Journal of Behavioral
Public Administration
Vol 4(1), pp. 1-13
DOI: 10.30636/jbpa.41.149

Research Article

Compared to whom? Social and historical reference points and performance appraisals by managers, students, and the general public

Amanda Rutherford*, Thomas Rabovsky*, Megan Darnley*

Abstract: Experimental studies in public administration often focus on samples of non-practitioner groups. In these cases, it is unclear whether findings from non-practitioner groups are generalizable to public managers. Some literature suggests that bureaucrats are likely to hold biases similar to the rest of the population while other research argues that bureaucratic expertise and training allow practitioners to make decisions in more strategic or rational ways. This study works within the literature of performance information to test for differences in responses to the same experiment among college students, citizens, and public managers in the context of U.S. K-12 education. Some differences were detected across groups, though results reveal largely similar findings which have implications for when and how scholars might rely on non-practitioner samples to consider the attitudes and behaviors of bureaucrats or elected policymakers.

Keywords: Sample comparisons, Survey experiment, Information framing, Perceived performance

Supplements: Open data

The growth of empirical studies that use some type of experimental methodology in public administration has been charted by multiple scholars. These experiments often study questions related to individual-level perceptions and behaviors of citizens or, to a lesser extent, policymakers. For example, in reviewing fourteen public administration journals from 1992-2014, Bouwman and Grimmelikhuijsen (2016) assessed 42 articles that included experiments and found survey experiments with between-subject designs to be most common. The average sample size among these studies was 471, with student samples or samples of the general public being most common. Similarly, Li and Van Ryzin (2017) considered 72 articles with randomized, quasi-, and natural experiments across twenty journals between 1957 and 2016. This review found that most experiments were published in the last decade and almost half (33 of 72) focused primarily on survey experiments. Ninety percent of the studies were conducted using students while six experiments focused on public managers.

These assessments of the literature illustrate that few experimental studies have been conducted on public sector workers (Roberts & Wenstedt, 2019; but see also Andersen, 2017 and Bellé, 2013a). This is due to a variety of possible reasons such as accessibility and cost. However, it is also the case that whether the attitudes and perceptions of practitioners reflect those of non-practitioner groups remains unclear. As such, the purpose of this study is to test whether public managers are subject to some of the same biases and heuristics as the rest of the population; in this case, we focus on biases stemming from information framing and performance gaps. On one hand, it may be the case that biases and heuristics are observed consistently across groups, as all people are required to make decisions and possess the same type of information processing system (e.g., the human

* O'Neill School of Public and Environmental Affairs Indiana University
Address correspondence to Amanda Rutherford at (aruther@indiana.edu)

mind) (Hilbert, 2012). In other words, there is reason to believe that managers and decision makers are not exempt from a range of biases (Schwenk, 1984; Das & Teng, 1999). However, there are also reasons to expect that public managers will respond differently to performance information, including feedback provided in a survey experiment, particularly in light of a wealth of research on bureaucratic socialization processes and professional norms (Maanen & Schein, 1977; Oberfield, 2014; Druckman & McDermott, 2008). From teachers to police officers to social workers, socialization is viewed as at least a mechanism for instituting shared customs and standards and often as a vital component for building a responsive and accountable workforce (Weber,1947).

The Use of Public Administrators in Public Administration Survey Experiments

Though not without limitations, there often exist legitimate reasons to involve students or citizens in experiments (Anderson & Edwards, 2015; Bouwman & Grimmelikhuijsen, 2016; Druckman & Kam, 2009). Such tradeoffs underscore the importance of theoretically considering whether public managers cognitively process information—in the case of current behavioral research, those provided through experimental manipulations—in a manner similar to or different from other groups. If public sector employees respond to experimental conditions in the same way as other groups, this would imply that non-practitioner samples can be used to construct valuable insights about practitioners. Further, scholars could more confidently assume that the ways in which one group responds to particular experimental conditions will be similar to other groups given common human cognition processes. This means scholars would have a priori grounds for claiming that the phenomenon they are interested in understanding is universal and can be studied through a variety of sample populations. In terms of theoretical implications, it also implies that public administration questions being studied through surveys (or perhaps other experimental designs) are those that tap general human traits, which mirrors arguments used regularly in the field of psychology for experiments focused on fairness, cooperation, reasoning, and induction (Moynihan, 2018; though see Heinrich, Maffioli, & Vazquez, 2010). Finally, parallel responses could indicate that educational training, socialization into a profession, and expertise in a policy area do not outweigh the biases and shortcuts to which individuals resort when processing information quickly, at least in the context of an experiment (e.g., Kanwisher, 1989; Rhee, Ryu, & Kim, 2012).

On the other hand, should an experiment conducted across multiple groups determine public sector managers have responses that differ in some meaningful way from other groups, then experiments conducted with citizens, students, policymakers, or other groups are less able to generalize to how public managers would interpret and respond to similar scenarios. More importantly, scholars would need to develop more detailed explanations for why public managers think and act the way they do given characteristics such as policy expertise, public service motivation, or agency culture. Scholarly work would need to determine circumstances in which common cognitive biases no longer hold, the conditions under which findings from non-practitioner samples can be generalized, and strategies for successfully fielding experiments with practitioner groups.

Existing studies centered on public managers as the population of interest provide valuable information to extend both theoretical knowledge and normative prescriptions in the field—the successes and failures of how performance systems are designed, what measures of performance public managers prefer, and whether changing accountability mechanisms influence managerial behavior. For example, Andersen and Jakobsen (2017) find that communication frames that align with the professional norms of bureaucrats will generate policy support while tapping dimensions outside of such norms could produce a deleterious effect. Bellé et al. (2017) run two experimental surveys with Italian bureaucrats to show the extent to which anchoring and halo effects might bias performance ratings. Importantly, more recent work by Liu, Stoutenborough, and Vedlitz (2017), Roberts and Wernstedt (2019), and Funezalida, Van Ryzin, and Olsen (2020) specifically consider whether public managers are subject to cognitive biases that are generally expected among non-practitioners. This work finds evidence of overconfidence stemming from perceived expertise among climate change officials, prospect theory and attribution bias among emergency managers, and equivalence framing effects among varied U.S. public service professionals, respectively. These studies reveal important information about the ways in which practitioners process information, but none explicitly compare practitioners and non-practitioners to empirically document what similarities and differences exist across groups. Instead, research has largely assumed that biases detected in one group can apply similarly to another such that whether the same conclusions would

have been reached had these surveys been conducted with populations internal and external to the bureaucracy remains an open question.

Testing for Differences through the Framing of Performance Information

In assessing whether public sector managers will respond to experiments in a manner similar to other groups, we focus on discussions of information framing, where experiments have been particularly popular. A considerable amount of literature has established that perceptions and attitudes influence employee and managerial decision making, work motivation, and other types of individual-level behavior in organizations (Bellé, 2013b; Capelo & Dias, 2009; Moynihan, 2006; Nielsen, 2014; Wise, 2004), though it is less clear whether different groups respond to information in similar or different ways. Instead, experiments involving historical and social comparison have primarily included either practitioner or non-practitioner populations to answer particular questions about perceptions of and responses to performance information (e.g., Barrows, Henderson, Peterson, & West, 2016; Hansen, Olsen, & Bech, 2015; Nicholson-Crotty, Nicholson-Crotty, & Webeck, S., 2019).

When evaluating how practitioner and non-practitioner groups compare, prior literature suggests at least two areas of commonality—e.g., opportunities for shared cognitive biases—in responses to performance information. First, negativity bias, or the notion that negative information will have a stronger effect on perceptions and decisions than positive information, has been detected in numerous studies on single populations (Nicholson-Crotty, Grissom, Nicholson-Crotty, & Redding, 2017; Nielsen & Moynihan, 2017; Salage, 2010). Given general claims that bureaucrats are susceptible to biases of their own (e.g., Mercer, 2005; Roberts & Wernstedt, 2019) and Fuenzalida et al.'s (2020) specific finding that negative information incites less favorable evaluations among public service professionals than logically equivalent positive information, it is likely that non-practitioners and practitioners alike will respond more strongly to negative rather than positive information. We expect this to be true whether performance information highlights historical data about the past performance of a public agency or focuses on how the institution compares to similar peer agencies (see additional discussion of peer groups and social or historical aspirations in Bendor, 2010; Cyert & March, 1963; Greve, 1998). This leads to the following hypothesis:

H1: Negativity bias will be detectable across practitioner and non-practitioner groups when respondents are presented with positive vs. negative experimental performance information.

Second, we expect that information related to peer performance will have a stronger effect on perceptions than information about historical performance across practitioner and non-practitioner groups. Pressure related to performance rating systems and rankings have made peer comparisons salient, and many government agencies are rewarded or punished based on performance relative to peers. In comparing the two types of performance data, Olsen (2017) found that social comparisons are twice as important for shaping citizen perceptions of performance, though both types of information mattered (see also Barrows et al., 2016; Charbonneau, Bromberg, & Henderson, 2015). For bureaucrats, Zhu and Rutherford (2019) find somewhat consistent evidence that gaps in performance with peers generates stronger reactions among managers via stated goals and preferences than historical gaps in the context of U.S. hospitals (see also Holm, 2018). While additional (more routine) exposure to performance information may temper the extent to which cognitive biases are observed among practitioners (e.g., Druckman & McDermitt, 2008), such expertise does not eliminate the possibilities of heuristics such as overconfidence or risk taking via prospect theory. Our second hypothesis can be stated as:

H2: Peer performance information will have a stronger effect than historical performance across practitioner and non-practitioner groups in an experimental setting.

It is also the case that the substantive effect size of performance information on perceptions could be quite different. As in the context of education used in this study, the nuances of performance are often understood differently by school administrators, parents and students, and the general population (Donegan & Trepanier-Street, 1998). This expected divergence is consistent with previous work on the influence of expertise in decision making processes. Nicholson-Crotty and Miller (2012), for example, highlight how bureaucratic expertise

influence the policy process, and additional work (e.g., Carpenter, 2001; Krause, 1996) suggests that bureaucratic experience and knowledge influence decision making in public agencies. More recently, Fuenzalida, Van Ryzin, and Olsen (2020) found that, while framing effects were present among public service professionals, they were about half the size of the effects detected for citizens in Olsen (2015); the authors attribute differences to potential mechanisms of work experience and training. In other words, the effects of information framing, while present, may be more muted among practitioners who are accustomed to certain types of performance information and who make decisions informed by specialized knowledge relevant to their job on a more regular basis (Liu et al., 2017). Our final hypothesis, broken into two components, is stated as:

H3a: The substantive size of reactions to positive and negative performance information will vary across practitioner and non-practitioner groups in an experimental setting.
H3b: The substantive size of reactions to peer and historical comparisons in performance information will vary across practitioner and non-practitioner groups in an experimental setting.

Experimental Design and Data

Before scholars can understand why similarities or differences exist in the cognitive biases displayed by practitioners and non-practitioners, we must first determine what similarities and differences exist. Though this will ideally require a series of experiments on multiple types of bias as well as replication experiments in varying contexts for any particular bias or heuristic, the analysis below can help push the behavioral literature forward to address these questions. Following the expectations above, we designed and implemented a survey experiment to capture the effects of social and historical comparisons, positive and negative performance information, and differences in the magnitude of responses across three distinct groups.[1] The experiment fielded for this study was circulated between October and December of 2019 and administered through Qualtrics. The first survey was sent via email to 1,678 school principals in the state of Indiana. From this group, 321 participated, yielding a response rate of 19%. Second, the survey was administered to 497 students enrolled at a university located in the state of Indiana. A total of 348 students participated in the survey, generating a response rate of 70%. Finally, the survey was listed on MTurk for residents of Indiana who are 18 or older; 633 total individuals completed this survey.

Respondents in each group were asked to act as the principal of a hypothetical school in the state of Indiana (see specific language for control and treatment groups in the appendix). Participants received information related to performance on the Indiana Statewide Testing for Educational Progress (I-STEP), the standardized test administered by the state each year. The performance information indicated the percentage of eighth grade students who were at least proficient in English/Language Arts. Performance data listed in the experiment reflected the actual performance of eighth graders in 2018.

Participants were randomly assigned to one of five groups. Two scenarios focused on social comparisons of performance with other schools (the state average for student performance), two focused on historical comparisons of performance of the school from previous years to the current year, and a control scenario provided no comparison for the performance information given. Social and historical scenarios each included one instance of comparatively higher or lower performance. Participants were asked to rate the hypothetical school's performance on a 1-100 scale where higher values indicated better performance.

In the social comparison scenarios, respondents were given information regarding the I-STEP and an average score for the hypothetical school. Information was then provided about the average passing rate in the state of Indiana that was either 10.5 percentage points higher or lower than the hypothetical school. For example, in the case in which the participant was selected into the higher performance scenario, the hypothetical school performance indicated 74.88% of students were proficient while the state average was 64.38%. The historical scenarios followed a similar logic, though the scenario focused on past performance of the hypothetical school. For example, participants in the lower historical performance group were told that 2019 results revealed that 53.38% of eighth grade students were at least proficient in English/Language Arts while, in 2018, 63.88% of eighth grade students were at least proficient. For the social low scenario, the phrase "This places your school in the bottom half of eighth grade performance" was included in the principal sample and not the MTurk or student sample. While this wording contains no new information for respondents (who received

numerical information indicating that the school was below the mean), it is possible that it may have impacted some of the responses. We return to this limitation in the discussion section.

Shifts in performance were determined based on state performance data observed in 2018. The standard deviation for students who were at least proficient was just over 18 percentage points across schools in the state. Changes from 2017 to 2018 within schools was approximately 4 percentage points. The 10.5 percentage point mark falls between these two points. Providing this consistent magnitude of change aligns with prior literature (Charbonneau & Van Ryzin, 2015; Webeck & Nicholson-Crotty, 2019) and allows for a clearer comparison of groups that is not skewed by variance in the magnitude of changes (though see Olsen, 2017, for example, for a discussion of the use of randomized numbers).

Figure 1 shows the distributions and summary statistics for each condition. It should be noted that variance in the size of control and treatment groups is due to the fact that some respondents started the survey and were randomized into groups but did not complete the survey such that their responses could not be included in the analysis. All respondents were asked a series of demographics questions that serve as control variables for our multivariate analysis. These included age, race/ethnicity, gender, and political ideology according to a five point Likert scale (1=very liberal, 3=middle of the road, 5=very conservative). While the use of these control variables is not strictly necessary given our random assignment experimental design, we include them in our multivariate analysis to aid in comparing the three samples. Table 1 below provides descriptive statistics for each sample. We also conducted randomization checks for each of the three samples (see appendix Table A2), which show that the treatment conditions are balanced for all three populations.

Figure 1: Histograms and Summary Statistics

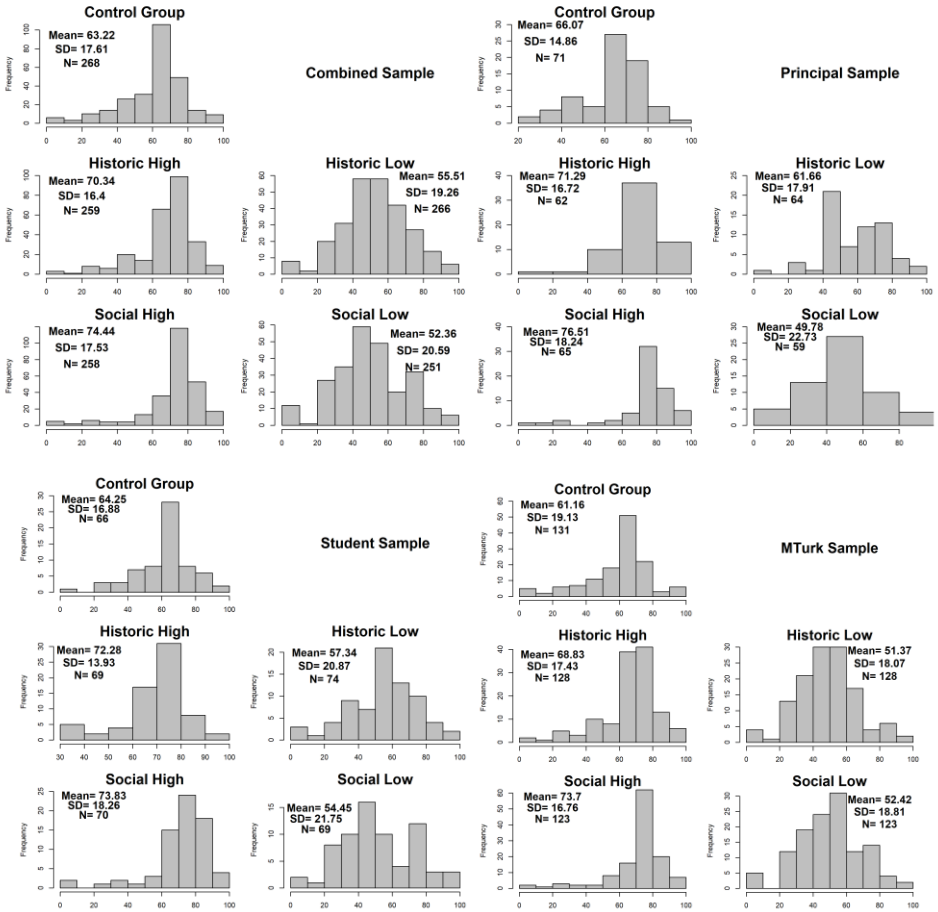


Table 1: Descriptive Statistics

Combined Samples

Variable	Mean	St. Dev.	Min	Max
Perceived Performance (DV)	63.19	20.12	0	100
Age	32.53	12.78	18	81
White	0.85	0.36	0	1
Male	0.46	0.50	0	1
Ideology (5=Very Conservative)	2.90	1.19	1	5

Principal Sample

Variable	Mean	St. Dev.	Min	Max
Perceived Performance (DV)	65.32	20.12	0	100
Age	46.50	7.61	31	66
White	0.93	0.26	0	1
Male	0.56	0.50	0	1
Ideology (5=Very Conservative)	3.12	1	1	5

MTurk Sample

Variable	Mean	St. Dev.	Min	Max
Perceived Performance (DV)	61.47	20.03	0	100
Age	34.80	10.83	18	81
White	0.85	0.36	0	1
Male	0.41	0.49	0	1
Ideology (5=Very Conservative)	2.87	1.26	1	5

Student Sample

Variable	Mean	St. Dev.	Min	Max
Perceived Performance (DV)	64.36	20.06	0	100
Age	19.55	1.74	18	39
White	0.80	0.40	0	1
Male	0.47	0.50	0	1
Ideology (5=Very Conservative)	2.79	1.17	1	5

Findings

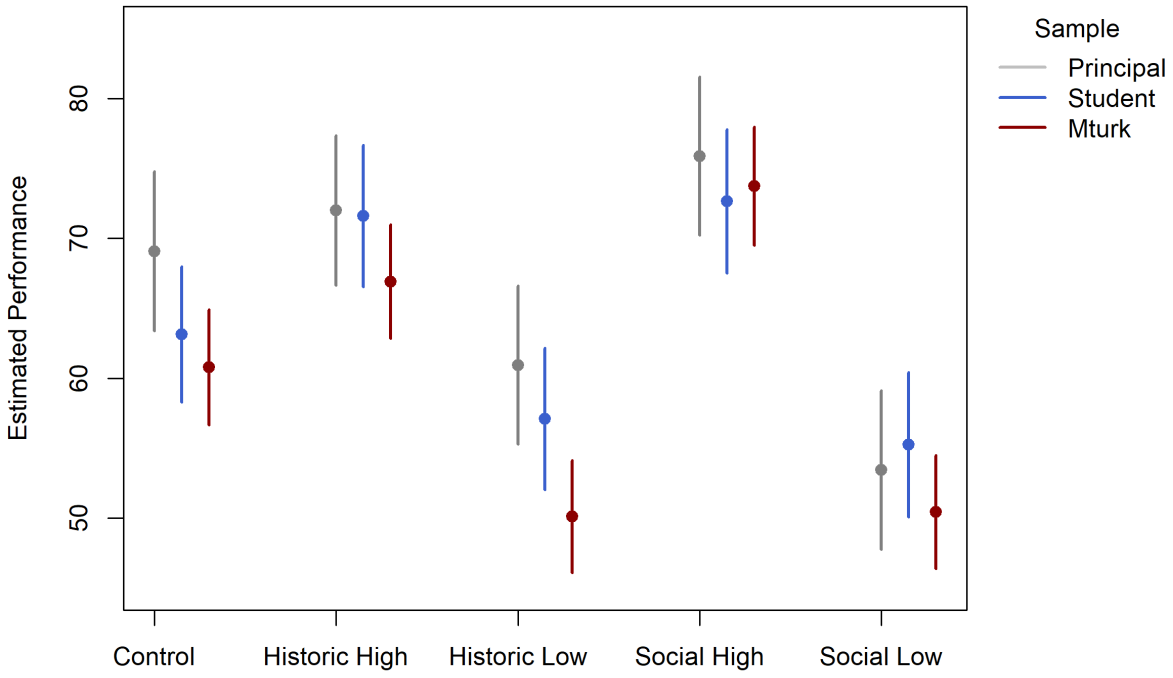
We use OLS regression to evaluate the effects of social and historical reference points and draw comparisons about these effects across the three samples. We also estimated two-way ANOVA both with and without an interaction. The primary advantage of OLS compared to ANOVA is that it allows us to evaluate the magnitude of effects in addition to statistical significance, which is important to assess potential asymmetric effects such as negativity bias. Because both the OLS and ANOVA results show the same substantive findings, we report the OLS models below and include the ANOVAs in the appendix (Tables A3-A5). Table 2 provides the results of OLS regressions while figure 2 graphically displays the point estimates and 95% confidence intervals for each group and treatment condition. Predicted values for this figure are generated from models 2-4 of Table 2 for each treatment condition, with all other variables set to mean or modal values.

Table 2: OLS Regression Results

	Combined Samples	School Principals	University Students	MTurk Respondents
	(1)	(2)	(3)	(4)
Historic high	**6.4*** **	2.9	**8.5** **	**6.1***
	(1.7)	(3.7)	**(3.2)**	**(2.5)**
Historic low	**-8.8*** **	**-8.1***	**-6.0+**	**-10.7*** **
	(1.7)	**(3.8)**	**(3.1)**	**(2.5)**
Social high	**10.7*** **	**6.8+**	**9.5** **	**13.0*** **
	(1.8)	**(3.8)**	**(3.2)**	**(2.5)**
Social low	**-10.5*** **	**-15.6*** **	**-7.9***	**-10.4*** **
	(1.8)	**(3.8)**	**(3.1)**	**(2.5)**
MTurk Sample	**-5.3** **			
	(1.7)			
Student Sample	-3.0			
	(2.4)			
White	-0.4	4.5	-3.3	0.5
	(1.6)	(4.5)	(2.5)	(2.3)
Male	-0.6	1.4	-0.5	-1.4
	(1.1)	(2.4)	(2.2)	(1.7)
Ideological Conservatism	**0.9+**	0.6	**1.9+**	0.4
	(0.5)	(1.2)	**(0.9)**	(0.6)
Age	-0.1	0.2	-0.4	**-0.2***
	(0.1)	(0.2)	(0.6)	**(0.1)**
Constant	**68.4*** **	**53.5*** **	**69.7*** **	**65.7*** **
	(3.9)	**(9.2)**	**(12.2)**	**(4.0)**
Observations	1,056	212	319	525
Adjusted R^2	0.181	0.163	0.139	0.198

Notes: Standard errors in parentheses + $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 01$

Figure 2 : Predicted Performance Assessments Across Samples and Treatment Groups



Overall, results are fairly similar across the samples, with positive effects for conditions indicating improved or above average performance and negative effects for conditions indicating poor or declining performance. In the combined sample (model 1 in Table 1), the control condition serves as the omitted category among randomized groups, and the principal sample serves as the omitted group to compare to the student and MTurk samples. MTurk respondents perceived performance to be, on average, 5.3 points lower than school principals, all else equal. The coefficient for the university student sample was also negative but not statistically significant. Among control variables, only political ideology is statistically significant such that conservativism is positively associated with performance, though the substantive effect is small (β=0.9). Age, which can serve as a proxy for experience among practitioners (Ng & Feldman, 2013) was not significant in either the combined model or the principal model.

To consider specific patterns for negativity bias and the weighting of social versus positive information, two approaches have been used in prior literature. The first occurs through comparing the absolute values of the coefficients reported in the models (e.g., James, 2011) while the second involves considering significance tests of the coefficients from the regression model (e.g., Nielsen & Moynihan, 2017).

Our first hypothesis is that negativity bias will be detected across all samples. When considering the absolute value of coefficients, we find evidence to support negativity bias in the cases of social and historical performance information among school principals as well as historical performance information in the case of MTurk respondents. For example, the MTurk historical high group has a coefficient of 6.1, while the historic low group has a coefficient of -10.7. Among school principals, the historic high group has a coefficient of 2.9, while the historic low group has a coefficient of -8.1, and the social high group has a coefficient of 6.8 while the social low group has a coefficient of -15.6. This suggests that poor performance, when framed as a comparison to other schools, has at least twice the effect that positive peer performance information has on perceptions of school performance for this particular group. For the student sample, no evidence of negativity bias

is detected as the absolute values for the high groups are actually larger than those for the low performance groups which does not support our first hypothesis.

When significance tests of the coefficients are considered, evidence of negativity bias is strongest for the historical scenarios in the school principal sample. Here, high historical performance is insignificant while low historical performance is negative and significant compared to the excluded control group category. For the MTurk and student samples, negative and positive information across historical and social comparisons is associated with statistically significant changes. Overall, less support is evident for negativity bias in this approach, and we are unable to clearly confirm support for hypothesis one. Interestingly, where negativity bias is most strongly detected suggests that principals might be more susceptible to the use of heuristics than non-practitioners. This points to the need for additional research to consider the ability of training and experience to taper the effects of at least some types of heuristics for practitioners.

When we turn to the strength of social or peer comparisons relative to historical comparisons, the absolute value approach provides mixed support for our second hypothesis. For example, MTurk respondents who were assigned to the historic high treatment condition rated school performance to be 6.1 points higher than the control group while those in the social high condition rated performance to be 13 points higher. This difference of 6.9 points between these treatments is statistically significant at $p<0.01$. On the other hand, the same is not true for low performance in the MTurk sample where the penalties for low performance are almost equivalent in size. When looking at the school principal sample, the negative impact of underperformance relative to other schools (-15.6) was larger than the penalty for declining performance over time (-8.1), and this difference of 7.5 points is statistically significant at $p<0.05$. For students, both positive and negative reactions were substantively similar across historical comparisons and across social comparisons, with no statistically significant differences between the two high or the two low performance treatments.

That said, in all cases except the comparison of high historic and social information in the case of principals, social and historical performance information maintain statistically significant coefficients. This limits the strength of support for the second hypotheses, though it also generally applies similarly across all groups which indicates one point of consistency across practitioners and non-practitioners. One explanation may be that the limited power of our samples prohibits greater distinctions between social and historical comparisons.[2]

Table 3: OLS Regression Results

	Interaction Model
Historic High	3.6 (3.8)
Historic Low	**-8.4* (4.0)**
Social High	6.1 (4.0)
Social Low	**-15.5*** (4.0)**
MTurk Sample	**-7.7* (3.4)**
Student Sample	**-6.8+ (4.0)**
White	-0.4 (1.6)
Male	-0.4 (1.1)
Ideological Conservatism	**0.9+ (0.5)**
Age	-0.1 (0.1)
Historic High * MTurk Sample	2.7 (4.6)
Historic Low * MTurk Sample	-2.3 (4.7)
Social High * MTurk Sample	6.9 (4.7)
Social Low * MTurk Sample	5.2 (4.7)
Historic High * Student Sample	4.9 (5.0)
Historic Low * Student Sample	2.3 (5.1)
Social High * Student Sample	3.9 (5.1)
Social Low * Student Sample	7.9 (5.1)
Constant	**70.8*** (4.6)**
Observations	1,056
Adjusted R^2	0.181

Standard errors in parentheses
+ $p < 0.1$; ***** $p < 0.05$; ** $p < 0.01$; *** $p < 01$

Finally, to formally test our third hypothesis—that there would be differential substantive effects of information across the three samples—we tested an interaction for sample and treatment conditions. As reported in Table 3, none of the interaction terms were statistically significant, and a nested F test revealed that the interaction did not improve the model ($F = 1.02$; $p = 0.42$). This suggests that respondents across groups do not considerably differ in the magnitude of their reactions to performance information, though again we note the potential limitations of detecting small group differences given our sample sizes.

Discussion

Our findings, overall, often show similar patterns across practitioner and non-practitioner groups. General patterns in the data suggest rewards in perceived performance when respondents were given positive historic or social information and penalties in perceived performance when respondents were provided negative historic or social information. The combination of social comparison with negative information appears to carry the most weight, though support for our hypotheses regarding the weighting of negative or social information is mixed. We do not find as much support for our hypothesis related to negativity bias as might be expected from earlier research. In our data, negativity bias appears to garner the most support among principals, which pushes against prior literature that assumes experience and expertise can help to mitigate the effect of heuristics. One possibility is that principals cognitively process information about schools in slightly different ways from non-practitioner groups; another might be that they simply pay more attention to the survey since it is related to a subject in which they are well versed. It may also be the case that at least some of the differences that we observed regarding negativity bias were due to the slight change in question wording discussed earlier that principals in the social low treatment received compared to their MTurk and student counterparts. If that is the case, then the modest differences across groups we observe may be even less pronounced than the data suggest. Finally, the size of each sample was not as large as might be ideal, which can increase the uncertainty of our estimates and make it more challenging to document clear similarities and differences across groups. Larger samples would be better suited for detecting differences that are not substantial in size.

When considering the third set of hypotheses about differences across groups, it may indeed be the case that cognitive biases and heuristics used in assessing information are fairly hardwired, meaning that commonly assumed mechanisms of training and socialization are limited in their ability to combat such tendencies. Of course, additional nuances exist. For instance, MTurk respondents, who often have considerable experience participating in online survey experiments, may have developed habits and tendencies that influence their responses over repeated interactions. Though this cannot be directly tested, we see that MTurk respondents have reactions to performance data and variance that is closer to practitioners than the student sample. Of course, we cannot immediately assume that the findings here extend beyond the context of Indiana or education. There may be other types of scenarios or biases that demonstrate more significant differences in treatment effects across stakeholder groups.

Another important question our study raises relates to the role of the baseline level of performance in shaping performance assessments. To maximize external validity, we used actual average scores on the I-STEP test as our baseline. It is plausible that at least some of our findings would have been different if the reported level of performance were lower or higher or if the deviation from comparisons was greater. More work is needed to better understand how performance assessments differ in instances where performance is, on average, interpreted to be relatively poor compared to when performance is generally good. Finally, while we were able to email school principals directly to request survey participation, we were only able to provide a survey link and suggested language to faculty who then asked their students to participate in the survey. We have little knowledge of the content and timing of these interactions.

Broadly, more research comparing responses across samples is needed to better understand when and how expertise and training will alter the ways in which individuals process experimental (or non-experimental) prompts. While various theoretical discussions in the field of public administration highlight the importance of bureaucratic expertise, particularly in decision making settings, our findings above do not provide consistent support for such arguments. Instead, our findings suggest that there may be at least some ways in which common cognitive processes across people challenge or mitigate practitioner experience or expertise. To fully understand whether such implications have merit, future work should replicate this and related experiments

across multiple groups in a variety of policy settings. Research that examines when and why stakeholders do or do not differ in their responses to performance information will not only improve the field's understanding of the underlying cognitive and psychological factors that drive these processes, but can also help in efforts to design more effective accountability systems.

As experimental methods have spread and become more commonplace in the field of public administration, scholars have begun examining important questions around a variety of approaches and best practices. Threats to external validity and generalizability are almost always present when conducting experiments, and efforts to address these issues are critical for advancing the behavioral approach. One such threat involves the potential for important differences to exist in how subjects from convenience samples react to surveys as compared to practitioners who have received specialized training and may have extensive experience analyzing and interpreting performance data. Our findings suggest that, at least in some circumstances, results from surveys of students or the general populace are largely reflected in practitioner samples, though there remain some questions about the conditions under which this is the case. As with most scientific endeavors, continued investment in studies replicating these comparisons are key to understanding how the field should move forward.

Notes

1. One weakness of this experiment is that it was not preregistered. Though preregistration is becoming a more common practice in public administration and is advantageous for avoiding a number of issues, a systematic review by Hansen and Tummers (2020) found only two of forty two field experiments were preregistered. Future iterations of this and other experimental work, whether in a lab or the field, should be encouraged to follow this practice.
2. An ANCOVA F-Test for 15 groups (5 treatment conditions, 3 samples) with 1,056 pooled observations testing for an estimated effect size of .25 with an alpha of 0.05 indicated the test was sufficiently powered (.9993) for detecting differences across groups. For models separated by sample, we calculate a power of 0.83 for the principal sample (n=212), 0.96 for the student sample (n=319), and .998 for the MTurk sample (n=525).

References

Andersen, S. C. (2017). From passive to active representation—experimental evidence on the role of normative values in shaping white and minority bureaucrats' policy attitudes. *Journal of Public Administration Research and Theory 27* (3): 400–414.

Andersen, S. C., & Jakobsen, M. (2017). Policy positions of bureaucrats at the front lines: Are they susceptible to strategic communication? *Public Administration Review 77*(1): 57-66.

Anderson, D. M., & Edwards, B. C. (2015). Unfulfilled promise: Laboratory experiments in public management research. *Public Management Review 17*(10): 1518-1542.

Barrows, S., Henderson, M., Peterson, P. E., & West, M. R. (2016). Relative performance information and perceptions of public service quality: Evidence from American school districts. *Journal of Public Administration Research and Theory 26*(3): 571-583.

Bellé, N. 2013a. Experimental evidence on the relationship between public service motivation and job performance. *Public Administration Review 73*(1): 143-153.

Bellé, N. 2013b. Leading to make a difference: A field experiment on the performance effects of transformational leadership, perceived social impact, and public service motivation. *Journal of Public Administration Research and Theory 24*(1): 109-136.

Bellé, N., Cantarelli, P., & Belardinelli, P. (2017). Cognitive biases in performance appraisal: Experimental evidence on anchoring and halo effects with public managers and employees. *Review of Public Personnel Administration 37*(3): 275-294.

Bendor, J. B. (2010) Bounded rationality and politics. Vol. 6. University of California Press.

Bouwman, R., & Grimmelikhuijsen, S. (2016). Experimental public administration from 1992 to 2014: A systematic literature review and ways forward. *International Journal of Public Sector Management 29*(2): 110-131.

Capelo, C., & Dias, J. F. (2009). A feedback learning and mental models perspective on strategic decision making. *Educational Technology Research and Development 57*(5): 629-644.

Charbonneau, É., Bromberg, D. E., & Henderson, A. C. (2015). Performance improvement, culture, and regimes. International Journal of Public Sector Management.

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

Carpenter SR. 2001. Alternate states of ecosystems: evidence and its implications. In: Press MC, Huntly N, Levin S, editors. Ecology: achievement and challenge. London: Blackwell.

Cyert, R. M., & March, J. G. (1963). A behavioral theory of the firm. Englewood Cliffs, NJ.

Das, T.K., & Teng, B-S. (1999). Cognitive biases and strategic decision processes: An integrative approach. *Journal of Management Studies 36*(6): 757-778.

Donegan, M. M., & Trepanier-Street, M. (1998). Teacher and parent views on standardized testing: A cross-cultural comparison of the uses and influencing factors. *Journal of Research in Childhood Education 13*(1): 85-93.

Druckman, J. N., & Kam, C. D. (2009). Students as experimental participants: A defense of the narrow data base. In Druckman, J.N. Green, D. P., Kuklinski, J.H., & Lupia, A. (eds.) *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press, pg. 41–57.

Fuenzalida, J., Van Ryzin, G.G., & Olsen, A. L. (2020). Are managers susceptible to framing effects? An experimental study of professional judgment of performance metrics. International Public Management Journal forthcoming.

Greve, H. R. (1998). Performance, aspirations, and risky organizational change. *Administrative Science Quarterly*: 58-86.

Hansen, J. A., & Tummers, L. (2020). A systematic review of field experiments in public administration. Public Administration Review.

Hansen, K. M., Olsen, A. L., & Bech, M. (2015). Cross-national yardstick comparisons: A choice experiment on a forgotten voter heuristic. *Political Behavior 37*(4): 767-789.

Heinrich, C., Maffioli, A., & Vazquez, G. (2010). A primer for applying propensity-score matching. Inter-American Development Bank.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin 138*(2): 211-237.

Holm, J. M. (2018). Successful problem solvers? Managerial performance information use to improve low organizational performance. *Journal of Public Administration Research and Theory 28*(3): 303-320.

James, O. (2011). Managing citizens' expectations of public service performance: Evidence from observation and experimentation in local government. *Public Administration 89*(4): 1419-1435.

Kanwisher, N. (1989). Cognitive heuristics and American security policy. *Journal of Conflict Resolution. 33*(4): 652-675.

Krause, G. A. (1996). The institutional dynamics of policy administration: Bureaucratic influence over securities regulation. *American Journal of Political Science* 1083-1121.

Li, H., & Van Ryzin, G. G. (2017). A systematic review of experimental studies in public management journals. In James, O., Jilke, S. R., & Van Ryzin, G. G. (eds.) Experiments in public management research: Challenges and contributions. Cambridge, United Kingdom ; New York, NY: Cambridge University Press, 20–36.

Liu, X., Stoutenborough, J., & Vedlitz, A. (2017). Bureaucratic expertise, overconfidence, and policy choice. *Governance 30*: 705-725.

Van Maanen, J. E., & Schein, E. H. (1977). Toward a theory or organizational socialization. Massachusetts Institute of Technology.

Moynihan, D. (2006). What do we talk about when we talk about performance? Dialogue theory and performance budgeting. *Journal of Public Administration Research and Theory: 16*(2): 151-168

Moynihan, D. (2018). A great schism approaching? Towards a micro and macro public administration. *Journal of Behavioral Public Administration. 1*(1): 1-8.

Ng, T. W., & Feldman, D. C. (2013). A meta-analysis of the relationships of age and tenure with innovation-related behaviour. *Journal of Occupational and Organizational Psychology 86*(4): 585-616.

Nicholson-Crotty, J., & Miller, S. M. (2012). Bureaucratic effectiveness and influence in the legislature. *Journal of Public Administration Research and Theory 22*(2): 347-371.

Nicholson-Crotty, S., Grissom, J. A., Nicholson-Crotty, J., & Redding, C. (2017). Disentangling the causal mechanisms of representative bureaucracy: Evidence from assignment of students to gifted programs. *Journal of Public Administration Research and Theory: 26* (4): 745–57.

Nicholson-Crotty, S., Nicholson-Crotty, J., & Webeck, S. (2019). Are public managers more risk averse? Framing effects and status quo bias across the sectors. *Journal of Behavioral Public Administration 2*(1).

Nielsen, P. A. (2014). Performance management, managerial authority, and public service performance. *Journal of Public Administration Research and Theory 24*(2): 431-

458.

Nielsen, P. A., & Moynihan, D. P. (2017). How do politicians attribute bureaucratic responsibility for performance? Negativity bias and interest group advocacy. *Journal of Public Administration Research and Theory 27*(2), 269-283.

Oberfield, Z. W. (2014). Becoming bureaucrats: Socialization at the front lines of government service. University of Pennsylvania Press.

Olsen, A. L. (2015). Citizen (dis) satisfaction: An experimental equivalence framing study. *Public Administration Review 75*(3): 469-478.

Olsen, A. L. (2017). Compared to what? How social and historical reference points affect citizens' performance evaluations. *Journal of Public Administration Research and Theory: J-PART. 27* (4): 562–80.

Rhee, H.-S., Ryu, Y.U., & Kim, C.-T. (2012). Unrealistic optimism on information security management. *Computers & Security 31*(2): 221-232.

Roberts, P.S., & K. Wernstedt. (2019). Decision biases and heuristics among emergency managers: Just like the public they manage for? *American Review of Public Administration 49*(3): 292-308.

Salage, T. O. (2010). A behavioral model of innovative search: Evidence from public hospital services. *Journal of Public Administration Research and Theory: 21*(1): 181-210.

Schwenk, C.R. (1984), "Cognitive simplification processes in strategic decision-making", *Strategic Management Journal, Vol. 5* No. 2, pp. 111-128

Webeck, S., & Nicholson-Crotty, S. (2019). How historical and social comparisons influence interpretations of performance information. *International Public Management Journal*, 1-24.

Weber, M. (1947). The theory of social and economic organization. Free Press.

Wise, L. R. (2004). Bureaucratic posture: On the need for a composite theory of bureaucratic behavior. *Public Administration Review. 64*(6): 669-680.

Zhu, L., & Rutherford, A. (2019). Managing the gaps: How performance gaps shape managerial decision making. *Public Performance & Management Review*. 1-33.