Compared to What?

Olsen, Asmus Leth

Publication date:
2015

Compared to What? How Social and Historical Reference Points Affect Citizens' Performance Evaluations

Asmus Leth Olsen [*]
Department of Political Science
University of Copenhagen
ajlo@ifs.ku.dk

First version: April 2013
This version: June 2015

Abstract

The question of what is "good" or "poor" performance is difficult to answer without applying a reference point–a standard for comparison. Citizens' evaluation of performance information will therefore tend to be guided by reference points. We test how reference points alter citizens' evaluation of organizational performance. Specifically, drawing on Herbert Simon, we test how citizens use historical (internal) and social (external) reference points when making relative comparisons: How important is performance relative to past performance? And how important is performance relative to the performance of other organizations? Two experiments are embedded within a large nationally representative sample of citizens (n=3443). The experiments assign historical and social reference points for performance data on education and unemployment to citizens. We find that citizens' performance evaluation is fundamentally a relative process. Interestingly, we show that social reference points are almost twice as important to citizens' evaluations as historical reference points. We find mixed evidence of a negativity bias in citizens' relative evaluations. The strong social reference point effects have implications for studying citizens' response to performance and how managers can frame and manipulate external performance data.

KEYWORDS: performance information · reference points · experiments · citizen satisfaction

[*] I am grateful to Oliver James, Fabrizio Gilardi, George Boyne, Soeren Serritzlew, and Gregg Van Ryzin for comments on earlier drafts on this paper. An earlier version was presented at the Midwest Political Association Meeting. Panel: Political Bias, Misinformation, and Conspiracy Theories. Chicago, 11th of April 2013. A more recent version was presented the 11th Public Management Research Conference in Madison, Wisconsin, June 20-23 2013. Funding: This research was made possible by a grant from the Master of Public Governance (MPG) at the University of Copenhagen and Copenhagen Business School in Denmark. Project title: The Psychology of the Public Sector and the experiments conducted in this article have received funding from the Center for Voting and Parties (CVAP), Department of Political Science, University of Copenhagen, Denmark.

> *...the only sound basis for decisions about numbers is numerical factual information about past experiences or the experiences of others–nothing more nor less than comparative statistics.*
>
> – Herbert A. Simon (1939: 106)

The abundance of quantitative performance information available to the public raises the fundamental question of how citizens make sense of all this data (Boyne et al. 2009; James 2011a; 2011b; Charbonneau and Van Ryzin 2015; Baekgaard and Serritzlew 2015; Marvel 2015; Olsen 2015). School rankings are offered to inform voice and exit in educational choice. Hospital report cards guide critical health care decisions. Monthly unemployment rates are featured prominently in the media to help citizens hold politicians accountable on their economic promises. However, the data does not speak for itself (Moynihan 2008, 108–109). In fact, Herbert Simon raised strong concerns about how average citizens were to make sense of performance data: "It is little wonder then that the layman today finds himself bewildered...Not only is it difficult for him to decide what his city should be doing, but it is even more difficult for him to decide whether it is being done well." (Ridley and Simon 1940, 1). Indeed, simple questions have no easy or straightforward answers: How good is a top-100 school? Is a hospital with a 90 percent satisfaction rate performing well? Can we be satisfied with an unemployment rate of 6 pct?

How do citizens map these absolute numbers onto a subjective scale of good and poor performance from which they can make informed decisions? While Herbert Simon was deeply concerned with this question, he also pointed to a potential solution: The necessity of evoking *comparisons* in performance evaluations as exemplified in the epitaph and in many other instances throughout his work (Simon 1937; 1938; 1939; Simon and Ridley 1938). The need for comparison of performance information was also a key issue in the early 20th century attempt to measure government performance (Upton 1915; Freeman 1954; Williams 2003) and is reflected today in the many studies on the role of benchmarking for managerial decision making (Ammons 1999; Ammons and Roenigk 2014; Askim 2007). However, only very recently has experimental work attempted to understand the importance of relative comparisons for citizens' performance evaluations (James 2011a; James and Moseley 2014;

Charbonneau and Van Ryzin 2015). Still, fundamental questions about reference points remain unanswered: How important is relative performance evaluation? And what types of reference points do citizens actually draw on for performance comparisons?

In this article, we study the role and effects of performance reference points for citizens' evaluation of public services. We conduct two vignette experiments embedded in large-scale representative sample of the Danish population ($n = 3443$). Specifically, we aim to answer two important questions derived from a reference-dependent view of citizens' judgment about performance information (Mussweiler 2003; Tversky and Kahneman 1991): When asked to evaluate some element of the social world many tend to respond: *compared to what*? This question points to the source of comparison–the reference point. We take Herbert Simons proposal to test and estimate how citizens rely on social and historical reference points (Simon 1937; 1939). Social reference points stress that comparison happens by looking at the performance of other organizations. It can be the neighboring school, the hospitals in another city, or the employment rate in other countries. Historical reference points denote a temporal comparison between current and previous performance. How did we do last month, in previous years, or in the past decade compared with today? The analysis presented here will allow us to compare the importance of these two fundamentally difference ways for how citizens make relative performance evaluations.

Second, we ask how losses and gains differ when using social and historical reference points. Political science has long been interested in the negativity bias (Lau 1982). In short, negativity bias denotes that negative information has more profound effects on attitudes and behavior than positive information of a similar magnitude (Baumeister et al. 2001). The negativity bias has also been found relevant for understanding how citizens respond to performance information (Boyne et al. 2009; James 2011a; James and Moseley 2014). In terms of implications, the negativity bias has been noted as the underlying driving force for blame avoidance among managers and politicians (Weaver 1986). The question is how the negativity bias works in relation to historical and social reference points. Is performing relatively poorly punished in the same manner as being relatively good is rewarded?

In order to test these questions we employ a novel survey experimental design. The survey experiments ask individuals to evaluate two types of performance information, namely: (a) school performance from grade average information and (b) municipal unemployment performance from local unemployment rates. As a unique feature we are not only randomly assigning the type of reference point that citizens are exposed to but also the relative distance between the organization's performance and the performance of the reference point. This provides a very direct causal test of how citizens' evaluations are altered as the relative difference in performance shifts. This is to our knowledge, the first experiment in public administration which allows to causally test how various levels of both historical and social relative performance affect citizens' evaluations of public services. Moreover, this design can easily be implemented by others in order to understand relative performance effects in other contexts and for other actors (e.g., policy makers and managers).

The results point to that relative performance evaluation is very important to citizens. Most strikingly, across both experiments, we find that social reference points have a stronger impact on citizens' evaluations than historical reference points. That is, citizens are more inclined to draw on a refererence point which indicates the performance of other organizations than a reference point which captures the historical performance of the very same organization they are evaluating. Findings on the negativity bias are more mixed and seem to depend more on type of performance measure than the actual reference point. However, interestingly, we find that relative performance has the strongest impact on evaluations in the near vicinity of the reference point.

The findings have a number of implications for both public administration research and the practice of managers. First, it adds to the century old call for empirical research into how the presentation of performance information affect citizens' perception of public services (Simon 1937; Freeman 1954). Second, it calls for further research into the downstream effects in terms of managers use of reference point as a means to give a more positive impression of their organization's performance (Moynihan 2008). The importance of social reference points raises concerns about subtle ways to game and manipulate the information that citizens are served (Smith 1995). Finally, understanding the importance of relative performance evaluations and the reference points they rely on have implications for how we study citizens' satisfaction with the public sector and democracy (Parks 1984; Van Ryzin 2004). These implications will be discussed in greater detail in the concluding section.

THEORY AND HYPOTHESIS

Reference Points and the Evaluation of Performance Information

Since the early days, public administration scholars have been concerned with how (if at all) citizens make meaningful use of performance information about public services (Willimott 1924; Ridley and Simon 1940; Freeman 1954; Williams 2004). For Herbert Simon the possibility of comparing the performance of an organization with some standard was seen as a key instrument for unlocking the potential of performance information (Simon 1937; 1939). However, since then surprisingly little empirical work has been done on citizens use of ”comparative statistics”. James (2011a) finds that good relative performance information raises citizen satisfaction while bad relative performance information tends to lower citizen satisfaction. James and Moseley (2014) find something similar in a field experiment and also suggests that relative comparisons of performance leads citizens' to attribute more responsibility to local service providers. Before outlining our hypothesis on the various sources of reference points, we will outline how the early focus on comparisons in public administration ties in with parallel developments in cognitive and social psychology.

The fundamental argument put forward here will be that citizens' judgment about public services is highly reference dependent: Citizens compare an organization's performance with some reference point in order to arrive at a relative measure of performance. Accordingly, from a formal definition a reference points is a: ”stimuli of known attributes that act as standards against which other categorically similar stimuli of unknown attributes are compared in order to gain information” (Yockey and Kruml 2009, 97). We can therefore think of a reference point as a standard or yardstick against which citizens can compare the performance information of a given organization. In performance management we are used to think of reference points more formally as benchmarks (Charbonneau and Van Ryzin 2015).

The most well-known account of reference dependent judgment is provided in prospect theory (Kahneman and Tversky 1979). Here reference dependence implies that citizens make judgment about a public organization's performance in terms of losses or gains compared to a reference point. An implication is that the same piece of absolute performance information can be evaluated differently depending on the framing of the reference point. In social and cognitive psychology, many have argued that individuals are deeply inclined to evaluate their own abilities and opinions in relative terms and not in absolute ones (Festinger 1954). Our perception of everything from light, sound, to colors will have a strong relative component (Kahneman and Tversky 1979; Kahneman 1992). The same holds for the social world where changes matter more to citizens' understanding of value than absolute levels (Markowitz 1952). Changes or differences contains a relative component: a comparison of the value at hand with the standard of the past. In fact, across a number of fields we find the notion that performance is coded as either a loss or gain, success or failure, negative or positive, and good or bad depending on how the object of evaluation falls relative to some reference point that informs expectations, aspirations, or norms (March and Simon 1958; Cyert and March 1963; Kahneman, Knetsch, and Thaler 1991; Greve 1998; Heath 1999; Meier et al. 2015). For instance, in organizational science, reference points can serve as goals or aspirational levels for the performance of organizations or individuals (March and Simon 1958). Comparing outcomes to some reference point for performance provides a measure of attainment discrepancy and places performance in a domain of either ”falling short” or ”making the cut” (March 1988).[1]

Some attribute this binary division of information around a reference point to bounded rationality: The reference point is a heuristic to overcome limited cognitive capabilities and the complexity of the information at hand (Simon 1955; James 2011a; Olsen 2013). In fact, in some of the early calls for comparative performance information for citizens, we also find an implicit reliance on a model of bounded rationality for human cognition: ”The human brain cannot absorb series of unrelated figures without a yardstick by which it can judge them. What is a yardstick of state costs? Since there is no objective standard, comparisons are the only means of arriving at relative judgment.” (Freeman (1954, 124). We can therefore view the reference point as a judgmental shortcut that allows citizens to form evaluations about abstract numerical performance information which they possess no deeper knowledge about (Mussweiler 2003). Judgment about an absolute performance number requires some

[1] In studies of consumer and citizen satisfaction, reference points are seen as informing the expectations which citizens have to public services (Oliver 1980). These expectations are important as they directly feed into citizens' subjective satisfaction with the public sector. In the widely applied expectations-disconfirmation model, satisfaction is formed by the discrepancy between expectations and the experienced performance (VanRyzin 2004; 2013).

scale-specific information. What is the possible interval of the scale? What is the mean of the scale and what is its distribution? With the help of comparison, the task shifts to comparing the abstract absolute performance of an organization with a reference point. However, comparison is not just a simplifying shortcut. Social psychologists have found that relative comparisons are efficient because they lead to faster decision making without resulting in worse decisions (Mussweiler and Posten 2009). Comparison has also been found to reduce judgmental uncertainty (Mussweiler and Epstude 2011).

In summary, we have multiple accounts of the importance of reference points spanning psychology, political science, organizational science, and public administration. Against this backdrop, we expect that relative comparisons should be important for citizens' judgment of organizational performance. Moreover, citizens' evaluation of performance information will be influenced by relative calculations even in the presence of absolute performance measures.

Where to Look for Comparison? Historical vs. Social Reference Points

At this point we have simply pointed out that reference points are what informs citizens' relative evaluation of performance information. However, as Levy (1997, 100) has noted, we are often left with ”a reference-dependent theory without a theory of the reference point”. There are can be many sources of reference points (Tversky and Kahneman 1981; Yockey and Kruml 2009). Herbert Simon was already in his early years arguing for comparisons ”across time” or ”with other organizations” as the most natural point of departure when seeking out a reference point for performance comparisons (Simon 1939). One of the earliest examples of this is found in Simon (1937, 525):

> ”Where is the data to be obtained? There are two possible sources: either from an examination of trends within a city over a period of years, or from an examination of data gathered from a number of cities or from different sections of the same city.”

In essence, Simon argues that comparisons are either (internal) with-in-subject over time or (external) between-subject across space. In the following we will refer to these as 1) historical reference points (i.e., comparisons across time) and 2) social reference points (i.e., comparisons between units).[2] Across a number fields, there has been a dual focus on both types of reference points. In psychology, social and historical reference points have been contrasted for individuals' self-evaluation (Robinson-Whelen and Kiecolt-Galser 1997). Crosby (1976) cites Sorokin (1925, 72) for a very illustrative quote on social and historical reference points wealth evaluations: "poverty or wealth of a man is measured, not by what he has at present but by what he used to have before or what others have". Contrasting social and historical reference points for organizational performance evaluation has a long tradition in organizational science (Greve 1998). Cyert and March (1963) viewed organizational aspirations as the combined product of the performance of others (social comparison) and an organization's own past performance (i.e. historical comparison). Decision makers are also expected to care about performance relative to that of other organizations as well as the organization's own historical performance (Heath 1999). As March and Simon notes (1958: 203–204):

> ”The level of satisfactory performance is likely to be very close to the actual achieved level of recent performance...Individuals adjust their criteria to the achieved levels of other individuals with whom they compare themselves, and to the levels that are established as norms by relevant reference groups. Organizations adjust their criteria to the levels achieved by other organizations."

Let's begin by looking closer at historical comparisons where individuals compare current performance with some past historical reference point. Here, we focus explicitly on historical comparison within a given organization in order to draw a sharp distinction to social comparison. Historical reference points emphasize that citizens care about what direction of change in performance an organization is experiencing. We find this very pronounced in the tradition of retrospective voting studies which emphasizes the calculation of differences between current and past performance as a means for prospective evaluation of the incumbent government (James and John 2007). For in stance, Hibbs (1982, 314) argued that citizens' evaluation of performance was driven by comparing the cumulative

[2] In studies of benchmarking these two reference points are often referred to as internal and external (Foltin 1999; Bird et al. 2005).

performance of the current administration with the performance of previous administrations. Historical reference points are clearly the most broadly used reference point for performance comparison in political science (Kayser and Peress 2012, 680). The status-quo-bias can be seen as another example of historical comparison (Samuelson and Zeckhauser 1988). Past performance is the status quo which current performance is evaluated against. If current performance is below past performance it will be seen as a worsening. In addition to Herbert Simon, we also find an emphasis on historical comparisons in Ridley (1937: 33) who notes that a good report of performance entails ”comparative data” where ”the present year’s accomplishments should be compared with those of previous years”. With the existing theoretical and empirical work in mind we propose the first hypothesis about relative performance evaluation:

> HISTORICAL REFERENCE POINTS (H_1): *Providing information about better past performance of an organization will lower citizens’ assessment of current performance and providing information about worse past performance of an organization will raise citizens’ assessment of current performance.*

Social reference points offer an alternative source of comparison. In social comparisons, individuals compare current performance with the performance simultaneously obtained by others. Whether individuals explicitly seek it out or not, the environment will inevitably provide information about the achievement of others (Chapman and Volkmann 1939). We simply cannot ignore the outcomes of similar entities. This broad idea has been framed differently across fields and research traditions. The idea is reflected in the theoretical traditions of bandwagon effects or ”keeping up with the Joneses” in early studies of consumer behavior and public opinion (Pierce 1940; Leibenstein 1950). Reference group theory also denotes comparison with *external others* when making judgment about oneself (Shibutani 1955). The research tradition on relative deprivation reflects a similar sentiment (Runciman 1961; Stark and Taylor 1991). Perhaps the most successful formulation of the idea is found in social comparison theory (Festinger 1954). Social comparison theory states that we have an urge to evaluate our own opinions and abilities by comparing them with the opinions and abilities of similar others. For instance, in studies of health, happiness, and wealth, there is strong evidence of relative comparison with others (Brickman and Campbell 1971). The health, happiness, and wealth we observe around us affect our personal self-assessment on these dimensions.

Initially, social comparison was seen as affected by the availability of absolute information. In other words, social comparison was a second best option in the absence of absolute information. Modern accounts of reference dependence find that relative social comparison can be relevant even if absolute measures are available (Moore 2007).[3] In economics and political science, we have seen some indirect application of social comparison processes which can be useful for our purpose. Salmon (1987) introduced the idea of social reference points with his yardstick theory for holding government accountable. He argues that historical comparisons will constitute a very noisy reference point for comparison with current performance. Over time, there will be exogenous disturbances. There can be business cycles and long-term trends which cannot be factored in when comparing the present situation with past times. Instead, citizens' are better off by applying neighboring jurisdictions or other similar social reference points as a means for comparison with current performance (Hansen et al. 2014). In public administration, Charbonneau and Van Ryzin (2015) found evidence of that citizens rely on national averages to compare individual organizational performance against. Along the same line we expect the following effect of social reference points:

> SOCIAL REFERENCE POINTS (H_2): *Providing information about better performance among other organizations will lower citizens' assessment of an organization's current performance and providing information about worse performance among other organizations will raise citizens' assessment of an organization's current performance.*

The Negativity Bias and Reference Points

The negativity bias has gained considerable attention in the study of citizens' attitudinal and behavioral responses to performance information (James and John 2007; Boyne et al. 2009; James 2011a; 2011b; James and Moseley 2015; Olsen 2015). The negativity bias stresses an asymmetrical response

[3] Goodman and Haisley (2007) pointed out that most social comparison research focus on how individuals apply information to others. That is, social comparison refers to self-other comparisons. Few studies apply the insights of social comparisons to organizational level analysis (Greve 1998, 60). However, social comparison theory point to the importance of general *comparison processes* and not exclusively on individual comparisons (Goodman and Haisley 2007, 115).

to positive and negative information where ”negative events are more salient, potent, dominant in combinations, and generally efficacious than positive events” (Rozin and Royzman 2001, 297). The final hypothesis focuses on differences in citizens' response to performance information depending on the relative performance being above or below the reference point. In economic voting studies, there are multiple accounts of a negativity bias in response to historical-temporal reference points. That is, a worsening economy damages the incumbent to a greater degree than an improving economy helps. For instance, Kinder and Kiewiet (1979) found evidence of decreasing voter support for the incumbent only if the unemployment rate went up. The negativity bias is also reflected in media coverage. Soroka (2006) found that negative economic performance is covered more widely in the media than positive economic performance of a similar magnitude.

In the study of performance information in public administration the findings have been more mixed–but with some indication of a negativity bias. Across, two studies James (2011a; 2011b) only found partial support for negativity bias in various field and survey experiments with performance data. In an observational study, Boyne et al. (2009) found that performance information has a asymmetrical effect on support for the incumbent among English local governments. Only bad performance is punished while good performance is not rewarded. Olsen (2015) find indirect evidence of a negativity bias as citizens respond more strongly to a negative worded ”dissatisfaction rate” than a logically similar ”satisfaction rate” in an equivalence framing experiment. These results align with James and Moseley (2014) who find that negative performance information has more profound effects across a number of dimensions.

The core question becomes how this asymmetry plays out in a context of relative performance evaluation. Reference points and the negativity bias have a close connection, because the reference point is likely to decide if an organizations performance is determined to be good or bad. Reference points make the valence of an organizations performance more salient. A core insight in prospect theory was an asymmetrical effect around the reference point with larger effects in the domain of losses than in the domain of gains (Kahneman and Tversky 1979; 1991). In a similar manner, we expect that the negativity bias implies that citizens respond less positively to ”relative good” performance compared with their response to ”relatively bad” of the same magnitude:

NEGATIVETY BIAS (H_3): *If an organization performance worse than the reference point, it will have a stronger effect on citizens' assessment of performance than if organizational performance is better than the reference point.*

DATA AND DESIGN

Two Experiments in a Large Nationally Representative Sample

We have outlined a set of causal predictions about how citizens evaluate public services given some specific variations in the relative performance. The causal nature of these predictions imply that an experimental test is the most appropriate design. With experiments we can directly manipulate and randomly assign the absolute and relative performance information which citizens are exposed to. The hypotheses outlined above will be tested with a set of experiments embedded in an online survey. Experimental designs have been applied in a recent string of research on how citizens apply performance information in their considerations (James 2011a; Charbonneau & Van Ryzin 2015; Baekgaard and Serritzlew 2015). Given the expectations outlined above, we need an experimental design which can both (*a*) compare the relative effect of social reference points with that of historical reference points, and (*b*) compare relatively negative performance with relatively positive performance of a similar magnitude. In order to do so, we employ a set of experiments which were conducted in a large nationally representative sample of the Danish population.

Participating subjects

The experiments were conducted in a survey fielded in YouGov's Danish online panel which consists of 40,000 Danes. [4] Respondents were recruited to participate via e-mail with an embedded link to the survey. In total, 3443 subjects participated in the experiments which yielded a response rate of 42%.[5] The sampling frame for the study was restricted to citizens between the age of 18 to 74. The sample was pre-stratified on gender, region, age, and party choice at the most recent national election (2011).

[4] The panel is recruited via both online, radio and newspaper ads as well as through telephone surveys. The survey was conducted between the 15th of October and the 22th of October (2012).

[5] In total 8.204 were invited to participate in the survey.

Table 1 below shows descriptive statistics of the sample and highlights a near nationally representative sample in terms of age, gender, education, and region of residence.

Table 1: Descriptive statistics

Variable	Pct.
Gender (male)	49.8%
Age (mean years)	50.5 (SD=15)
Education	
High school or less	18.4%
Vocational training	24.7%
Short-cycle tertiary	12.9%
Medium-cycle tertiary	28.6%
Long-cycle tertiary	15.5%
Geographical region	
Capital area	24.7%
Zealand	15.3%
Southern Denmark	23.9%
Middle Jutland	23.8%
Northern Jutland	12.3%

CAWI (computer-assisted web interviewing) survey. Note: n=3443.

In comparison with other recent experiments in the field this sample diverge on two important dimensions. First, it is a very large sample (n=3443) which allows us to obtain the necessary statistical power for disentangling many components in the same experimental design–including various reference points and varying distances between an organization's performance and the reference point performance. Secondly, by employing a nationally representative sample we provide strong external validity of the results for people with diverse backgrounds and experiences with performance information. We hereby make sure that the findings we may uncover are not contingent on some specific characteristic of the sample but can be generalized to the public-at-large.

Two Experimental Vignettes

The two experiments focus on school grades and local government unemployment rates as examples of performance data. School grades provide an example of a performance measure which represents a specific public service. Education performance information is widely published today in most developed countries. Unemployment rates are, on the other hand, an example of a more general economic performance indicator which dominates media coverage of national or local performance (Soroka 2006). The actual content and setup of the two experiments differed. However, for both studies the experimental design entails randomized treatments at two levels. The first level of treatments randomly assign subjects to either: (1) absolute performance, (2) absolute performance and an absolute social reference point, and (3) absolute performance and an absolute historical reference point. The first level of treatment allows for both comparing the effects of absolute performance with the effect relative to some reference points, and for determining variations in relative effects for both social and historical reference points. The second level of treatment randomizes the actual numerical content of the performance information which the subjects were provided with. The second level provides a unique setting for understanding how citizens process relatively bad vs. relatively good performance information. It is also an optimal robustness check for understanding sensitivities in evaluations for different degrees of relative performance.[6]

Experiment I: Citizens' Evaluation of School Grade Averages

The first experiment deals with the research question in a setting of education performance data. In Denmark, the Ministry of Education has released unadjusted school grade averages for all elementary schools over the course of the last ten years. Subjects were randomly assigned one of three different treatments providing subjects with different absolute performance and/or different reference points. An overview of the different conditions is provided in table 2. Under the first treatment, subjects are only provided with the absolute grade average of an unnamed school (n=1156). This is a form of control state without any reference point information where we can observe the effect for difference levels of absolute performance. In the social comparison treatment, the subjects are presented with the absolute grade average of an unnamed school along with a municipal grade average for all schools in the same municipality (n=1148). Finally, for the historical comparison treatment subjects were presented with the average grade of a school along with the previous year's average grade level of the same school (n=1139). Operationalizing the historical treatment is more straightforward than the

[6] As the survey experiment comprises two experiments their order was randomized to avoid sequence effect across the different treatments.

social comparison as it is only a matter of determining the temporal lag on the reference point. Here, a one year lag is used as it resembles the actual rate of publication of grade averages in Denmark. The social comparison case is trickier as comparison with others can happen in numerous ways. In social comparison theory, it is argued that individuals seek out comparison with others that are believed to have similar characteristics which affect performance (Goethals nad Darley 1977). Recent studies of benchmarking has also relied on similar operationalizations of national or local averages (James 2011; Charbonneau and Van Ryzin 2015). Here, I apply schools from the same municipality which both are likely to share some common characteristics and which are also located close to each other.[7]

Table 2: Experimental design: School grade averages with varying reference points

Baseline question:		
Each year, the Ministry of Education releases a grade average for all schools in the country. How well do you think this school is doing?		
Treatment frames:		
Absolute level only (n=1156)	*Social comparison* (n=1148)	*Historical comparison* (n=1139)
Treatment texts:		
The school has a grade average of x.	The school has a grade average of x_1. The grade average for all schools in the same municipality is x_2.	The school has a grade average of x_1. The schools grade average was last year x_2.
Numerical treatments:		
$x \in N(\mu = 6.5, \sigma = 1.0)$	$x_1 \in N(\mu = 6.5, \sigma = 1.0)$, $x_2 \in N(\mu = 6.5, \sigma = 1.0)$, $x_1 \perp x_2$	$x_1 \in N(\mu = 6.5, \sigma = 1.0)$, $x_2 \in N(\mu = 6.5, \sigma = 1.0)$, $x_1 \perp x_2$

Note: 6.5 percent was the national grade average for all schools at the time of the study.

At the second stage of treatment, the actual school grade values are drawn from a normal distribution with an average of 6.5 and a standard deviation of 1.0. Higher grade averages correspond to better results.[8] The mean grade average corresponds to the actual grade average obtained for all schools in 2011. The distribution reflects, to some degree, the school grade averages people would experience in their everyday life. School grade averages will therefore constitute a performance measure which most individuals have some familiarity with.[9] For each subject, a grade value is drawn independently

[7] In Denmark, public schools are governed by municipalities and school choice is most likely to happen within the boundaries of the same municipality.

[8] The Danish grade scale is a 7-point scale with values from worst to best: -3, 0, 2, 4, 7, 10, and 12. It is directly comparable to an American scale where 12 corresponds to an A, and both 0 and -3 represent an F.

[9] The averages are presented with one decimal which corresponds to how the government and the media would normally report them. Individual grade averages for students are also presented this way in the final exam transcripts.

from a distribution with the above mentioned mean and standard deviation. That is, for the absolute treatment subjects are presented with only one value drawn randomly from this distribution. For the two reference point treatments, subjects are provided with two values randomly drawn from the same distribution: one for the absolute level of the unnamed school, and one for the social or historical reference point. For instance, some subjects have been told that the municipality has an average of 5.5 and the average for all other schools in the municipality is 6.8. In a similar manner, other subjects one of more than 50 different combinations of relative performance. This implies that the findings will be valid and robust for a large range of different levels of both absolute and relative performance. Often survey experiments only present subjects with a few sets of various values or written descriptions.

As the dependent variable subjects were asked to provide their assessment of the school. For all conditions, the subjects should score their response on a slider scale from 0 to 100 where 0 was labeled *very bad* and 100 was labeled *very good*. A graphical presentation of the scale is provided in appendix A (figure 3). Across all conditions the average response was 52.83 with a standard deviation of 20.18.

Experiment II: Citizens' Evaluation of Municipal Unemployment Rates

The second experiment is similar to the first one in terms of the general setup. However, the substantive setting is different as citizens are asked to evaluate municipal unemployment performance based on information about local unemployment rates. The wording of treatment conditions are outlined in table 3. In the baseline condition, subjects are only provided with the absolute unemployment rate in an unnamed municipality (n=1155). In the social comparison treatment, the subjects are presented with the absolute unemployment rate of the unnamed municipality along with a hypothetical national unemployment rate (n=1142). Finally, in the historical reference point treatment, subjects were provided the unemployment rate of the unnamed municipality along with previous years' absolute unemployment rate in the same municipality (n=1146).

At the second stage of treatment, subjects are assigned a randomly drawn unemployment rate. For the unemployment rates, the values are drawn from a normal distribution with a mean of 6.3% and a standard deviation of 1.0%. The average corresponds to the national unemployment rate in

Table 3: Experimental design: Local unemployment rates with varying reference points

<table>
<tr><th colspan="3">Baseline question:</th></tr>
<tr><td colspan="3">How well do you think that this municipality is doing in terms of unemployment?</td></tr>
<tr><th colspan="3">Treatment frames:</th></tr>
<tr><td><i>Absolute level only</i> (n=1555)</td><td><i>Social comparison</i> (n=1142)</td><td><i>Historical comparison</i> (n=1146)</td></tr>
<tr><th colspan="3">Treatment texts:</th></tr>
<tr><td>The municipality has an unemployment rate of x%.</td><td>The municipality has an unemployment rate of x_1%. The unemployment rate in the rest of the country is x_2%.</td><td>The municipality has an unemployment rate of x_1%. The municipality had last year an unemployment rate of x_2%.</td></tr>
<tr><th colspan="3">Numerical treatments:</th></tr>
<tr><td>$x \in N(\mu = 6.3, \sigma = 1.0)$</td><td>$x_1 \in N(\mu = 6.3, \sigma = 1.0)$, $x_2 \in N(\mu = 6.3, \sigma = 1.0)$, $x_1 \perp x_2$</td><td>$x_1 \in N(\mu = 6.3, \sigma = 1.0)$, $x_2 \in N(\mu = 6.3, \sigma = 1.0)$, $x_1 \perp x_2$</td></tr>
</table>

Note: 6.3% was the average national unemployment rate at the time of the study.

Denmark available at the time of the study. For each treatment status, unemployment rates are drawn independently from a distribution with the above parameters. For the relative frames, two independent values were drawn from this distribution: one for the unnamed municipality, and one for the national average (social reference point) or last year's unemployment rate (historical comparison). The same response scale was used as in the experimental school vignette. The average response was 50.63 with a standard deviation of 20.93.

If the random assignment of subjects have worked as intended, we would expect groups to be probabilistically similar. Table 7 in the appendix B shows descriptive statistics of the numerical treatment variables and selected background variables across the three main treatment groups for both experiments. The first three variables are the second stage of randomization included in the experiment which randomly assigned values of absolute and reference point performance across the main treatment groups. Naturally, we do not want the main treatment groups to differ in terms of the numerical treatment. If that was the case, then we would not be able to seperate the effect of the type of reference point from the actual absolute and relative level of performance. As expected there is not variation in the performance assigned across the treatment groups. The relative performance variable simply capture the difference between the absolute performance and the reference point performance. Looking at background characteristics of the subjects also shows no systematic variation between treatment groups.

EMPIRICAL FINDINGS

Relative Performance Information: Effect of Social and Historical Reference Points

First, we turn to the task of comparing the effects of social and historical reference point across the two experimental vignettes. Results from both experiments are shown in table 4 below. Here we can estimate and compare the causal effects of the randomized performance numbers across the different treatment groups. Column A shows the effect of the treatments where subjects were only provided the absolute level of the organization being evaluated. In columns marked B the models for the social reference point treatments are provided. Finally, columns C provides the results from the historical reference point treatment conditions. As all independent variables are randomized we can interpret the coefficients as causal effects. Furthermore, as the a the same scale is used as dependent variable we can compare effect sizes between treatment groups and across experiments.

Results from the school grade experiment shows the following: In the absolute performance treatment in column A the absolute grade average is highly positively correlated with citizens' evaluation of school performance. This simply indicates that the subjects evaluated the unnamed school around 8.5 points higher for each average grade point. It tells us that the subjects attended to the information of grade averages as one would expect: better performing schools are given better evaluations by the subjects (James 2011a).[10] In the next two models the numerical treatment of the reference point is introduced. For the social reference point treatment group we observe a substantially strong and significant negative effect of the other schools' grade average. Simply put, citizens' evaluation of a school's performance worsens by -7.5 points for each higher grade point average among the reference point schools. This provides strong support for the importance social reference points for citizens evaluation of performance (H_2). Moving on to the historical reference point treatment group we also find a negative effect. However, at -4.3 the effect is substantially weaker by 3.2 points. It provides support for the importance of historical reference points (H_1), but also indicates that they might be substantively weaker than social reference points. This difference is also notable via the much higher adjusted R^2 for the social reference point condition compared with the historical one (R^2 .34 vs. R^2 .26).

[10] Graphs of all the simple correlations between the absolute performance and the evaluation is shown in Appendix C.

In table 5 the difference between the social and historical reference points are tested directly. This is done by pooling the data from the two treatments and interacting the reference point value variable with a dummy indicating the type of reference point (i.e, social or historical). In the first column we can note that the 3.2 points difference is highly significant. Overall, we find strong support for the importance of both reference points, but also that social reference point affected citizens' evaluation of school performance substantially more than the historical reference point.

Results from the unemployment experiment shows the following: As expected the absolute treatment in column A in table 4 shows a significant negative effect of a municipality's unemployment rate on citizens' evaluation of municipal unemployment performance. In this context, the interpretation of the coefficients is reversed as they now measure a negative outcome (unemployment rates) as opposed to the positive measure of school grade averages. This being said, the effect of about -4.6 is much lower than we saw in the school case. The same holds for the R^2. Exposure to the absolute measures of unemployment rates had a much more noisy effect on the subsequent evaluation of the unnamed municipality. It indicates that citizens are less sure of how the unemployment information affects their impression of unemployment performance. One explanation could be that citizens recognize that unemployment rates are more affected by exogenous factors than school grade averages. They are therefore more reluctant to evaluate unemployment performance based on the unemployment rate alone. However, turning to the reference points we can see how they become an important aid for evaluating the more ambiguous level of unemployment in absolute terms:

The next two columns of models show the effect of the two reference point treatments. The social reference point of the national unemployment rate shows a positive effect of the same magnitude as the negative effect of the municipality's own unemployment rate. That is, citizens' evaluate the target municipality's unemployment performance 7.8 points better if the national unemployment rate increases by 1 percentage point (H_2).

Table 4: Absolute, Social, and Historical Comparisons

	Experiment and Treatment Groups:					
	School Grade Average			Municipal Unemployment Rate		
	A: Absolute	B: Social ref.	C: Historical ref.	A: Absolute	B: Social ref.	C: Historical ref.
Absolute Performance	8.46^{***}	10.45^{***}	9.09^{***}	-4.62^{***}	-7.76^{***}	-5.79^{***}
	(0.48)	(0.51)	(0.51)	(0.61)	(0.54)	(0.56)
Reference Point Performance		-7.51^{***}	-4.29^{***}		7.16^{***}	4.34^{***}
		(0.52)	(0.50)		(0.52)	(0.55)
Constant	-3.57	35.86^{***}	20.87^{***}	79.34^{***}	53.64^{***}	61.57^{***}
	(3.18)	(4.65)	(4.70)	(3.90)	(4.75)	(4.91)
Observations	1,156	1,148	1,139	1,155	1,142	1,146
R^2	0.21	0.35	0.26	0.05	0.25	0.13
Adjusted R^2	0.21	0.34	0.26	0.05	0.25	0.12
Residual Std. Error	16.45	17.65	17.18	21.24	17.67	19.15
F Statistic	306.10^{***}	301.59^{***}	200.77^{***}	58.25^{***}	192.52^{***}	82.68^{***}

Note: OLS estimates. SEs in parentheses. $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 5: Different Effects of Social and Historical Reference Points

	Experiment:	
	School Grade Average	Municipal Unemployment
Absolute Performance	9.79^{***}	-6.73^{***}
	(0.36)	(0.39)
Ref. Point Performance	-4.27^{***}	4.36^{***}
	(0.50)	(0.53)
Ref. Point Type (1 = Social)	23.75^{***}	-20.14^{***}
	(4.73)	(4.83)
Ref. Type $*$ Ref. Performance	-3.21^{***}	2.78^{***}
	(0.72)	(0.76)
Constant	16.19^{***}	67.33^{***}
	(4.08)	(4.14)
Observations	2,287	2,288
R^2	0.31	0.19
Adjusted R^2	0.31	0.19
Residual Std. Error	17.43	18.45
F Statistic	254.95^{***}	135.51^{***}

Note: OLS estimates. SEs in parentheses. *p<0.05; **p<0.01; ***p<0.001

A similar pattern is found for the historical comparison treatment where the subjects were exposed to last year's unemployment rate for the target municipality: Citizens' respond negatively to higher unemployment in the target municipality and positively to higher levels of past unemployment (H_1). Interestingly, we also here find the social reference point to be much strong than the effect of 4.3 points found for the historical reference point. In table 5 the difference between the social and historical reference points are tested directly. The difference in effect between two reference points amount to 2.8 and is highly significant. Furthermore, as in the school case we find much higher explanatory power for the social reference point than the historical one (R^2 .25 vs. R^2 .12).

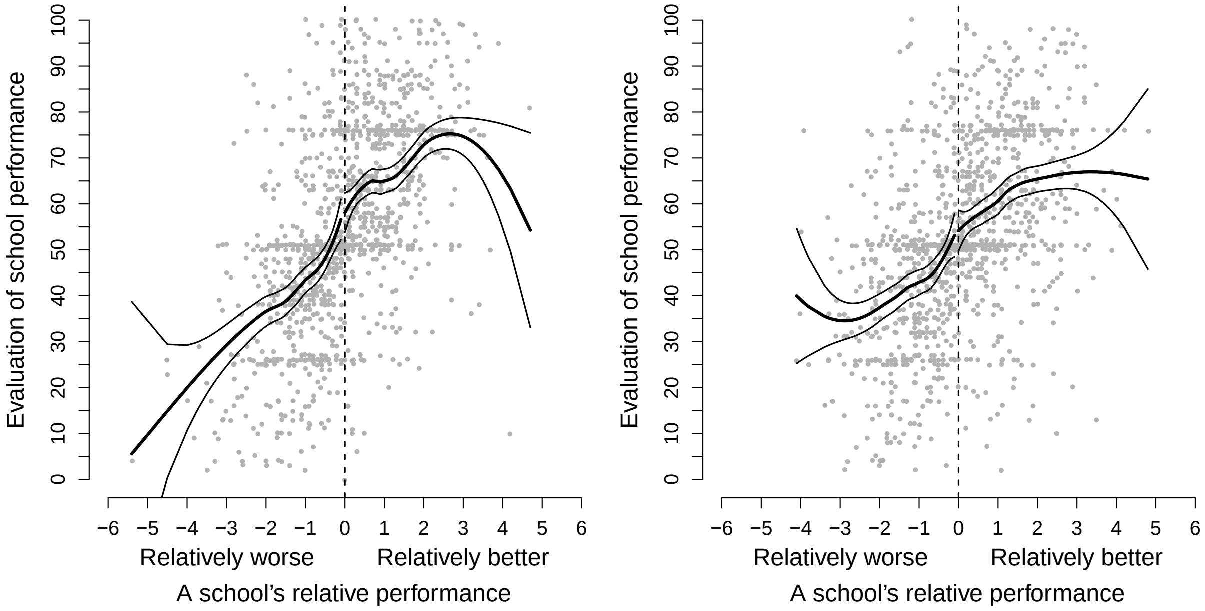
The increases in explanatory power for models with social reference points is not due to differences in how much subjects consider each treatment. Trimmed mean response times for the social grade average treatment was 17.9 seconds, and 18.5 seconds for the historical case. In the unemployment experiment the response time was 14.7 seconds for the social treatment and 15.1 seconds for the historical treatment. In other words: the social reference points provided slightly faster response times and much less noise in evaluations. This is in line with psychological research showing faster and more efficient judgment via social comparison (Mussweiler and Epstude 2009). Finally, it is worth noting that while performance information contribute less to the evaluation of unemployment relative to the case of grades, we can see how the relative importance of reference points is higher for the unemployment case. One way to capture this is to compare the relative strength of the absolute information and the reference point. For the case of unemployment these two are much more closely aligned in magnitude. This indicates that relative performance becomes even more important if the absolute performance measure is more ambiguous.

Taken together we find a very robust and substantially important pattern across both experimental vignettes: Citizens are very much attuned to reference point information and they are substantially and significantly more affected by social reference point than historical ones.

Negativity Bias and Reference Dependence

We now turn to the question of a negativity bias in citizens' relative performance evaluations (H_3). In figures 1 and 2 the results are plotted. The x-axis shows the relative grade difference between the organization and the reference point. Negative values denote that the organization's performance is numerically lower than the reference point. The fitted lines are smooth lowess fits with 95% confidence intervals in order to capture non-linear trends. Separate lines are plotted for relatively good and bad performance to allow for differences in the slope for "relatively good" and "relatively bad" performance. The differences in slopes are more formally tested in table 5. The table includes the measure of relative performance (i.e. difference bewteen absolute performance and reference point performance) as shown in the figures. It also contains a simple dummy variable capturing if absolute performance is higher than the reference point. Finally, it includes an interaction between the two variables which captures differences in the effect of relative performance depending on absolute performance being above or below the reference point.

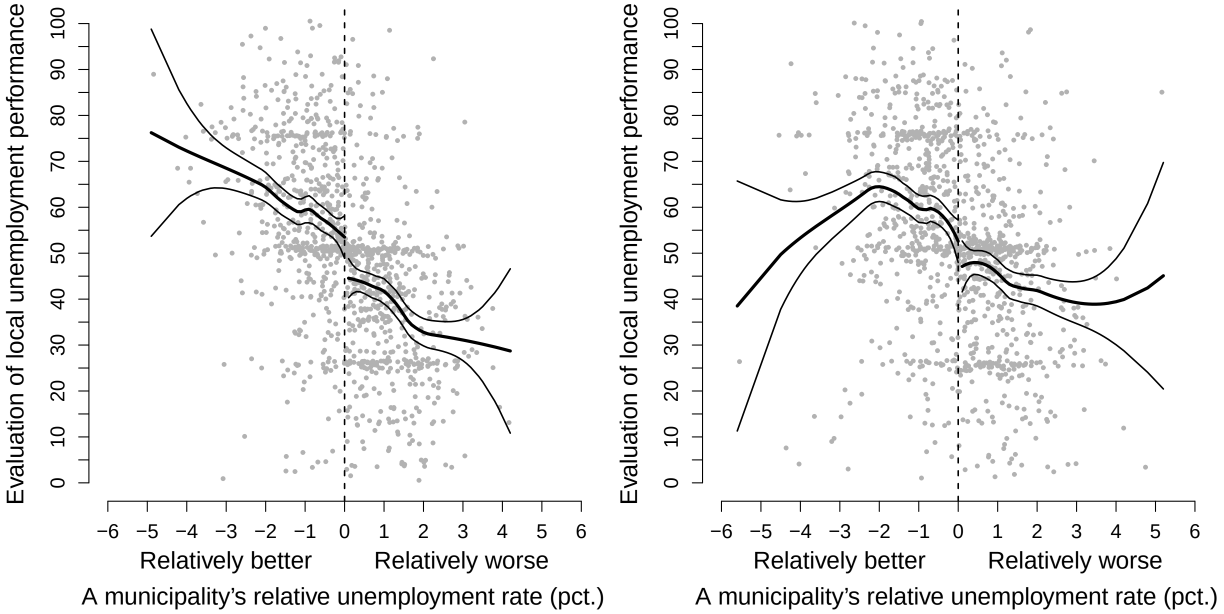
First, we turn to the grade experiment in figure 1. In the social reference point treatment to the left there is a stronger reaction in the negative domain indicated by a more steep trend line below



(a) **Social Reference Point (N=1148)**: Citizens' evaluation of school performance and the relative social performance, i.e, *a school's grade average minus the average for all schools in the same municipality.*

(b) **Historical Reference Point (N=1139)**: Citizens evaluation of school performance and the relative historical performance, i.e, *a school's grade average minus it's average last year.*

Figure 1: School Grade Average Experiment: Randomly Assigned Relative Performance



(a) **Social Reference Point (N=1142)**: Citizens' evaluation of local unemployment and the relative social performance the municipality, i.e, *a municipality's unemployment rate minus the national rate.*

(b) **Historical Reference Point (N=1146)**: Citizens' evaluation of local unemployment and the relative historical performance of the municipality, i.e, *a municipality's unemployment rate minus last years rate.*

Figure 2: Municipal Unemployment Rate Experiment: Randomly Assigned Relative Performance

the reference point as compared to above the reference point. For relatively good grades (i.e. school grades above the reference point) the effect of an extra grade point is 5 points. However, for relatively bad grades (i.e. school grades below the reference point) an extra grade point affects evaluations with about 9 points. This difference in slopes is substantial and significant (cf. table 5, column I). It supports the hypothesis of a negativity bias: a worsening af performance relative to the reference point has a stronger effect on evaluations than a relative improvement of performance beyond the reference point. For the case of the historical reference point the findings are more mixed. There is still a tendency for the slope in the positive domain to be less steep than in the negative domain. However, there is weaker evidence of a negativity bias with a difference in slopes of only 2.4 points. Next, we turn to the unemployment experiment reported in figure 2. Here negative values indicate that the municipality's unemployment rate is lower than the reference point. In the social comparison case the difference in slopes is close to zero and insignificant. In the historical case there is also no clear evidence of a negativity bias. In summary, we find mixed evidence of a negativity bias with support in two of the four treatment groups.

Finally, a note on the functional form for how relative performance affects citizens' evaluations. Looking at the graphs, marginal effects of relative performance seem larger closer to where relative performance is zero. That is, where the absolute performance of the organization is equal to the reference point. In the unemployment case the graphs show a jump in evaluations as relative performance crosses zero. In the grade experiment there are clear s-shaped curves. Changes in relative performance seem more important closer to the reference point than further aways from it. Interestingly, this decreasing sensitivity to relative performance as the difference in performance increases is in line with the prediction of decreasing marginal effects in prospect theory (Kahneman and Tversky 1991). It underscores the importance of comparisons as a strong predictor of citizens' evaluation of performance when the organization's performance is closer to the reference point.

Table 6: Negativity Bias

	Experiment and Treatment Groups:			
	School Grade Average		Local Unemployment Rate	
	B: Social ref.	C: Historical ref.	B: Social ref.	C: Historical ref.
Relative Performance	9.25^{***}	6.09^{***}	-4.63^{***}	-1.76^{*}
	(0.85)	(0.86)	(0.87)	(0.87)
Absolute > Ref. Point	6.61^{***}	6.52^{***}	-8.62^{***}	-8.96^{***}
	(1.73)	(1.73)	(1.72)	(1.87)
Interaction	-4.20^{***}	-2.40^{*}	-0.61	-1.63
	(1.24)	(1.19)	(1.25)	(1.31)
Constant	54.07^{***}	50.25^{***}	54.30^{***}	57.74^{***}
	(1.16)	(1.19)	(1.16)	(1.24)
Observations	1,148	1,139	1,142	1,146
R^2	0.35	0.24	0.27	0.14
Adjusted R^2	0.35	0.24	0.27	0.14
Residual Std. Error	17.61	17.38	17.49	18.97
F Statistic	204.47^{***}	122.42^{***}	139.42^{***}	63.69^{***}

Note: OLS estimation. Std. errors in parentheses. $^{*}p < 0.05$;$^{**}p < 0.01$;$^{***}p < 0.001$.

DISCUSSION AND CONCLUSION

More than seventy five years ago Herbert Simon (1937) argued for the importance of relative performance evaluation–comparing absolute performance to either the past performance of an organization or to the performance of other organizations. We have argued that Simon's view is important for understanding how citizens make up their mind about the performance of public organizations when faced with absolute numerical performance information. Across two experimental vignettes embedded in a large nationally representative sample of citizens, we find evidence that both vindicates and extends on Simon's view of relative performance information:

First, we find strong support for the notion that citizens' evaluation of an organization is affected not only by absolute performance but also the performance relative to an available reference point. When faced with randomly assigned absolute performance of a public organization along with an absolute measure of a historical or social reference point, citizens will tend to base their judgment on the difference between the absolute performance of the target organization and the reference point. That is, when faced with two absolute performance measures citizens are likely to compute their difference in order to arrive at a relative measure of performance discrepancy. Performance evaluation has a fundamental relative component (James 2011a; James and Moseley 2014).

Second, the analysis showed that not all reference points are of equal importance. Specifically, social reference points were found to affect citizens' evaluation to an even greater extent than historical reference points. This was both the case when the social reference point was local (i.e., other schools in the same municipality) as well as when the social reference point was national (i.e., the national unemployment rate). In both cases, the relative social reference point effect was almost twice as large as the effect for the historical reference point. This finding is in line with experimental results from the US where (some) social reference points also have had more profound effects on evaluations than historical ones (Charbonneau and Van Ryzin 2015).

Finally, the analysis found mixed evidence of a negativity bias in citizens' evaluations of relative performance. In two of the four experimental conditions, we found a much larger marginal effect on evaluations if performance was worse than the reference point. This supports the idea of a negativity bias in citizens' evaluations (Rozin and Rozyman 2001). It confirms the impression that we have yet to fully specify under which conditions a negativity bias is present (Olsen 2015). Interestingly, there was also some indication of a decreasing sensitivity to relative performance further away from the reference point. In addition, citizens' evaluation of performance was stronger in close vicinity of the reference point – in particular in the unemployment case. This speaks to the importance of the reference point as a fundamental yardstick for coding absolute performance as either good or bad depending on it being above or below the reference point (March and Simon 1958).

The results presented here are not without limits. In a real world setting, citizens will be exposed to multiple reference points when evaluating an organization. Often there will be multiple opportunities for various simultaneous historical and social comparisons. What if an organization has improved it's performance compared to last year but is still below in a salient social comparison? Maybe citizens' judgment will be more ambivalent and characterized by mixed feelings (Kahneman 1992). The present study confronted citizens with various reference points in isolation and compared their relative impact. Future studies should seek to disentangle the effects of multiple reference points and move the question of reference points into venues that more directly mimic the complexities of political-administrative settings. This being said, a survey experiment as presented here is an important first step to provide some initial support for Simon's propositions about the importance of relative performance evaluation.

The findings offer three broader implications for public managers and future research: First, the findings contribute to the almost century old ambition of providing an experimental foundation for how citizens' are affected by the presentation of performance information (Upton 1915; Freeman 1954; Williams 2003). While this effort at first hand may seem like a technical detail, it is in fact the very core of the accountability role that performance data often is intended to serve: Enhancing this role requires a deep empirical understanding of how the public-at-large makes up their mind about public services when confronted with performance data (Lee 2006a; 2006b). The findings here stress that the selection of reference points for benchmarking is fundamental to how citizens are affected by performance data. Reference points enduce a shift in how citizens evaluate the absolute performance of an organization. Reference points also give performance data a potent effect on citizens evaluations without increasing the time or effort spend. It is thus one of the key variables which we should focus on as we progress to understand the various dimensions of performance data which affect citizens' evaluation of public services.

Second, from a research stand point the findings at the citizen-level raises important questions about downstream effects on how managers and policy makers can shift reference points in order to change the perception among citizens' of how well their organization is performing (Moynihan 2008: 107–109). Early on, Simon and Ridley (1938: 467) noted this in their discussion of municipal reporting: "Figures don't lie, but liars do figure, and the citizen has very little defense in this field, as in other realms of reporting and advertising, against deliberate deception or misinformation". With citizens' reliance on social comparisons, public managers are incentivized to keep up with the performance that citizens observe in other organizations (Hansen et al. 2014). The natural research question then becomes if managers aim to affect the reference point against which their own performance is compared against? One interpretation of the findings would be that social reference points offer a better opportunity for framing current performance than historical reference point do. Managers will have a harder time forming their own history of performance than they will have pointing to other poorly performing organizations which may shed a more positive light on the manager's own organization. In addition, from the negativity bias in social comparisons, we may expect that managers aim for downward comparisons, i.e., to point to the performance of organizations doing worse than their own. Taken together, the findings provide a micro-level foundation for how citizens respond to relative comparisons which in turn can help generate hypotheses about how managers and policy makers respond to and affect the presentation of their own organization's performance. Here lies a potential for understanding more subtle ways of gaming and manipulating performance data than we traditionally have thought of (Smith 1995; Kelman and Friedman 2009).

Third, the results provide a potential explanation for many puzzles concerning both the difficulty of increasing citizen satisfaction and the weak correlation between objective and subjective measures of performance (Parks 1984; Boyne 2003). If performance evaluations have a strong relative component, then the provision of performance information can induce a form of 'hedonic treadmill' as has been found for perceptions of personal wealth and happiness (Brickman 1971). If evaluations are relative then satisfaction with service remain constant as long as everyone improves on a given metric. Moreover, the strong social comparison effect might also help explain why objective measures of performance can show improvements without being reflected in subjective perceptions of performance at the citizen level (Van Ryzin 2004). For instance, large increases in absolute objective performance will only partly affect subjective evaluation of the organization as long as the social reference point also has improved it's performance.

Taken together, we now know that citizens' evaluation of an organizations from performance data is fundamentally relative. Citizens base their judgment on a comparison between absolute performance and an available reference point. This is particular true if the reference point provides the performance of other similar organizations. Citizens' performance evaluation is fundamentally a relative process.

References

Albert, Stuart. 1977. Temporal comparison theory. *Psychological Review* 84(6): 485–503.

Ammons, David N. 1999. A proper mentality for benchmarking. *Public Administration Review* 59: 105–9.

Ammons, David N., and Dale J. Roenigk. Benchmarking and Interorganizational Learning in Local Government. *Journal of Public Administration Research and Theory* 25(1): 309-335.

Ansolabehere, Stephen, Marc Meredith, and Erik Snowberg. 2013. Asking About Numbers: Why and How. *Political Analysis* 21(1): 48–69.

Askim, Jostein, Aage Johnsen, and, Knut-Andreas Christophersen. 2008. Factors behind organizational learning from benchmarking: Experiences from Norwegian municipal benchmarking networks. *Journal of Public Administration Research and Theory* 18(2): 297–320.

Boyne, George A. 2003. What is public service improvement? *Public Administration* 81(2): 211–227.

Boyne, George A., Oliver James, Peter John, and Nicolai Petrovsky. 2009. Democracy and Government Performance: Holding Incumbents Accountable in English Local Governments. *Journal of Politics* 71(4): 1273-1284.

Brickman, Philip, and Donald T. Campbell. 1971. Hedonic relativism and planning the good society. In Adaptation-level theory: A symposium, ed. M. H. Apley. Michigan: Academic Press, 287–305.

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is Stronger than Good. *Review of General Psychology* 5(4): 323–370.

Baekgaard, Martin and Soeren Serritzlew. 2015. Interpreting performance information: Motivated reasoning or unbiased comprehension. *Public Administration Review*.

Chapman, D. W., and J. Volkmann. 1939. A social determinant of the level of aspiration. *The Journal of Abnormal and Social Psychology* 34(2): 225–238.

Charbonneau, tienne, and Gregg G. Van Ryzin 2015. Benchmarks and Citizen Judgments of Local Government Performance: Findings from a survey experiment. *Public Management Review* 17(2): 288–304.

Cyert, Richard M., and James G. March. 1963. *A behavioral theory of the firm* Englewood Clis, New Jersey: Prentice-Hall International Series in Management and Behavioral Sciences in Business Series.

Crosby, Faye. 1976. A model of egoistical relative deprivation. *Psychological review* 83(2): 85–13.

Freeman, Roger A. 1954. How can voters understand? *National Municipal Review* 43(3): 123–129.

Festinger, Leon. 1954. A theory of social comparison processes. *Human Relations* 7(2): 117–140.

Foltin, Craig. 1999. State and Local Government Performance: It's Time to Measure Up! *The Government Accountants Journal* 48: 40–46.

Hansen, Kasper M., Asmus Leth Olsen, and Mickael Bech. 2014. Cross-national Yardstick Comparisons: A Choice Experiment on a Forgotten Voter Heuristic. *Political Behavior*.

Heath, Chip. 1999. Goals as Reference Points. *Cognitive Psychology* 38(1): 79–109.

Hibbs, Douglas A. 1982. On the Demand for Economic Outcomes: Macroeconomic Performance and Mass Political Support in the United States, Great Britain, and Germany. *Journal of Politics* 44(2): 426–462.

Goethals, George R., and John M. Darley. 1977. Social comparison theory: An attributional approach. In: Social comparison processes: Theoretical and empirical perspectives, eds. J. M. Suls, and R. L. Miller. Washington, DC: Hemisphere. 259–278.

Goodman, Paul, and Emily Haisley. 2007. Social comparison processes in an organizational context: New directions. *Organizational Behavior and Human Decision Processes* 102(1): 109–125.

Greve, Henrich R. 1998. Performance, Aspirations, and Risky Organizational Change. *Administrative Science Quarterly* 43(1): 58–86.

James, Oliver. 2011a. Managing Citizens Expectations of Public Service Performance: Evidence from Observation and Experimentation in Local Government. *Public Administration* 89(4): 1419-1435.

James, Oliver. 2011b. Performance Measures and Democracy: Information Effects on Citizens in Field and Laboratory Experiments. *Journal of Public Administration Research and Theory* 21(3): 399-418.

James, Oliver, and Peter John. 2007. Public Management at the Ballot Box: Performance Information and Electoral Support for Incumbent English Local Governments. *Journal of Public Administration Research and Theory* 17(4): 567-580.

James, Oliver, and Alice Moseley. 2014. Does Performance Information about Public Services Affect Citizens' Perceptions, Satisfaction, and Voice Behaviour? Field Experiments with Absolute and Relative Performance Information. *Public Administration* 92(2): 493–511.

Kahneman, Daniel, and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 47(2): 263–291.

Kahneman, Daniel, J. L. Knetsch, and R. H. Thaler. 1991. Anomalies: The endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives* 5(1): 193–206.

Kahneman, Daniel. 1992. Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes* 51(2): 296–312.

Kayser, Mark A., and Michael Peress. 2012. Benchmarking across Borders: Electoral Accountability and the Necessity of Comparison. *American Political Science Review* 106(03): 661–684.

Kelman, Steven, and John Friedman. 2009. Performance Improvement and Performance Dysfunction: An Empirical Examination of Distortionary Impacts of the Emergency Room Wait-Time Target in the English National Health Service. *Journal of Public Administration Research and Theory* 19(4): 917–46.

Lau, Richard R. 1982. Negativity in Political Perception. *Political Behavior* 4(4): 353–377.

Lee, Mordecai. 2006a. Empirical experiments in public reporting: Reconstructing the results of survey research, 19411942. *Public Administration Review* 66(2): 252-262.

Lee, Mordecai. 2006b. The history of municipal public reporting. *International Journal of Public Administration* 29(46): 453-476.

Leibenstein, H. 1950. Bandwagon, Snob, and Veblen Eects in the Theory of Consumers' Demand. *The Quarterly Journal of Economics* 64(2).

March, James G., and Herbert A. Simon. 1958. *Organizations*. New York: Wiley

March, James G. 1988. *Introduction: A Cronicle of speculations about organizational decision-making*. Oxford, UK.

Markowitz, Harry. 1952. The utility of wealth. *The Journal of Political Economy* 60(2): 151–158.

Marvel, John D. 2015. Unconscious Bias in Citizens Evaluations of Public Sector Performance. *Journal of Public Administration Research and Theory*.

Meier, Kenneth J., Nathan Favero, and Ling Zhu. 2015. Performance Gaps and Managerial Decisions: A Bayesian Decision Theory of Managerial Action. *Journal of Public Administration Research and Theory*.

Moynihan, Donald P. 2008. *The Dynamics of Performance Management: Constructing Information and Reform*. Washington, D.C: Georgetown University Press

Moore, Don A. 2007. Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes* 102(1): 277–291.

Mussweiler, Thomas. 2003. Comparison processes in social judgment: mechanisms and consequences. *Psychological review* 110(3): 472–489.

Mussweiler, Thomas, and K. Epstude. 2009. Relatively fast! Efficiency advantages of comparative thinking. *Journal of Experimental Psychology: General* 138(1): 1–21.

Mussweiler, Thomas, and A. C. Posten. 2011. Relatively certain! Comparative thinking reduces uncertainty. *Cognition* 122: 236–240.

Nielsen, Paul A., and Martin Baekgaard. 2015. Performance information, blame avoidance, and politicians' attitudes to spending and reform: Evidence from an experiment. *Journal of Public Administration Research and Theory* 25(2): 545–570.

Olsen, Asmus L. 2013. Leftmost-digit-bias in an enumerated public sector? An experiment on citizens judgment of performance information. *Judgment and Decision Making* 8(3): 365-371.

Olsen, Asmus L. 2015. Citizen (Dis)Satisfaction: An Equivalence Framing Study. *Public Administration Review* 75(3): 469–478.

Upson, L. D. 1915. The value of municipal exhibits. *National Municipal Review* 4(1): 65–69.

Parks, Roger B. 1984. Linking objective and subjective measures of performance. *Public Administration Review* 44(2): 118-127.

Pierce, Walter M. 1940. Climbing on the Bandwagon. *The Public Opinion Quarterly* 4(2): 241–243.

Robinson-Whelen, Susan, and Janice Kiecolt-Glaser. 1997. The Importance of Social Versus Temporal Comparison Appraisals Among Older Adults. *Journal of Applied Social Psychology* 27(11): 959–966.

Rozin, Paul, and Edward B. Royzman. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* 5(4): 296–320.

Ridley, Clarence E. 1937. Annual appraisal of municipal reports. *National Municipal Review* 26(1): 31–35.

Ridley, Clarence E., and Herbert A. Simon 1940. The citizen looks at his local government. *Social Education*, February edition.

Runciman, Walter G. 1961. Problems of Research on Relative Deprivation. *European Journal of Sociology* 2(2): 315–323.

Salmon, Pierre. 1987. Decentralisation as an Incentive Scheme. *Oxford Review Of Economic Policy* 3(2): 24–42.

Samuelson, William, and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1(1): 7–59.

Shibutani, Tamotsu. 1955. Reference Groups as Perspectives. *American Journal of Sociology* 60(6): 562–569.

Soroka, Stuart N. 2006. Good News and Bad News: Asymmetric Responses to Economic Information. *Journal of Politics* 68(2): 372-385.

Smith, Peter. 1995. On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18(2): 277-310.

Simon, Herbert. A. 1937. Comparative statistics and the measurement of efficiency. *National municipal review* 26(11): 524–527.

Simon, Herbert A. 1938. Inter-city contets. *Municipal Year Book*: 210–216.

Simon, Herbert. A. 1939. The administrator in search of statistics. *Public Management* 21: 106–109.

Simon, Herbert. A. 1955. A behavioral model of rational choice. *The quarterly journal of economics*: 99–118.

Simon, Herbert. A., and C. E. Ridley 1938. Trends in Municipal Reporting. *The Public Opinion Quarterly* 2(3): 465–468.

Sorokin, Pitirim. 1925. *Sociology of revolution*. Philadelphia: Lippincott

Stark, Oded, and J. E. Taylor. 1991. Migration incentives, migration types: The role of relative deprivation. *The economic journal* 101(408): 1163–1178.

Tversky, Amos, and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211 (4481): 453–458.

Tversky, Amos, and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics* 106(4): 1039–1061.

Van Ryzin, Gregg G. 2004. Expectations, performance, and citizen satisfaction with urban services. *Journal of Policy Analysis and Management* 23(3): 433–448.

Van Ryzin, Gregg G. 2013. An Experimental Test of the Expectancy-Disconfirmation Theory of Citizen Satisfaction. *Journal of Policy Analysis and Management* 32(3): 597–614.

Weaver, R. Kent. 1986. The Politics of Blame Avoidance. *Journal of Public Policy* 6(4): 371-398.

Williams, Dan W. 2003. Measuring government in the early twentieth century. *Public Administration Review* 63(6): 643–659.

Willimott, J. F. (1924). Public reports and public opinion. *National Municipal Review* 13(8): 421–424.

Yockey, M. D. and Kruml, S. M. 2009. Everything is Relative, but Relative to What? Defining and Identifying Reference Points. *Journal of Business and Management* 15: 95–109.

Appendix A

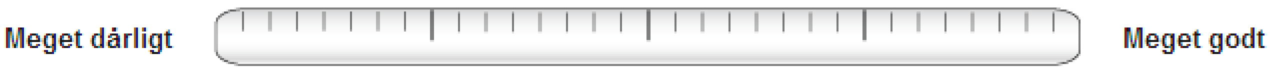


Figure 3: Screen caption of the exact response scale used in the experiment. The scale varies from ”very bad” (”meget dårligt”, 0) to ”very good” (”meget godt”, 100).

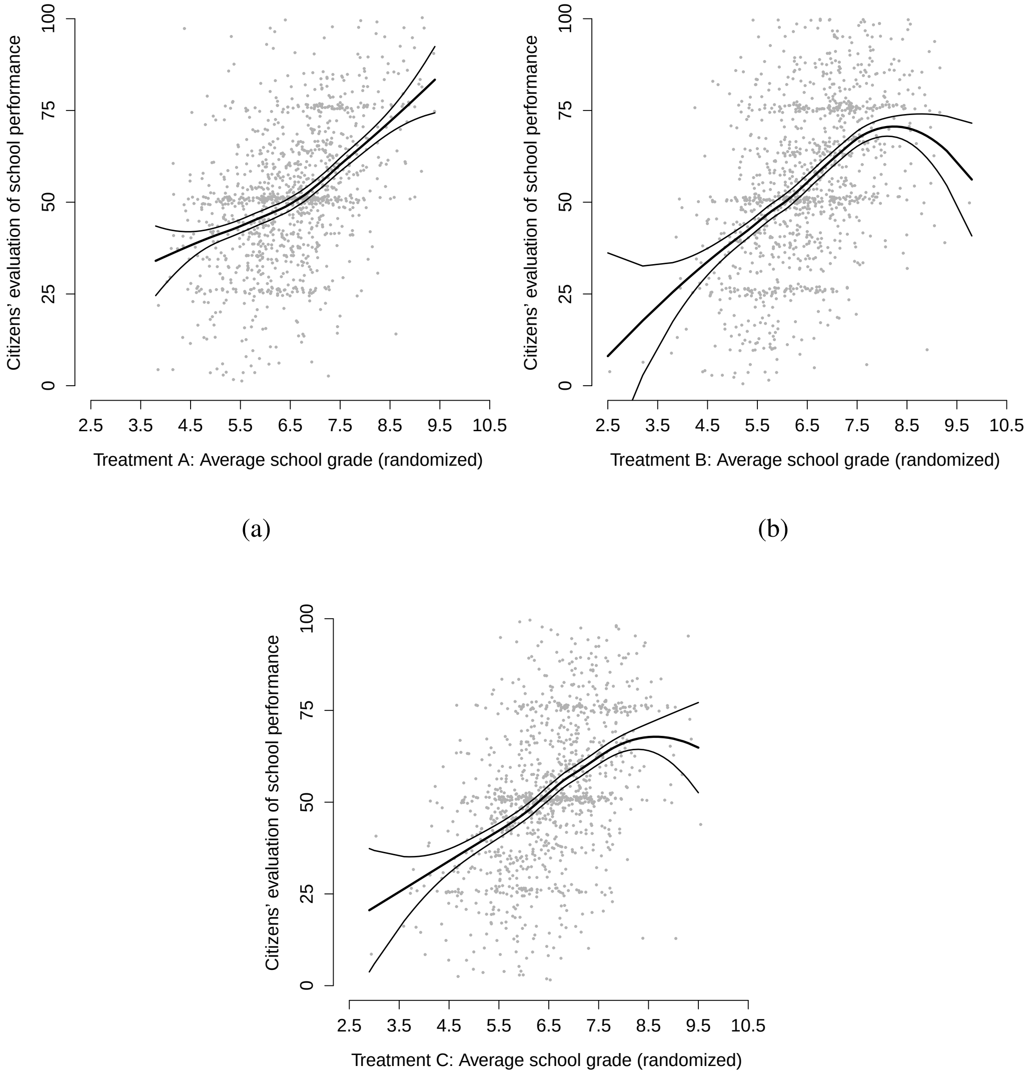
Appendix B

Table 7: Randomization Check across Experiment and Treatment Groups

	Experiment and Treatment Groups:					
	School Grade Average:			Municipal Unemployment Rate:		
	Absolute	Social	Historical	Absolute	Social	Historical
Absolute performance	6.5	6.5	6.5	6.4	6.4	6.3
Ref. point performance	–	6.5	6.5	–	6.3	6.3
Relative performance	–	0.01	0.03	–	0.06	-0.02
Gender (%)	50.5	50.2	48.6	51.9	48.8	48.5
Age (years)	50.1	50.5	50.1	50.7	50.3	50.5
Left-wing voter, 2011 (%)	53.6	51.8	54.5	54.3	50.8	54.8
Education	4.7	4.8	4.9	4.8	4.8	4.8
Private sector emp. (%)	31.9	34.5	31.4	33.3	33.5	31.0
Copenahgen area (%)	26.3	24.0	23.6	24.0	25.4	24.7

Note: Means for each experimental group.

Appendix C

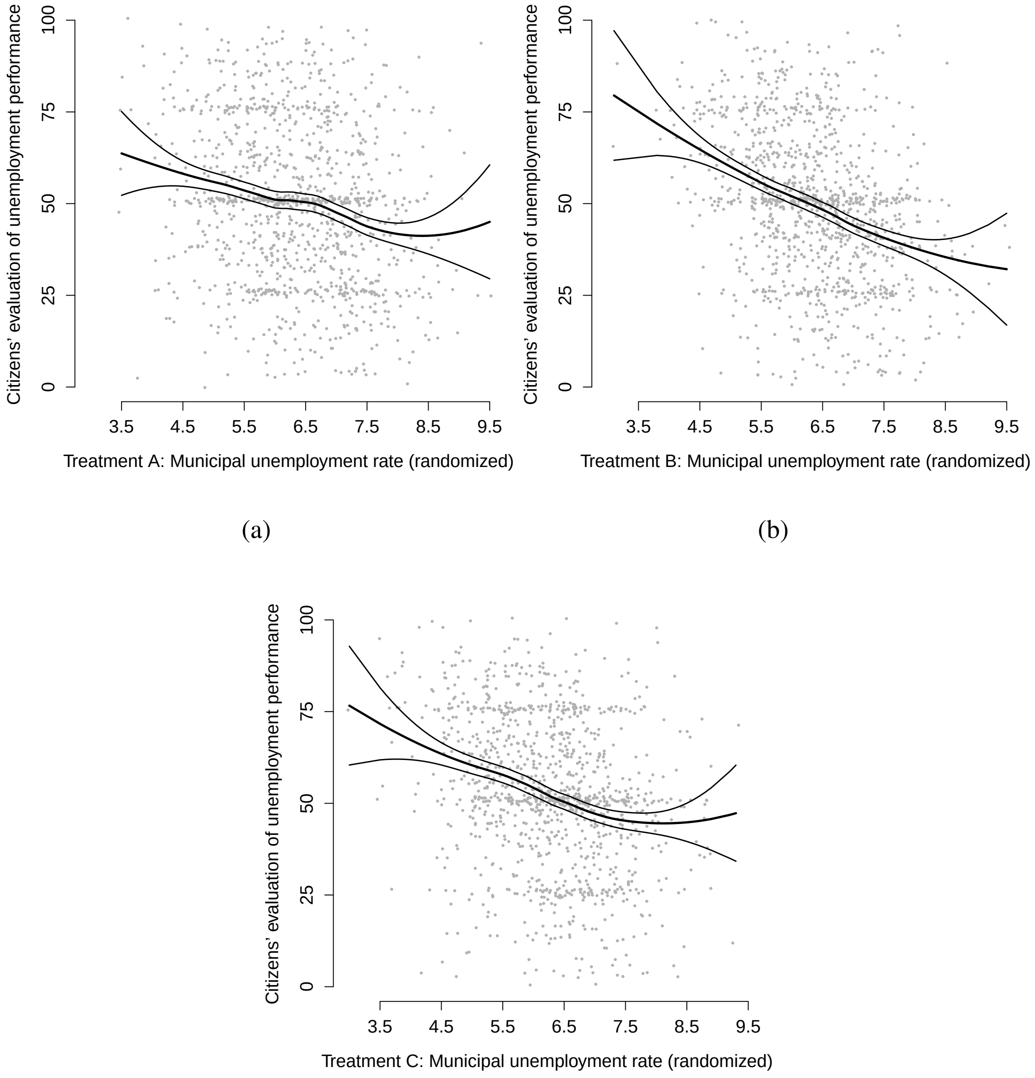


(a)

(b)

(c)

Figure 4: School grade experiment: the plots show the correlation between the absolute grade for the target school under the three treatment frames and the evaluation provided by the citizens. The flexible fit is a lowess estimation.



(a)

(b)

(c)

Figure 5: Unemployment rate experiment: the plots show the correlation between the unemployment rate for the target municipality under the three treatment frames and the evaluation provided by the citizens. The flexible fit is a lowess estimation.