PUBLIC PERFORMANCE & MANAGEMENT REVIEW
2021, VOL. 44, NO. 3, 451–470
https://doi.org/10.1080/15309576.2020.1835679

Routledge
Taylor & Francis Group

Evidence of an Output Bias in the Judgment of Public Performance: A Replication and Extension

Gregg G. Van Ryzin[a], Ashley Grosso[b], and Étienne Charbonneau[c]

[a]Rutgers, School of Public Affairs and Administration; [b]Rutgers The State University of New Jersey; [c]Ecole nationale d'administration publique a Montreal

ABSTRACT

Despite calls for an evidence-based focus on outcomes as a way to enhance accountability for public performance, findings from a prior study suggest that the public may be more impressed by high frequency (low cost) but ambiguous outputs (such as people served) rather than more meaningful but costly outcomes (causal effects). We attempt to replicate and extend the investigation of this output bias through a pair of survey experiments involving judgments about two evidence-based, highly effective social programs: one, an HIV/AIDS prevention program (adapted from the prior study), the other, a program for special needs high school students (Check and Connect). Our findings confirm that respondents viewed both programs more favorably when given information about mere outputs (people served) in comparison with more rigorous outcomes (causal effects). We then tested an extension of the Check and Connect experiment in which we modified the framing of cost and performance information in ways that reduced the tendency toward an output bias. We speculate on the possible mechanism that may lead to an output bias, and we discuss the implications of our findings for evidence-based public policy and management.

KEYWORDS

accountability; evidence-based policy; experimental methods; government performance; program evaluation

The performance movement has long called for a focus on evidence-based *outcomes* or results over the reporting of mere *outputs*. This call for a focus on outcomes comes from those concerned with government performance and accountability (Boyne et al., 2006; Davies & Nutley, 2000; Kettl, 2006; Moynihan, 2008; National Performance Management Advisory Commission, 2010) as well as those seeking to enhance the performance of nonprofit organizations (Kim et al., 2017; Lee & Clerkin, 2017; Morino, 2011). But an experimental study by Grosso et al. (2017) suggests that the general public may be influenced more by frequent yet ambiguous outputs (people served) than by less frequent but more meaningful outcomes (causal effects)—with important implications for evidence-based policy-making and democratic accountability. In this article, we report on a set of

CONTACT Gregg G. Van Ryzin vanryzin@rutgers.edu Rutgers, School of Public Affairs and Administration, 111 Washington Street, Newark, NJ 07102, USA.

three survey experiments that attempt to replicate this prior study as well as extend the investigation of how the public judges real social programs for which rigorous outcome evidence exists. All three experiments examine a key question: *How does reporting evidence about the outcomes (or causal effects) of a social program, in contrast to the reporting of mere outputs, influence judgments of the program?*

In the first experiment, we replicate a vignette in which respondents are asked to judge the effectiveness and efficiency of an HIV/AIDS program in California with randomly assigned real information about outputs, outcomes, and costs. This effort responds in part to calls for more replication in public management research (Walker et al., 2017). Results of our replication parallel previous findings, suggesting that respondents—contrary to performance doctrine—react more favorably to high frequency but ambiguous outputs (at-risk clients served) in comparison with more meaningful outcomes (HIV infections prevented). In the second experiment, we extend the paradigm to a program for high school students with special needs called Check and Connect, again randomly assigning real information about outputs, outcomes, and costs. We find even stronger evidence of a bias toward high frequency but ambiguous outputs (students served) over more significant outcomes (students graduating high school who would not have graduated otherwise). In both experiments, providing information on the economic benefit to society of producing an outcome (preventing HIV, graduating high school) leads respondents to judge the programs overall as being much more effective and efficient, but it does nothing to eliminate (and may even strengthen) people's output bias. But in a third experiment, we modified the wording and framing of the outputs, outcomes, and costs for Check and Connect in ways that seem to make outcomes more convincing to the public.

We suspect this tendency toward an *output bias* (as we call it) in the public's interpretation of evidence about social programs remains widespread, influencing not only laypeople but perhaps policymakers and public managers as well (although future research is needed to confirm this speculation). We propose some possible mechanisms that may lead to an output bias, and we discuss the implications of our findings for evidence-based public policy and management. In so doing, we respond to Olsen's (2015) call in the pages of this journal for more attention to the psychology of numbers in the study of performance information and to a growing body of experimental work on behavioral public performance (James et al., 2020).

Outputs and outcomes

In the long literature on performance measurement and program evaluation, there is widespread agreement on the importance of rigorous

evidence about the outcomes, or causal effects, of government programs (Crane, 1998; Davies & Nutley, 2000; Gueron & Rolston, 2013; Nussle & Orszag, 2014). Indeed, the emphasis on seeking rigorous evidence of outcomes has continued to grow in both the government and the nonprofit sectors (Morino, 2011). The Institute for Educational Sciences has maintained its What Works Clearinghouse since 2002, Congress created the Commission on Evidence-Based Policymaking in 2016, and private organizations such as Results for America and Arnold Ventures continue actively to promote a focus on evidence-based policy and practice. As Kettl (2006) observes: "If there is convergence anywhere on the [public] management reform front, it is the central role that many nations have created for performance data, especially about program outcomes" (p. 82).

Indeed, the movement for evidence-based policy is concentrated mainly on rigorous evaluation of outcomes, or causal effects, that go beyond the mere tracking and reporting of outputs. Hatry (2006) provides a useful definition that highlights the important distinction between outputs and outcomes of a program:

> Outputs are things that the program's personnel have done, not changes to outside people or changes that outside organizations have made. . . Outcomes are the events, occurrences, or changes in conditions, behavior, or attitudes that indicate progress toward a program's mission and objectives. Thus, outcomes are linked to the program's (and its agency's) overall mission—its reason for existing. (Hatry, 2006, pp. 16–17)

As we use the term here, in line with Hatry's definition, an outcome can be seen as synonymous with *impact, result, causal effect,* or other similar terms used in a general way to refer to a change in the world produced by a policy or program. We recognize that others may draw a distinction between *outcome* and *impact*, with only the latter referring to a net causal effect; but for simplicity, we use the term outcome also in a causal sense. As Hubbard (2014) writes in *Moneyball for Government*: "When it comes to government programs, we often have a lot of data about how much they cost or how many people they employ—what are often called inputs. We may also know how many people they serve and in what ways—often named outputs. The trouble is that these data don't tell us much about how the program is (or isn't) changing people's lives" (p. 14). This quote is indicative of a general consensus in the field that performance measurement and program evaluation must go beyond the tracking and reporting of mere inputs and outputs. Importantly, Hubbard goes on to explain that the assessment of genuine *outcomes* has improved thanks to the increasing use of randomized controlled trials (RCTs). Indeed, the use of RCTs, a method borrowed from the health sciences, to evaluate government and nonprofit programs began as early as the 1960s and grew in importance

over the years, especially during the period of welfare reform experimentation in the 1980s and 1990s (Gueron & Rolston, 2013). By now, the RCT has become a widely accepted gold standard in government and the nonprofit sector for rigorous causal evidence about the outcomes of social programs (Bloom, 2005; Cartwright & Hardie, 2012; Crane, 1998; Doleac, 2019).

In its final report to the president and congress, the Commission on Evidence-Based Policymaking explained its mandate this way: "Taxpayers and policymakers should receive credible information to know and understand how well the programs and policies they fund achieve their intended goals . . . Without the use of evidence in our democracy, we are only guessing at whether government programs and policies are achieving their intended goals" (Commission on Evidence-Based Policymaking, 2017, p. 106). The clear implication in this line of argument is that, given rigorous evidence, taxpayers and policymakers will make better decisions and arrive at more correct conclusions about program efficiency and effectiveness. Note the mention of taxpayers—or ordinary citizens, the focus of our study—in addition to policymakers, in the Commission's statement about the target audience for such rigorous evidence.

Mechanisms of misunderstanding

In sum, the movement toward a more evidence-based, outcome-focused approach to performance measurement and program evaluation rests in part on the assumption that the public and other audiences recognize and value rigorous outcome evidence. At the very least, the public, stakeholders and other audiences must be able to distinguish between outcomes (causal effects) from more ambiguous outputs. However, making this distinction may not be easy for many people. As mentioned, the study by Grosso et al. (2017), which involved experimentally varying information about the outputs, outcomes, and costs of a real HIV/AIDS prevention program in California, found that participants tended to exhibit a bias toward valuing more frequent yet ambiguous outputs (clients served) over less frequent but more meaningful outcomes (causal effects). What are the possible mechanisms that might help explain this apparent output bias? In other words, what would lead people to be more persuaded or convinced by ambiguous outputs over more meaningful outcomes? Here we suggest several possible mechanisms.

One mechanism at work may be the difficulty many people have with *counterfactual thinking,* which requires cognitive effort and complex reasoning (Roese & Olson, 1995/2014). For example, a public health program may provide free flu vaccines to 1,000 residents of a community, but how

many cases of the flu does such a program prevent? Some careful thought and effort is required in answering this question. Only some of the 1,000 people would have come down with the flu, and the vaccine itself may be only partially effective. Imagining what would have happened to the 1,000 people without the flu vaccine (the counterfactual) requires a fairly complex act of analytical imagination, and—indeed—may be observable only from the results of a rigorous RCT.

Another possible mechanism may be the tendency for *substitution* of simpler or more available information for more complex information. Because interpreting evidence correctly often requires complex counterfactual thinking, as just discussed, citizens may engage in substitution as a shortcut and thus simply interpret outputs as if they were outcomes. Indeed, substitution is a frequent heuristic in human judgment and decision making (Baron, 2000; Gilovich et al., 2002). For example, when hearing that 1,000 people received a flu vaccine from the local health department, citizens may simply assume that this means that 1,000 cases of the flu were prevented. This interpretation makes initial intuitive sense until reflection reveals that it rests on the unrealistic assumption that everyone vaccinated would come down with flu and that the vaccine is 100% effective.

An additional mechanism could be tied to the fact of the much greater *frequency* of outputs relative to outcomes. Take again the example of 1,000 people receiving a flu vaccine from the health department. Only about 10% of people will get the flu in a given year, and the flu vaccine is only about 50% effective (these numbers are approximately correct in magnitude, according to the U.S. Centers for Disease Control and Prevention). This implies that only about 100 people were actually at risk of getting the flu to begin with and (given the efficacy of the vaccine) only about 50 actual flu cases would be prevented. If people are susceptible to a kind of more-is-better heuristic, they may simply view 1,000 flu vaccines administered as a more impressive or more convincing number than 50 cases of the flu prevented.

When cost information is added to the picture, people may also succumb to a kind of *sticker shock* at the relatively high cost of producing program outcomes relative to outputs (Schueler & West, 2016). To continue with the flu vaccine example, say the program costs $20,000 in total and serves 1,000 people, which implies a fairly low cost of just $20 per person served (per output). But again, say only 100 or those 1,000 people were actually at risk of getting the flu and, given the partial efficacy of the vaccine, only 50 cases of the flu would be prevented (the causal effect). In this case, the cost of preventing a case of the flu (an outcome) is the $20,000 total cost divided by 50 cases prevented, or $400 per flu case prevented—a heftier price

tag. This sticker shock, as it were, from revealing the true price of producing an outcome may lead people to generally look more favorably upon the much lower-cost outputs—even without knowing how effective the outputs truly are at causally reducing the number of flu cases.

Taken together, these mechanisms suggest that the public and other audiences may tend to judge outputs *more* favorably because they are more easily interpretable, more frequent, less costly, and perhaps even misunderstood to be outcomes. In contrast, people will tend to judge outcomes *less* favorably because they are more complex to interpret, fewer in number, and relatively expensive, despite being rigorously demonstrated causal effects of a program or intervention. These propositions have important and rather counterintuitive implications for our investigation of how reporting evidence about the outcomes (or casual effects) of a social program, in contrast to the reporting of mere outputs, influences judgments of the program.

Study 1: Design and method

To probe this question, we first replicate a modified survey-vignette experiment from Grosso et al. (2017) in which respondents are asked to evaluate the effectiveness and efficiency of the HIV Transmission Prevention Project (HTPP) in California, with randomly assigned real information about outputs, outcomes, and costs. By using somewhat different measures and an independent online sample, this replication can be considered as an empirical generalization (Walker et al., 2019). In the second vignette, we extend the paradigm to a program for high school students with special needs called Check and Connect, again randomly assigning real information about outputs, outcomes, and costs. This more conceptual replication with a different policy context enables us to better evaluate the robustness and external validity of our findings of an output bias. We discuss both programs and the design of the experimental vignettes below.

HIV Transmission Prevention Project (HTPP)

We used a modified (shortened) version of an experimental vignette from Grosso et al. (2017) that was based on a report by the California Department of Health Services (2006), entitled *Economic Evaluation of California's Prevention Case Management Intervention for HIV-Positive and HIV-Negative Persons: The HIV Transmission Prevention Project (HTPP).* Program staff at 11 sites in California implemented HTPP, which was an intervention developed to reduce risk behaviors through one-on-one sessions built around incremental steps toward long-term behavior change.

N=840 US adults			
California has one of the largest populations of people living with HIV/AIDS in the United States. AIDS is a disease caused by the HIV virus, which can be spread through sexual contact, sharing needles for injection, and from mother to child during pregnancy, childbirth and breastfeeding. Without treatment, AIDS severely weakens the immune system and leads to opportunistic infections and even death. Since there is no cure for AIDS, the California Department of Health Services launched the HIV Transmission Prevention Project (HTPP). HTPP was implemented in several high-need locations in the state and involved one-on-one sessions with at-risk people to encourage long-term behavior change and prevent HIV infections.			
Random assignment to 1 of 4 conditions →			
[A] Outcomes	*[B] Benefit to society + Outcomes*	*[C] Outputs*	*[D] Benefit to society + Outputs*
HTPP cost $551,478 and the program successfully prevented HIV infection in 11 people. This represents a cost of $50,134 per infection prevented.	The estimated lifetime cost of treating a person with AIDS is over $250,000. HTPP cost $551,478 and the program successfully prevented HIV infection in 11 people. This represents a cost of $50,134 per infection prevented.	HTPP cost $551,478, and the program served 104 people. This represents a cost of $5,303 per person served.	The estimated lifetime cost of treating a person with AIDS is over $250,000. HTPP cost $551,478, and the program served 104 people. This represents a cost of $5,303 per person served.
Dependent variable (3-items)			
Based on this information, do you agree or disagree that the HTPP program is (1) an effective program, (2) an efficient program, (3) a good investment of tax dollars. (Strongly disagree=1, to Strongly agree=5)			

Figure 1. HIV Transmission Prevention Project (HTPP) Experimental Vignette (Study 1).

In addition, the sessions provided an opportunity to identify and address clients' basic needs, such as housing or medical care. The report's estimates of the number of HIV infections prevented relied on a Bernoulli-process model that included the number of sexual partners and the proportion of unprotected (versus protected) acts of intercourse, measured by self-administered questionnaires shortly after enrollment into the program as well as 6, 12, and 18 months later. It is important to point out that the HTPP evaluation was *not* an RCT, although it was a causal impact evaluation. A benefit-cost analysis was included in the HTPP report based on program budgets and published medical care cost data. Following the vignette-construction advice of Rooks and colleagues (2000), we chose to represent a limited amount of information in each vignette to reduce cognitive overload and enhance the validity and reliability of respondents' judgments.

The full text of the vignette, including the four experimental arms, is presented in Figure 1. Respondents were randomized to one of the four arms, which represent a 2×2 factorial variation of information. Arms A and B represent the *outcome* scenarios, as they mention, "the program successfully prevented HIV infection in 11 people." In contrast, arms C and D represent the *output* scenarios, as they only mention, "the program served 104 people." Arms B and C include the lifetime cost of treating a person with AIDS (over $250,000 at the time of the HTPP evaluation), which provides information on the benefits to society of preventing HIV infection in a typical individual.

Check and Connect

Check and Connect is a dropout prevention program for high school students with learning, emotional, and behavioral disabilities that identifies

N=840 US adults			
Check and Connect is a government-funded dropout prevention program for high school students with learning, emotional, or behavioral disabilities. Students typically enter the program in 9th grade and are assigned a "monitor" (usually a special education teacher) who works with them year-round as a mentor, adviser, and service coordinator. Monitors carry an average caseload of approximately 35 students, regularly track each student's behavior and academic performance, and convey a strong message to both students and parents about the importance of completing high school.			
Random assignment to 1 of 4 conditions →			
[A] Outcomes	*[B] Benefit to society + Outcomes*	*[C] Outputs*	*[D] Benefit to society + Outputs*
Check and Connect cost $350,200 and helped 34 students graduate high school who would not have graduated otherwise. This represents a cost of $10,300 per student graduating because of the program.	The total benefit to society associated with high school graduation is about $90,000 on average (from increased taxes and reduced social support and incarceration). Check and Connect cost $350,200 and helped 34 students graduate high school who would not have graduated otherwise. This represents a cost of $10,300 per student graduating because of the program.	Check and Connect cost $350,200 and served a total of 206 students. This represents a cost of $1,700 per student served by the program.	The total benefit to society associated with high school graduation is about $90,000 on average (from increased taxes and reduced social support and incarceration). Check and Connect cost $350,200 and served a total of 206 students. This represents a cost of $1,700 per student served by the program.
Dependent variable (3-items)			
Based on this information, do you agree or disagree that the Check and Connect program is (1) an effective program, (2) an efficient program, (3) a good investment of tax dollars. (Strongly disagree=1, to Strongly agree=5)			

Figure 2. Check and Connect Experimental Vignette (Study 1).

at-risk students in 9th grade and provides them with a monitor (usually a special education teacher) to track their behavior. In addition to testing whether our original HTPP health program findings would be replicated with an educational program, we selected Check and Connect because its target population (high school students) are presumably less stigmatized than people at risk of HIV/AIDS (Schneider & Ingram, 1993). The vignette we constructed was based on information from an RCT evaluation of Check and Connect published by Sinclair et al. (2005) and also summarized by The Arnold Ventures' website, Social Programs That Work (2017). We used cost information from an earlier version of the information on the Social Programs That Work website (at the time of our research design), which was $1,700 per student (a figure subsequently updated by Social Programs That Work). Our estimates of the benefit to society of someone completing high school comes from a separate RAND Corporation research brief (Carroll & Erkut, 2009), which conservatively estimated a social benefit to society of about $90,000 (rounding up slightly from $89,000 to simplify the number and to adjust for the somewhat older date of the RAND brief) from increased tax collection and reduced social services and incarceration.

By using these various sources of information on Check and Connect, and following the basic structure of the HTPP vignette, we constructed the experimental vignette for Check and Connect shown in Figure 2. As before, respondents were randomized to one of four arms in the experiment, which represent a 2×2 factorial variation of information. Arms A and B represent the *outcome* scenarios, as they mention that the program "helped 34 students graduate high school who would not have graduated otherwise." This wording attempted to make the notion of a counterfactual and resulting causal effect somewhat more explicitly stated than in the

original HTPP vignette. Arms C and D represent the *output* scenarios, as they mention only that Check and Connect, "served a total of 206 students." Again, these are the real numbers from the published evaluation. Arms B and C provide additional information on the lifetime benefit to society of a person completing high school ($90,000) from the RAND brief.

In both the HTPP and Check and Connect experiments, after viewing the vignette, respondents were asked to indicate the extent to which they agreed or disagreed that the program is: (1) an effective program, (2) an efficient program, and (3) a good investment of tax dollars. The 1–5 response scale (from 1 = strongly agree, to 5 = strongly disagree) was reverse coded for purposes of analysis (from 1 = strongly disagree, to 5 = strongly agree) so that higher scores correspond to more favorable judgments of the program. Responses to these agree-disagree items constitute the dependent variables in both experimental vignettes. For the main analyses, we use a summated, standardized scale of the three items (which are highly correlated) to simplify the presentation of graphs and tables. Additional analyses of each item separately are provided in the Appendix.

Study 1: Data and results

Data for Study 1, which included both the HTPP and Check and Connect vignettes, came from an online sample of $n = 840$ U.S. adults obtained through invitations sent to the Qualtrics research panel in February, 2018. (The study was approved by the Rutgers University Arts and Sciences IRB, Study ID: Pro2018000090, January 24, 2018.) Quotas were established for region, sex, age, and race based on national estimates from the American Community Survey. Thus, although not a probability sample of the U.S. adult population, the sample matches the U.S. adult population on these key characteristics. It should be noted that the HTPP vignette and Check and Connect vignette were two of 10 short survey experiments embedded in the same online questionnaire, with the order of experiments randomized for each respondent. Data were analyzed (unweighted) with Stata 16.

Figures 3 and 4 show the results of the two experiments, and Table 1 presents the corresponding two-way ANOVAs. The dependent variable in these graphs and analyses is a standardized scale of the three agree-disagree items (an effective program, an efficient program, and a good investment of tax dollars), with reliability $\alpha = 0.91$ in the HTPP experiment and $\alpha = 0.89$ in the Check and Connect experiment. Beginning with the HTPP vignette, as Figure 3 shows, the societal benefit information (the $250,000 lifetime cost of treating a person with AIDS) results in significantly more favorable judgments about the program ($F = 13.38$, $p < 0.01$, from Table 1).

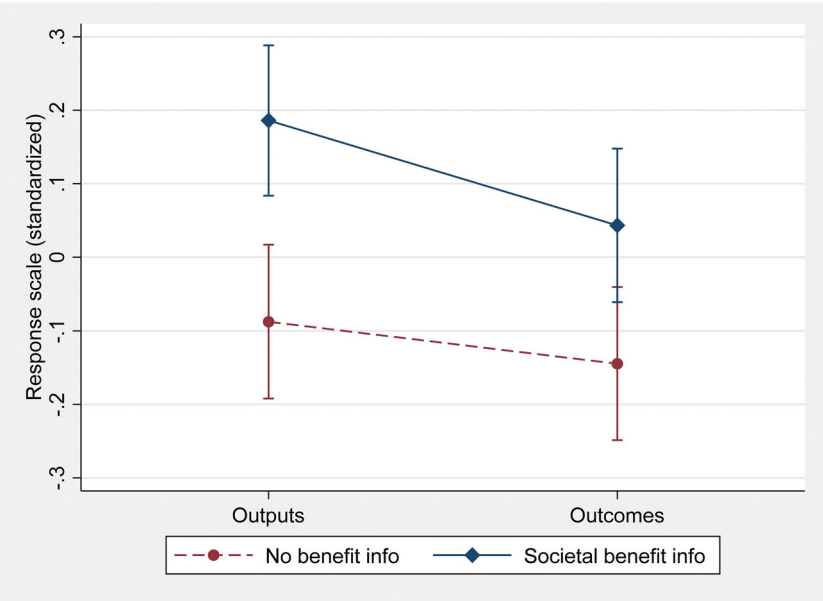

Figure 3. Mean responses for HTPP vignette (Study 1).

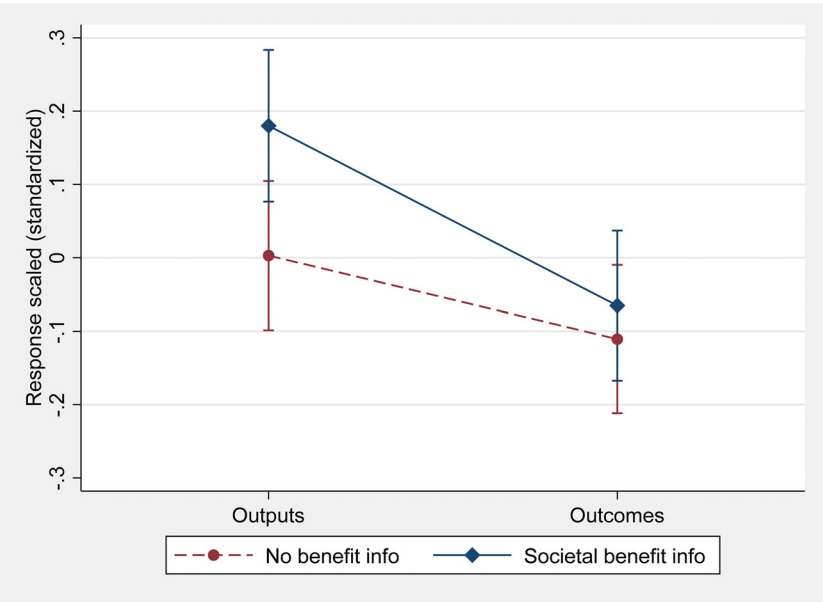

Figure 4. Mean responses for Check and Connect vignette (Study 1).

Similar to the findings from the original study by Grosso et al. (2017), the presentation of outputs leads to somewhat more favorable views of the program than the presentation of outcomes, although the main output-outcome difference is not quite significant statistically ($F = 2.51$, $p = 0.11$,

Table 1. Two-way ANOVA (Study 1).

	HTPP vignette					Check and Connect vignette				
Source	Partial SS	*df*	*MS*	*F*	Prob>*F*	Partial SS	*df*	*MS*	*F*	Prob>*F*
Model	13.82	3	4.61	5.52	0.00	10.12	3	3.37	4.19	0.01
Outcome	2.10	1	2.10	2.51	0.11	6.74	1	6.74	8.38	0.00
Benefit	11.16	1	11.16	13.38	0.00	2.59	1	2.59	3.22	0.07
Outcome × Benefit	0.38	1	0.38	0.46	0.50	0.91	1	0.91	1.13	0.29
Residual	696.60	835	0.83			670.71	834	0.80		
Total	710.43	838	0.85			680.83	837	0.81		
Observations (*n*)	839					838				
R^2	0.02					0.01				

two-tailed test, from Table 1). An analysis of agreement with just the item "a good investment of tax dollars" (see Appendix), however, reveals a significant main effect for outputs over outcomes ($F = 4.73$, $p = 0.03$). As in the original study, there is no significant interaction effect ($F = 0.46$, $p = 0.50$), although it does appear that the output bias is somewhat stronger when the societal benefit information is provided.

Figure 4 presents the results of the Check and Connect experiment. Again, providing the societal benefit information (the $90,000 lifetime benefit to society of someone completing high school) leads to somewhat more favorable judgments of the program ($F = 3.22$, $p = 0.07$, from Table 1), although the effect is not as large as in the HTPP experiment. However, the presentation of outputs, compared to outcomes, leads to much more favorable judgments of the Check and Connect program ($F = 8.38$, $p < 0.01$, from Table 1). This is a much stronger output bias than in the HTPP experiment—despite the fact that we phrased the Check and Check outcome arms in even more explicit counterfactual terms to help with understanding. An analysis of the individual scale items (see Appendix) suggests that the strongest output bias is again from agreement with the item "a good investment of tax dollars" ($F = 17.30$, $p < 0.01$). Although it is not significant statistically, there is the appearance of an interaction effect in that the output over outcome difference is somewhat more pronounced when societal benefit information is provided. This parallels the finding from the HTPP experiment and suggests that, even when given specific information on the benefit to society of an outcome, citizens may still judge outputs more favorably than outcomes.

Study 2: Design and method

We decided to run a second study to probe the extent to which rewording and reframing the output and outcome information in our vignettes might mitigate the observed output bias. In our first study, we focused mainly on directly replicating the vignette from Grosso et al. (2017) for the HTPP

N=1105 US adults					
Check and Connect is a government-funded dropout prevention program for high school students with learning, emotional, or behavioral disabilities. Students typically enter the program in 9th grade and are assigned a "monitor" (usually a special education teacher) who works with them year-round as a mentor, adviser, and service coordinator. Monitors carry an average caseload of approximately 35 students, regularly track each student's behavior and academic performance, and convey a strong message to both students and parents about the importance of completing high school.					
Random assignment to 1 of 6 conditions →					
[A] Outcomes	*[B] Benefit to society + Outcomes*	*[C] Outputs*	*[D] Benefit to society + Outputs*	*[E] Outcomes + Outputs*	*[F] Benefit to society + Outputs + Outcomes*
Check and Connect cost $305,200 and, according to a rigorous evaluation, caused 34 students to graduate high school who would not have graduated otherwise.	The total benefit to society associated with high school graduation is about $90,000 on average (from increased taxes and reduced social support and incarceration). Check and Connect cost $305,200 and, according to a rigorous evaluation, caused 34 students to graduate high school who would not have graduated otherwise.	Check and Connect cost $350,200 and served a total of 206 students.	The total benefit to society associated with high school graduation is about $90,000 on average (from increased taxes and reduced social support and incarceration). Check and Connect cost $350,200 and served a total of 206 students.	Check and Connect cost $350,200 and served a total of 206 students. According to a rigorous evaluation, the program caused 34 students to graduate high school who would not have graduated otherwise.	The total benefit to society associated with high school graduation is about $90,000 on average (from increased taxes and reduced social support and incarceration). Check and Connect cost $350,200 and served a total of 206 students. According to a rigorous evaluation, the program caused 34 students to graduate high school who would not have graduated otherwise.
Dependent variable (3-items)					
Based on this information, do you agree or disagree that the Check and Connect program is (1) an effective program, (2) an efficient program, (3) a good investment of tax dollars. (Strongly disagree=1, to Strongly agree=5)					

Figure 5. Check and Connect Experimental Vignette (Study 2).

program and extending it to the Check and Connect program, which as noted above showed an even stronger output bias. However, there were aspects of the framing of outputs, outcomes, and costs in these vignettes that may have encouraged a more negative judgment of outcomes: namely, the vignettes provide respondents with the per-unit cost of an output or an outcome. In most programs, as mentioned earlier with the example of the flu vaccine, the per-unit cost of an outcome is many times more expensive than for an output. Thus, in our second study, we wanted to test judgments of Check and Connect without the explicit unit-cost of an output or outcome shown to respondents. We also tried to strengthen even more the wording of an outcome as a rigorous causal effect. Moreover, in our first study, we did not test a version of the vignette in which *both* output and outcome information were provided jointly. For our second study, therefore, the vignette has six arms (A–F) representing a 2×3 factorial variation of information, as shown in Figure 5.

As can be seen from Figure 5, the introductory wording is the same as before, but each arm of the vignette provides only the total cost of the program ($350,200) and then the number of students served (206) or students graduating because of the program (34). It is possible, of course, for respondents to calculate or at least estimate the unit-cost of an output or outcome, but again this information was not explicitly given to them (as it was in Study 1). Moreover, the causal language for the outcome versions of the vignette (arms A, B, E, and F) was strengthened as follows: *According to a rigorous evaluation, the program caused 34 students to graduate high school who would not have graduated otherwise.* Thus, we added "according to a rigorous evaluation" and explicitly stated that the program "caused" the students to graduate (rather than just "helped" them to graduate, as in

Study 1). The aim of these modifications was to enhance the persuasiveness of the outcome information. Arms E and F of the vignette provide *both* output information (206 students served) *and* outcome information (34 students graduating high school because of the program). We did not have a specific expectation of the effects of these combined output + outcome vignettes, which were included more for exploratory reasons. In general, we wanted to examine how providing a more complete picture of the program, with both output and outcome information, would shape respondents' evaluative judgements.

Study 2: Data and results

Data for Study 2 came from an online sample of $n = 1105$ U.S. adults, with responses obtained through invitations sent to the Qualtrics research panel in July 2019. (The study was approved by the Rutgers University Arts and Sciences IRB, Study ID: 2019000960, May 14, 2019.) As before, quotas were established for region, sex, age, and race based on national estimates from the American Community Survey. The Check and Connect vignette for Study 2 was the second of three experiments embedded in the same online questionnaire. Data were again analyzed (unweighted) with Stata 16.

Figure 6 shows the results of the Study 2 Check and Connect vignette experiment, and Table 2 presents the corresponding two-way ANOVA. The dependent variable in the graph and ANOVA is the same standardized scale of three agree-disagree items used in Study 1 (an effective program,

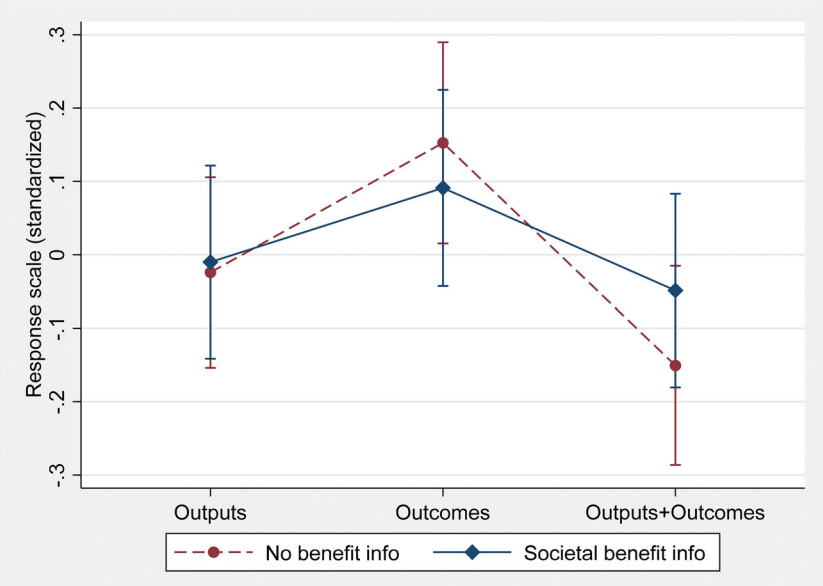

Figure 6. Mean responses for Check and Connect vignette (Study 2).

Table 2. Two-way ANOVA (Study 2).

	Check and Connect vignette				
Source	Partial SS	*df*	*MS*	*F*	Prob>*F*
Model	9.90	5	1.98	2.39	0.04
Outcome	8.79	2	4.39	5.32	0.01
Benefit	0.09	1	0.09	0.11	0.74
Outcome × Benefit	1.17	2	0.59	0.71	0.49
Residual	884.38	1070	0.83		
Total	894.28	1075	0.81		
Observations (*n*)	1076				
R^2	0.01				

an efficient program, and a good investment of tax dollars), with reliability $\alpha = 0.90$. As can be seen in Figure 6, providing information on the societal benefit of high school graduation does little to influence judgments of the program, in contrast to Study 1. This suggests that the per-unit cost information in Study 1, as might be expected, facilitated a more direct benefit-cost comparison on the part of respondents. In any event, as Table 2 shows, there is no main effect of the social benefit information in Study 2. There is, however, a significant main effect of the factor representing outcomes versus outputs ($F = 5.32$, $p = 0.01$). As Figure 6 shows, the presentation of rigorous outcomes (causal effects) now leads to a *more* favorable judgment of the Check and Connect program, the reverse of Study 1. This suggests that removing the per-unit cost information and strengthening the language about causal evidence in Study 2 helped overcome the output bias evident in Study 1. Curiously, however, presenting the full picture of both outputs and outcomes results in the least favorable evaluation of the Check and Connect program. We will suggest some possible interpretations of this finding, which indicates a possible output bias, in the next section.

Discussion

Our experiments replicate and extend the findings of Grosso et al. (2017), which suggested that citizens tend to judge a social program more favorably when given information about mere outputs (such as clients served) contrasted with more meaningful outcomes (causal effects). We found this *output bias,* as we call it, in a fairly close replication of the original study by using a vignette based on California's HTPP, although the effect was small and not as statistically significant as in the original study (even though our sample size, $n = 839$, was slightly larger than in the original study, $n = 774$). We extended the paradigm to another evidence-based program for special needs high school students, Check and Connect, and found the same pattern with an even stronger and more statistically significant difference in people favoring outputs over outcomes. In Study 2 with

a new sample, however, we were able to reverse the output bias by using a modified version of the Check and Connect vignette that dropped the explicit mention of the dollar-cost of an output or outcome and strengthened the statement about rigorous causal evidence of an outcome. With these modifications, respondents presented with outcomes (causal effects) judged the program more favorably than those provided only with mere outputs (students served).

Why would people judging this kind of evidence generally tend to value ambiguous outputs compared to more meaningful outcomes, as we found in Study 1? As we suggested earlier, people may have difficulty with counterfactual thinking; they may engage in substitution by interpreting outputs as if they were outcomes; or the larger number of outputs relative to outcomes might trigger a more-is-better heuristic. However, as Study 2 suggests, the most likely explanation can be found in the high per-unit cost of an outcome compared to an output, resulting in a kind of outcomes sticker shock that seems to influence judgments of the program. In the case of Check and Connect, the program cost $10,300 per student graduating because of the program (outcome) but only $1,700 per student served by the program (output). When this explicit unit-cost information was dropped from the vignette in Study 2, respondents judged the program more favorably given (strengthened) outcome information rather the mere output information. This interpretation is reinforced by the analysis of individual scale items (see Appendix) that shows the largest effects in Study 1 come from the scale item related most directly to cost (a good investment of tax dollars). Thus, it appears that being too transparent about the costs of producing an outcome may have detrimental effects on how people judge a social program.

On the other hand, interestingly, providing per-unit cost information, as in Study 1, still seems to have the advantage of facilitating people's ability to perform intuitive benefit-cost calculations. Indeed, there were consistent positive effects on judgments of the program in Study 1 when information on societal benefits was given. However, in Study 2, which did not provide the unit-cost of an outcome or output, judgments of the program were not enhanced at all by telling respondents about the societal benefit or payback of producing an outcome (high school graduation). We speculate that this may be due to the added difficulty of performing an intuitive benefit-cost calculation given the lack of explicit unit-cost information in Study 2 (although this information could be calculated or estimated with some basic mental math). Thus, while avoiding unit-cost information seems to place outcomes in a more favorable light, it also appears to limit people's ability to judge the societal benefit of a program relative to its costs.

A curious finding from Study 2 was that providing a more complete picture of Check and Connect, by giving *both* output *and* outcome information together, resulted in the least positive judgments of the program. This finding was unexpected but might be explained by several of the prior hypothesized mechanisms behind an output bias more generally. In particular, presenting both outputs and outcomes may highlight the relatively small number of outcomes produced by Check and Connect, namely, only 34 students graduating because of the program despite 206 students served. Counterfactual thinking is required to recognize that some share of the 206 students would graduate anyway, without Check and Connect, so that 34 of 206 is not a gross graduation rate but rather a net increase in the number of graduations causally attributable to the program. However, this counterfactual way of thinking about these numbers does not come easily, as discussed previously. In any event, these results suggest that the fuller presentation of both outputs and outcomes does not seem to lead to a more favorable judgment of even a highly effective social program like Check and Connect.

Our study clearly has some limitations. To begin, our vignettes involved brief descriptions of relatively unknown social program in the context of an online survey. Thus, we do not know how people would respond to more in-depth information on outputs, outcomes, and costs or to information regarding programs that they know better or perhaps benefit from more directly. Another limitation is that our sample, although nationwide in scope and balanced with the U.S. population (in terms of region, age, race, and gender), remains a voluntary sample and not a true probability sample. Moreover, we do not know from our study how a population of public managers or other professionals, with more substantive policy experience and practical knowledge, might similarly judge this kind of output, outcome, and cost information. In addition, while our experiments focused on health and education, it remains unknown how our results would generalize to other policy areas.

Despite these limitations, our findings still have some important implications for theory as well as the direction of future research. As noted earlier, our findings respond to Olsen's (2015) call for more work in the field of performance measurement on the psychology of numbers and to a growing body of research on behavioral public performance (James et al., 2020). As regards future research along these lines, similar experiments should be done with public managers or policymakers to test the extent to which they too may be susceptible to an output bias when interpreting evidence of this kind. Relatedly, future studies could examine whether individuals with higher numeracy may be less susceptible to rating outputs more favorably than outcomes, echoing some of the recent results by Baekgaard and

Serritzlew (2020). Future research could also examine how political engagement moderates the effects of outcome, output, and cost information on citizens' perceptions, as evidence suggests political engagement influences responses to varied presentations of performance information (Piotrowski et al., 2019). Because graphically displaying metrics and statistics seems to result in higher rates of intention to use performance information (Ballard, 2020), additional research could also be conducted to assess whether an output bias remains when presenting output and outcome information in a visual format rather than as text (as in our experiments). Future studies might also probe the effects of output and outcome information in discrete choice experiments, which Bellé and Cantarelli (2018) argue have been overlooked as a method for public management research.

Finally, we believe our results have potentially important implications for evidence-based policy and management, especially to the extent that such evidence is meant to enhance public support and democratic accountability. The finding of a tendency toward an output bias in the public's judgments about social programs suggests that rigorous evidence of causal effects (such as provided by RCTs) is not necessarily self-explanatory and may need careful framing and additional explanation to enhance public understanding. Specifically, our findings suggest that the structural features of outcomes—namely, that they are less frequent and more expensive than outputs—tend to lead to a systematic bias in people's judgments about government performance. The resulting output bias could potentially extend to public managers and policymakers, not just citizens—although future research, as noted earlier, would be needed to confirm this speculation. However, as Study 2 suggests, there seem to be ways to frame outcome information in more persuasive terms. Still, it may be that we all experience at least some difficulty with the task of interpreting evidence about the outputs, outcomes, and costs of social programs. Understanding and recognizing the possibility of an output bias represents an important first step in learning how better to frame or explain performance metrics and evidence in ways that encourage greater appreciation of rigorous causal evidence by the public and perhaps other key audiences as well.

Notes on contributors

Gregg G. Van Ryzin is professor in the School of Public Affairs and Administration, Rutgers University-Newark (USA). His work employs experimental and behavioral approaches to various issues in public management, including citizen satisfaction, coproduction, performance measurement, representative bureaucracy, and organizational behavior. He is author (with Dahlia Remler) of *Research Methods in Practice* (SAGE) and editor

(with Oliver James and Sebastian Jilke) of *Experiments in Public Management Research* (Cambridge).

Ashley Grosso is assistant professor in the Department of Urban-Global Public Health, School of Public Health, Rutgers University (USA) and a member in residence at the Rutgers Institute for Health, Health Care Policy and Aging Research. Her research is focused on social and structural determinants of health among people living with or at risk for HIV infection. She holds a Ph.D. from the Rutgers School of Public Affairs and Administration.

Etienne Charbonneau is Canada Research Chair in Comparative Public Management at École Nationale d'Administration Publique, Montreal (Canada). His current research focuses on public management, accountability and surveillance. He holds a Ph.D. from the Rutgers School of Public Affairs and Administration.

References

Baekgaard, M., & Serritzlew, S. (2020). Those who understand it will not be persuaded: A performance information paradox. *International Public Management Journal*, *23*(1), 138–160. https://doi.org/10.1080/10967494.2018.1461152

Ballard, A. (2020). Promoting performance information use through data visualization: Evidence from an experiment. *Public Performance & Management Review*, *43*(1), 109–128.

Baron, J. (2000). *Thinking and deciding*. Cambridge, UK: Cambridge University Press.

Bellé, N., & Cantarelli, P. (2018). Randomized experiments and reality of public and non-profit organizations: Understanding and bridging the gap. *Review of Public Personnel Administration*, *38*(4), 494–511.

Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytic approaches*. Russell Sage Foundation.

Boyne, G. A., Meier, K. J., O'Toole, Jr, L. J., & Walker, R. M. (Eds.). (2006). *Public service performance: Perspectives on measurement and management*. Cambridge, UK:Cambridge University Press.

California Department of Health Services (2006). *Economic evaluation of California's prevention case management intervention for HIV-positive and HIV-negative persons: The HIV Transmission Prevention Project (HTPP)*. California Department of Health Services.

Carroll, S. J., & Erkut, E. (2009). *How taxpayers benefit when students attain higher levels of education*. RAND Corporation.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Commission on Evidence-Based Policymaking. (2017). *The promise of evidence-based policymaking*. Washington, DC.

Crane, J. (Ed.). (1998). *Social programs that work*. Russell Sage Foundation.

Davies, H. T., & Nutley, S. M. (Eds.). (2000). *What works: Evidence-based policy and practice in public services*. Policy Press.

Doleac, J. L. (2019). "Evidence-based policy" should reflect a hierarchy of evidence. *Journal of Policy Analysis and Management*, *38*(2), 517–519. https://doi.org/10.1002/pam.22118

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Grosso, A., Charbonneau, É., & Van Ryzin, G. G. (2017). How citizens respond to outputs, outcomes, and costs: A survey experiment about an HIV/AIDS program. *International*

Public Management Journal, *20*(1), 160–181. https://doi.org/10.1080/10967494.2016.1143425

Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. Russell Sage Foundation.

Hatry, H. P. (2006). *Performance measurement: Getting results* (2nd ed.). Urban Institute Press.

Hubbard, G. (2014). The pursuit of evidence. In J. Nussle & P. Orszag (Eds.), *Moneyball for government*. Results for America.

James, O., Moynihan, D., Olsen, A. L., & Van Ryzin, G. G. (2020). *Behavioral public performance: How people make sense of government metrics*. Cambridge University Press.

Kettl, D. F. (2006). *The global public management revolution*. Brookings Institution Press.

Kim, M., Charles, C., & Pettijohn, S. L. (2017). Challenges in the use of performance data in management: Results of a national survey of human service nonprofit organisations. *Public Performance & Management Review*, *42*(5), 1085–1111.

Lee, C., & Clerkin, R. M. (2017). Exploring the use of outcome measures in human service nonprofits: Combining agency, institutional, and organizational capacity perspectives. *Public Performance & Management Review*, *40*(3), 601–624.

Morino, M. (2011). *Leap of reason: Managing to outcomes in an era of scarcity*. Venture Philanthropy Partners.

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

National Performance Management Advisory Commission. (2010). *A performance management framework for state and local government: From measurement and reporting to management and improving*. National Performance Management Advisory Commission.

Nussle, J., & Orszag, P. (Eds.). (2014). *Moneyball for government*. Results for America.

Olsen, A. L. (2015). The numerical psychology of performance information: Implications for citizens, managers, and policymakers. *Public Performance & Management Review*, *39*(1), 100–115.

Piotrowski, S., Grimmelikhuijsen, S., & Deat, F. (2019). Numbers over narratives? How government message strategies affect citizens' attitudes. *Public Performance & Management Review*, *42*(5), 1005–1028.

Roese, N. J., & Olson, J. M. (Eds.). (1995/2014). *What might have been: The social psychology of counterfactual thinking*. Psychology Press.

Rooks, G., Raub, W., Selten, R., & Tazelaar, F. (2000). How inter-firm co-operation depends on social embeddedness: A vignette study. *Acta Sociologica*, *43*(2), 123–137. https://doi.org/10.1177/000169930004300203

Schneider, A., & Ingram, H. (1993). Social construction of target populations: Implications for politics and policy. *American Political Science Review*, *87*(2), 334–347. https://doi.org/10.2307/2939044

Schueler, B. E., & West, M. R. (2016). Sticker shock: How information affects citizen support for public school funding. *Public Opinion Quarterly*, *80*(1), 90–113. https://doi.org/10.1093/poq/nfv047

Sinclair, M. F., Christenson, S. L., & Thurlow, M. L. (2005). Promoting school completion of urban secondary youth with emotional or behavioral disabilities. *Exceptional Children*, *71*(4), 465–482. https://doi.org/10.1177/001440290507100405

Social Programs That Work (2017). *Evidence summary for check and connect*. https://evidencebasedprograms.org/programs/check-and-connect/

Walker, R. M., Brewer, G. A., Lee, M. J., Petrovsky, N., & Van Witteloostuijn, A. (2019). Best practice recommendations for replicating experiments in public administration.

Journal of Public Administration Research and Theory, *29*(4), 609–626. https://doi.org/10.1093/jopart/muy047

Walker, R. M., James, O., & Brewer, G. A. (2017). Replication, experiments and knowledge in public management research. *Public Management Review*, *19*(9), 1221–1234. https://doi.org/10.1080/14719037.2017.1282003

Appendix

Separate ANOVAs for Scale Items (only *F*-tests shown).

	Effective program		Efficient program		Good investment of tax dollars	
Source	*F*	Prob>*F*	*F*	Prob>*F*	*F*	Prob>*F*
HTPP, Study 1						
Model	3.75	0.01	5.72	0.00	4.77	0.00
Outcome	0.59	0.44	1.85	0.17	4.73	0.03
Benefit	10.40	0.00	14.80	0.00	8.70	0.00
Outcome × **Benefit**	0.17	0.68	0.34	0.56	0.66	0.42
Observations (*n*)	839		838		838	
R^2	0.013		0.0202		0.0169	
Check and Connect, Study 1						
Model	1.40	0.24	3.06	0.03	7.77	0.00
Outcome	1.62	0.20	5.86	0.02	17.30	0.00
Benefit	0.60	0.44	2.91	0.09	5.65	0.02
Outcome × **Benefit**	2.04	0.15	0.50	0.48	0.54	0.46
Observations (*n*)	838		837		838	
R^2	0.005		0.0109		0.0272	
Check and Connect, Study 2						
Model	3.82	0.00	1.29	0.26	1.40	0.22
Outcome	7.65	0.00	2.99	0.05	3.17	0.04
Benefit	0.62	0.43	0.00	0.99	0.02	0.89
Outcome × **Benefit**	1.76	0.17	0.28	0.76	0.38	0.69
Observations (*n*)	1075		1072		1074	
R^2	0.013		0.006		0.002	