. https://doi.org/10.1111/puar.13590.

Weimer, David and Aidan Vining. 2017. *Policy Analysis: Concepts and Practice*. New York: Taylor & Francis.

Westwood, Sean J., and Erik Peterson. 2020. "The Inseparability of Race and Partisanship in the United States." *Political Behavior* 44: 1125–47. https://doi.org/10.1007/s11109-020-09648-9.

Winter, Nicholas, Tyler Burleigh, Ryan Kennedy, and Scott Clifford. 2019. A Simplified Protocol to Screen out VPS and International Respondents Using Qualtrics Available at SSRN 3327274.

Zhang, Bingbing, and Sherice Gearhart. 2020. "Collecting Online Survey Data: A Comparison of Data Quality among a Commercial Panel & MTurk." *Survey Practice* 13(1). https://doi.org/10.29115/SP-2020-0015.

AUTHOR BIOGRAPHIES

Kenneth J. Meier is a Distinguished Scholar in Residence and Director of the Summer Diversity Academy at the School of Public Affairs American University, Professor of Bureaucracy and Democracy at Leiden University (the Netherlands), and a Professor of Public Management at the Cardiff School of Business, Cardiff University (Wales). His research interests include public management, the role of bureaucracy in democratic systems, comparative public administration, behavioral approaches to public administration, and virtually everything else. Email: kmeier@american.edu

Jourdan Davis is an Assistant Professor in the Department of Political Science and Public Administration at University of North Carolina at Charlotte. Her research interests include behavioral public administration, fairness, justice, citizen evaluations, and public policy. She has published in Public Administration Review, The Review of Black Political Economy, International Public Management Journal, and the American Review of Public Administration. Email: jdavi445@uncc.edu

Xiaoyang Xu recently received her Ph.D. in Public Administration and Public Policy from American University. She is an incoming postdoctoral fellow at the Better Government Lab, McCourt School of Public Policy, Georgetown University. Her research interests include theories of representative bureaucracy, gender and public administration, public management, and budgeting. Email: xx0266a@american.edu

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Meier, Kenneth J., Jourdan Davis, and Xiaoyang Xu. 2023. "Effectiveness, Efficiency, and Equity Tradeoffs in Public Programs: A Citizen Experiment." *Public Administration Review* 1–16. https://doi.org/10.1111/puar.13690