The Effect of Peer Comparisons on Polluters:

A Randomized Field Experiment among Wastewater Dischargers

Dietrich Earnhart
Department of Economics
University of Kansas
earnhart@ku.edu

Paul J. Ferraro
Carey Business School and the Department of Environmental Health and Engineering, a joint department of the Bloomberg School of Public Health and the Whiting School of Engineering
Johns Hopkins University
pferraro@jhu.edu
ORCID: 0000-0002-4777-5108

September 30, 2020 [Accepted version]
Environmental & Resources Economics version at https://doi.org/10.1007/s10640-020-00522-0

Abstract

Peer comparisons combine descriptive and injunctive messages about social norms. In experiments, these comparisons have encouraged pro-environmental behaviors among consumers. Consumers, however, are not the only sources of environmental externalities. Firms and other organizations also damage the environment. Yet organizations may not respond to peer comparisons in the same way that consumers respond because organizations have different objectives, constraints, and decision-making processes. In a pre-registered field experiment with 328 municipal wastewater treatment facilities in Kansas, we randomly sent some facilities a certified letter that contrasted, using text and a graphic, each facility's discharge behavior to the behaviors of other facilities in the state. We estimate the effect of these peer comparisons on the degree to which the recipient facilities complied with discharge limits under the U.S. Clean Water Act. On average, letter recipients reported discharge ratios 8 % lower than non-recipients in the eighteen-month period after letters were sent (95 % CI [-15 %, -1 %]), although we cannot detect an effect in all post-treatment quarters. We believe that the results warrant further experimental replications and extensions to examine the cost-effectiveness of reducing pollution through peer comparisons.

Keywords: randomized controlled trial, regulatory compliance, social comparisons, nudges, voluntary approaches, pollution, environmental management

1. Introduction

To protect and improve the natural environment, regulators typically depend on a mix of mandatory and voluntary approaches. Proponents of voluntary approaches have argued that these approaches can be more cost-effective and less conflict-ridden than mandatory approaches, particularly in contexts where environmentally-damaging behaviors are hidden from regulators. Yet skeptics of voluntary approaches question their effectiveness.

Recently, a growing movement has highlighted a new breed of voluntary approaches to help achieve environmental policy goals (Byerly *et al.*, 2018; Klotz *et al.*, 2019; Palm-Forster *et al.*, 2019) and other social policy goals (Thaler and Sunstein, 2008; Benartzi *et al.*, 2017). This movement, with its focus on "nudges" and "choice architecture," draws its insights from social psychology, behavioral economics, and other behavioral sciences. Proponents argue that self-enforcing, voluntary behavioral change can be achieved cost-effectively by leveraging attributes of human decision-making that deviate from the traditional neo-classical economics model of human behavior. These attributes include so-called behavioral "anomalies", such as loss aversion, probability weighting, emotions, pro-social preferences, norms, and cognitive constraints.

We study one increasingly popular approach that is inspired by theories from social psychology: providing peer comparisons to decision-makers (Festinger, 1954; Schultz *et al.*, 2018). Peer comparisons, which are also called social comparisons, combine injunctive and descriptive messages about social norms to encourage behavior change towards the norms. In the context of environmental protection, the injunctive message describes the socially desired level of an environmental behavior (e.g., low water use). The descriptive message describes how the decision-maker's behavior compares to the behavior of peers (e.g., proportion of peers that use

less water). In randomized field experiments, peer comparisons have been reported to change environmentally-relevant behaviors (see references in Section 3).

The behavioral impacts of peer comparisons are modest, but they are also inexpensive. To construct the comparisons, program administrators draw on already available, or easily acquired, data. To deliver the comparisons, administrators use postal or digital delivery. Given the strength of the evidence base, some regulators permit the use of peer comparisons as energy demand management tools (e.g., Oracle's OPower product). If peer comparisons were equally effective in affecting the behaviors of polluters, they could serve as complements to pollution regulations by encouraging noncompliant polluting facilities to comply with their regulatory limits and by encouraging compliant facilities to reduce their emissions below their regulatory limits.

Yet the empirical evidence for the effectiveness of peer comparisons comes almost entirely from the context of consumers. Given that the objectives, constraints, and decision-making processes of consumers differ from those of polluting facilities, the behavioral effects of peer comparisons may also differ. Few experimental field studies estimate peer comparison impacts on behavior outside of the consumer context. Most of these studies focus on doctors (e.g., Reeves, 2012; Sacarny *et al.*, 2016), with a couple of studies focusing on farmers (Wallander, Ferraro and Higgins, 2017; Chabé-Ferret *et al.*, 2019). The evidence in favor of peer comparisons changing behavior in these experiments is mixed, at best.

However, the notion that peer comparisons can affect the behavior of polluting facilities is plausible. Humans manage these facilities. Thus the findings from prior experiments with individual consumers may generalize to facility managers. Indeed, "naming and shaming" approaches in the environmental context are predicated, in part, on environmental managers

expressing social preferences.[1]

On the other hand, the humans that make decisions in these facilities typically make repeated, high-stakes decisions (e.g., pollution abatement investments), often embedded in a larger organizational structure (i.e., group decision-making) and often subject to pressures to keep costs low. These same attributes have been reported in lab and field experiments to induce behavior that is more consistent with the conventional, neo-classical economic model of self-interested, payoff-maximizing agents (see Section 2 for more details). Thus, whether the results of prior studies of peer comparisons in the environmental context generalize to polluting facilities is an open question.

To help scholars and practitioners begin to answer this question, we ran a randomized field experiment with a registered, pre-analysis plan.[2] With a sample of 328 municipal wastewater treatment facilities in Kansas, we randomly assigned facilities to a treatment group or a control group. Facilities in the treatment group received a peer comparison message, while the control group received nothing (status quo condition). Over a year later, we estimated the effect of peer comparisons on the degree to which facilities complied with discharge limits under the U.S. Clean Water Act.

On average, the peer comparisons reduced the ratio of reported discharges to permitted discharge limits by 8 % (95 % CI [-15 %, -1 %]). Nevertheless, given the modest sample size, the confidence intervals on this estimate are wide, and both the estimate and the confidence intervals are sensitive to changes in the empirical estimation strategy. We thus view these results as suggestive, rather than definitive. Our exploratory analyses do not detect any differences in

1 Such approaches may also work via extrinsic motivations mediated through investors, consumers, and other stakeholders (Konar and Cohen, 1997; Khanna, Quimio and Bojilova, 1998; Delmas, Montes-Sancho and Shimshack, 2007; Bennear and Olmstead, 2008)
2 https://osf.io/6u7t8/

behavioral responses conditional on a facility's size or compliance history. These analyses do suggest, however, that facilities whose discharges are under greater regulatory scrutiny, and thus required to report discharges monthly, may be more responsive to the peer comparisons.

Our study contributes to four literatures. First, it contributes to the large literature on voluntary approaches to environmental protection. Second, it contributes to the small literature on the effectiveness of nudges and changes in choice architecture among producers (i.e., non-consumer decision-makers). Outside of the medical or agricultural contexts, we find few peer-reviewed experimental studies of any behavioral psychology-inspired interventions in which a business or other organization is the unit of randomization. Third, our study contributes to the literature on drivers of regulatory compliance and over-compliance, which focuses on inspections and enforcement actions (Stafford, 2002; Earnhart, 2004a; Shimshack and Ward, 2005; Telle, 2009; Earnhart and Segerson, 2012) or external pressure, such as customer, investor, or community pressure (Henriques and Sadorsky, 1996; Pargal and Wheeler, 1996; Dasgupta, Hettige and Wheeler, 2000; Becker, 2004; Earnhart, 2004b; Gangadharan, 2006; Earnhart and Segerson, 2012). Our study helps to shed light on whether intrinsic motivations are also a possible driver. Finally, our study contributes to the inchoate literature using field experimental designs to estimate the effects of interventions on polluting facilities (Duflo *et al.*, 2013; Telle, 2013).

The next section describes the study context. Section 3 develops a conceptual framework. Section 4 describes the experimental design. Section 5 describes the hypotheses and econometric estimators. Section 6 reports results. Section 7 discusses the results. Section 8 concludes.

2. Study Context

Our study population comprises municipal wastewater treatment facilities, which are permitted within the U.S. Clean Water Act's (CWA) National Pollutant Discharge Elimination System (NPDES). The CWA seeks to protect and restore surface water quality, primarily by controlling wastewater discharges from facilities (i.e., point sources). Each regulated facility is issued a NPDES permit, which describes the facility's pollutant-specific discharge limits. Each limit is based on the stricter of two possible limits: a limit for a given industry, based on Effluent Limitation Guideline standards, or a limit based on water quality, which ensures the receiving waterbody meets ambient water quality standards. Given spatial and temporal variation in ambient water quality, discharge limits differ across facilities and time. To comply with these limits, facilities can use any abatement method.

Facilities are required to monitor their discharges on a regular basis and report them to regulatory agencies in the form of Discharge Monitoring Reports (DMRs). Facilities submit DMRs monthly or quarterly. In Section 6, we describe in more detail the reasons why some facilities are required to submit monthly rather than quarterly; in general, agencies require monthly reporting from facilities that are generating large or particularly polluted volumes of discharge, discharging into sensitive water bodies, operating their treatment systems near their design limits, or failing to achieve compliance. To monitor and enforce compliance with the CWA, the U.S. Environmental Protection Agency (EPA) and authorized state agencies inspect facilities and take enforcement actions against non-compliant facilities. These inspections, however, are not designed to verify the DMRs. Inspectors generally do not take water samples; if they do take samples, they commonly measure the concentration of pollution in a discharge stream only over a short period of time, such as a single day. Very few samples cover a month of

discharges. Rather than using inspections for verification, agencies primarily use inspections as a vehicle for gathering evidence for future enforcement actions (Wasserman, 1984) and offering compliance assistance (Earnhart, 2004b).

The NPDES distinguishes between major facilities and minor (or "non-major") facilities. Major facilities tend to be larger and discharge more wastewater. In our sample, major facilities are required to submit DMRs monthly. Federal guidelines prompt EPA and state agencies to scrutinize major facilities more than minor facilities (Earnhart, 2009; Earnhart and Segerson, 2012; Earnhart and Harrington, 2014). For example, guidelines dictate that major facilities are inspected once every two years, whereas minor facilities need only be inspected once every five years. In most states, state agencies conduct the inspections.

Our sample comes from the population of municipal wastewater facilities in the state of Kansas between 2015 and 2018. In Kansas, all municipal wastewater treatment facilities are, to our knowledge, owned by local governments. Regardless of ownership, facilities may be operated by governments or private contractors. (Our data do not allow us to differentiate the two types of potential operators) These facilities may handle household and industrial wastewater. They are permitted and inspected by the Kansas Department of Health and Environment. Our focus on a single sector is consistent with other empirical studies of point source pollution (e.g., Laplante and Rilstone, 1996; Earnhart, 2004a, 2004b, 2009; Earnhart and Harrington, 2014).

As the pollutant outcome measure, we selected the most commonly regulated pollutant from wastewater dischargers: biological oxygen demand (BOD), which is a measure of organic water pollutants and is one of five conventional pollutants on which the EPA has traditionally focused its efforts. Specifically, we focus on the 5-day biological oxygen demand measure, which is widely viewed as a proxy measure for the overall quality of facility discharges. This pollutant

was chosen based on feedback from an engineering consultant and two engineering professors with specialties in wastewater treatment. In addition to serving as a single, widely recognized dimension of pollution, the BOD measure can be affected by facility managers in the short-term (<1 year) via changes in treatment processes and technologies. Moreover, all wastewater treatment facilities in our sample face a BOD limit. As is true for most municipal wastewater treatment facilities operating in the US, the limits are concentration-based, rather than quantity-based, and expressed in milligrams of BOD per liter (mg/L) of effluent. A concentration-based limit, which is independent of facility size, allows for easy comparability across facilities and over time.

To construct a measure of facility discharge behavior, we use the ratio of the facility's reported BOD discharge in its DMR to the facility's permitted BOD limit, a ratio that we call the "discharge ratio". The discharge ratio measures the extent of compliance. It captures both improvement toward compliance and improvement beyond compliance (Earnhart, 2004a, 2004b, 2009; Earnhart and Segerson, 2012; Earnhart and Harrington, 2014). We obtained data on BOD limits and reported discharges from the EPA DMR Pollutant Loading Tool (since updated and renamed the Water Pollutant Loading Tool). Facility permits place limits on the maximum and average BOD discharges. For each reporting period (e.g., month), we calculated the discharge ratio separately for the maximum limit and the average limit. Then we averaged the two discharge ratios to generate a facility-period-specific discharge ratio. See online Material and Methods (MM) for more details on BOD, our data sources, and the discharge ratio. Based on discharges reported in 2016, the most compliant facility generated a discharge ratio of 0.039; i.e., over-complying with its limits by 96.1 % [=(1-0.039)×100]. The least compliant facility generated a discharge ratio of 1.92, i.e., violating its limits by 92 % [=(1.92-1)×100]. The median

facility generated a discharge ratio of 0.438, over-complying by 56.2 %. The next section describes how this discharge ratio could be affected by peer comparisons.

3. Conceptual Framework

To guide our analysis, we develop a simple conceptual framework, which adapts a standard economic model of crime (Becker, 1968) to the case of regulatory compliance. A risk-neutral manager operates facility *i* in period *t*, and seeks to maximize expected utility, where utility, U_{it}, depends on the facility's production value, Π_{it}, so that $U_{it}=U(\Pi_{it})$. In each period *t*, the manager decides on a discharge level, D_{it}, relative to the NPDES-permitted limit, L_{it}.

Given that the limit is set by the regulatory authority, we can reframe the manager's decision problem as choosing the extent of compliance, $C_{it} = D_{it} / L_{it}$, where C_{it} is the discharge ratio from Section 2. In each period, the manager pays abatement costs of *k*, which are decreasing in C_{it}. If the manager exceeds the regulatory limit, L_{it}, the regulator has the authority to impose a fine, *F*. The probability of receiving a fine is denoted as *p*. Both F and p are weakly increasing in C_{it}. The manager chooses C_{it} to maximize expected utility: $E[U]= \Pi - k(C) - p(C)*F(C)$.

This model captures the traditional perspective on environmental enforcement, in which the regulator can induce better regulatory compliance only by increasing *p* or *F*. Moreover, the model predicts that, with $k > 0$, a firm never chooses a value of $C < 1$. Yet, $C < 1$ is the norm, rather than the exception, in the NPDES context (Earnhart and Segerson, 2012). In our sample of Kansas wastewater dischargers, 324 out of 328 facilities had a pre-treatment discharge ratio less than 1.

To explain this norm, scholars have extended this model in ways that highlight other factors that may influence a manager's compliance decision. In one research thread, scholars have

identified other contextual factors to add to the model, like stochastic discharges and non-regulatory external pressures. For example, a facility's discharges are typically subject to random fluctuations, due to weather conditions, human error, and equipment failure, among other factors. Earnhart and Segerson (2012) develop a theory of a manager's abatement choice when discharges are stochastic. When facing stochastic discharges, managers may choose to over-comply to insure against potential penalties driven by randomly high discharges. In another research thread, scholars have suggested alternative behavioral models that incorporate pro-social preferences or include the costs of acquiring information and paying attention to this information. For example, a large body of research suggests that some humans are pro-social (e.g., altruists, reciprocators, conformists) or boundedly rational (e.g. inattentive, costly information acquirers). The simplest way to expand our model to account for pro-social behaviors would be to add a "moral cost to pollution" in the decision problem (Ferraro and Price, 2013). Depending on its functional form, this cost may mimic conformism with a reference level of behavior, warm glow altruism, pure altruism, or reciprocal altruism. Two simple ways to incorporate bounded rationality in the model are: (1) add a cost function that captures the costs of acquiring information about the pollution abatement cost function, or (2) add noise to the model (akin to the stochastic discharge model of Earnhart and Segerson, 2012).

Scholars argue that pro-sociality and cognitive boundedness can be leveraged to cost-effectively improve policy and program outcomes (Benartzi *et al.*, 2017). In other words, the same factors that help explain $C < 1$ can be leveraged to push C even lower. One popular way of leveraging these attributes is through the delivery of peer comparisons. In the context of polluting facilities, a peer comparison combines an injunctive norm, which describes the socially desired behavior of low discharge ratios, with a descriptive norm, which identifies how the

facility's discharge ratio compares to the ratios of the facility's peers. Figure 1 shows the mechanisms through which these comparisons can affect behavior. Although the magnitudes of these mechanism effects are debated (Ferraro and Miranda, 2013), peer comparisons are reported to affect individual (consumer) decision-making in a wide range of contexts, including environmental contexts such as energy use, water use, and recycling (e.g., Allcott, 2011; Ferraro, Miranda and Price, 2011; Ayres, Raseman and Shih, 2013; Ferraro and Price, 2013; Allcott and Rogers, 2014; Bernedo, Ferraro and Price, 2014; Henry, Ferraro and Kontoleon, 2019).[3]

Yet, as noted in the Introduction, whether peer comparisons similarly affect the compliance decisions of polluting facility managers is unknown. In contrast to "naming and shaming" approaches, which make explicit or implicit *public* comparisons, the peer comparison intervention is a *private* message, targeted at intrinsic motivations of facility managers. Nakamura et al. (2001) report that managerial attitudes towards environmental protection are correlated with facilities' management choices, a result that is consistent with intrinsic motivations in the form of pro-social preferences. Other studies report that the regulator's enforcement style affects environmental compliance (Winter and May, 2001; Short and Toffel, 2010; Earnhart and Glicksman, 2015), a result that is consistent with reciprocating preferences. Moreover, our experience in this sector suggests that professional norms influence facility managers who take pride in their performance. Facility managers may also extract information about optimal facility management practices from a peer comparison, a response that is consistent with the Porter Hypothesis (Porter, 1991). This hypothesis posits that flexible environmental policies may prompt regulated facilities to discover previously unexploited opportunities for cost reductions or revenue improvements, thereby improving facility profitability (Gabel and Sinclair-Desgagné, 1998, 2001). In sum, peer comparisons could

3 Not all studies, however, have detected an effect (e.g., Chong et al., 2015; Andor et al., 2018).

plausibly affect facility manager behavior via intrinsic motivations (i.e., via information constraints or non-monetary attributes in the utility function of a facility manager).

[**Figure 1. Potential Mechanism through which Peer Comparisons could Affect Pollution Discharges** here]

On the other hand, in contrast to the average consumer, facility managers are technical experts, who may make decisions in groups (e.g., with co-workers, consultants or municipal public works directors) and are subject to high-stake decisions and cost pressures, particularly for decisions like investments in pollution abatement processes. In laboratory and field experiments, these attributes correlate with behaviors that appear closer to predictions from traditional economic models (Smith and Walker, 1993; List, 2003; Charness and Sutter, 2012; Maciejovsky *et al.*, 2013; Alevy, Landry and List, 2015; Meub and Proeger, 2018).

Thus, whether and how peer comparisons can affect facility manager decisions is an empirical question. The next section describes an experimental design that aims to contribute to answering this empirical question.

4. Experimental Design

4.1. Treatment Message

Each facility in the treatment group received a letter that defined the discharge ratio, reported the facility's own discharge ratio, and explained how its own ratio compared to the ratios of other municipal facilities in the state. To communicate these ideas, the letter provided a visual (graphical) display and narrative text. To communicate an injunctive norm without explicitly stating that "pollution is bad", we used a "negative" framing (Ferraro and Price, 2013). We

communicated the facility's percentile as follows: "X % of Kansas municipal facilities comply with their discharge limits to a greater extent than your facility complies with your limits," where X is the facility's percentile in the distribution of discharge ratios. In the graphical display (Figure 2), the text states: "X % of all Kansas municipal facilities comply to a greater extent than your facility."

To construct the peer group, we used the entire population of municipal wastewater facilities in the state, which generates the greatest amount of variation in discharge ratios (thus, high discharge ratio facilities will clearly see they are in the tail of the distribution). Our limited sample size prevented us from experimenting with the peer group composition across multiple treatment messages. We chose to use the entire population because prior studies report behavioral impacts from peer comparisons using peer groups of heterogeneous individuals living in large geographic areas (e.g., Ferraro and Price, 2013) and no study has quantified the relative effects of variations in peer group composition versus perceived deviation from the norm (i.e., the more homogenous the peer group, the greater the "acceptability" of the peer group, but the smaller the average perceived deviation from the norm, making the overall effect on behavior ambiguous).

[**Figure 2. Peer Comparison Message** here]

The letter identified the authors as the senders, with both university logos on the letter and the University of Kansas logo on the envelope. In our analysis, we assume that the recipients do not believe that the authors have any regulatory authority or ability to influence a regulator or the media; in other words, we assume any behavioral response by a recipient does not stem from a

change in beliefs about the probability or intensity of regulatory enforcement or citizen action. Our analysis also assumes that the peer comparisons contained new information and that they were read and understood by the facility managers. To increase the likelihood that the letters reached their intended recipients, we sent the letters by certified U.S. mail, which requires a signature from a human recipient (not necessarily the person named on the envelope). To assess whether the information would be new, we relied on three sources of information: (1) our engineering consultant and academic informants; (2) our understanding of the time and knowledge required to access and process the DMR data to produce the peer comparisons; and (3) interviews with a dozen facility managers after we had collected the endline data. The information from these sources is consistent with our assumption that the peer comparisons provided new information to the facility managers (see MM for details).

4.2. Randomization

To be able to reduce discharges, a manager must oversee a facility that has discharges. Thus we aimed to select facilities that were actively discharging into surface water. From a population of 361 municipal wastewater facilities in Kansas, 328 facilities (91%) had a discharge reporting pattern between April 2015 and December 2016 that implied they were likely to report a discharge during the post-treatment assignment period (see MM for details on criteria). Our sample size of 328 facilities is similar to other pollution compliance field experiments (N=534 with a 1:4 treated:control ratio in Telle, 2013; N=473 in Duflo et al., 2013).

In April 2017, we blocked-randomized NPDES contact names to treatment and control groups. We randomized contact names, rather than facilities, because 13 contact names were associated with more than one facility; these names were associated with 28 facilities, with two

contacts associated with three facilities and the rest associated with two facilities. By randomizing at the level of the facility contact name, we aimed to avoid interference among treated and control facilities (spillovers) that could arise if two facilities associated with the same contact person were assigned to different groups. To the best of our knowledge, the contact people are not clustered within larger organizations that manage multiple facilities.

Based on analysis of the 2015 and 2016 data, we chose the following blocking covariates, measured at the facility level: the quartile of average 2016 BOD discharge ratio, a binary indicator of the facility's EPA designation (major vs minor), and a binary indicator of the facility's frequency of discharge reporting (monthly vs quarterly). For the few NPDES contact people associated with more than one facility, we used the average values of the facilities associated with the contact (rounded up for binary variables). Thus, the blocking covariates connect to a contact name.

[**Figure 3. Kansas Municipal Wastewater Treatment Facilities in Experiment** here]

In August 2017, we sent each member of the treatment group a certified letter with the peer comparison for the facility. We sent no letter to the control group. See Figure 3 for the spatial distribution of the facilities. Table 1 presents descriptive statistics for the sample overall and by treatment arm. As expected, given the block randomization, the treated and control facilities do not meaningfully differ across observable attributes.

[**Table 1. Descriptive Statistics** here]

4.3. Pre-treatment and Post-treatment Periods

The pre-treatment period is the first quarter Q1:2015 through Q2:2017. The post-treatment

period is Q3:2017 through Q4:2018. We downloaded post-treatment data in April 2019, when data for Q4:2018 became available. Given that our sample comprises a mix of monthly and quarterly reporters, we chose, as the primary analysis in our pre-analysis plan, to aggregate all reported monthly discharges to the quarterly level, taking the quarterly average of the monthly values (see MM). As a robustness check, we also estimate effects separately for the quarterly and monthly reporters.

If we expect that facilities can react immediately to the peer comparisons, we should include the entire post-treatment period in our analysis for maximum statistical power. However, if we expect that facilities need time to make process changes, we should use only the later quarters in the post-treatment period for maximum statistical power. Prior to launching the experiment, we were uncertain about the speed at which changes would be made.

In addition to estimating the treatment effect using the entire post-treatment period (Q3:2017 – Q4:2018), we also estimate the effect using a "later-quarter" definition of the post-treatment period. To avoid accusations of "cherry picking" in choosing a later-quarter definition, we pre-registered one. We would use the third quarter of 2018, which was a year after treatment assignment, and use the neighboring quarters to impute any missing third quarter observations. If a fourth quarter observation were available, we used it to impute the missing third quarter value; if not, we looked to the second quarter and then to the first quarter. If no observations were available in 2018, we left the value as missing (see pre-analysis plan for more details). Using these rules, 47 facilities had missing third quarter values that we could impute, and 32 facilities had missing post-treatment periods (i.e., no reports in the 2018 calendar year; see sub-section 5.5).

4.4. Hypotheses

Based on our conceptual framework, we pose three, pre-registered hypotheses:

Hypothesis H1 (Peer comparisons reduce reported discharge ratios): If the traditional (Becker-inspired) model in Section 3 captures facility manager decision-making, peer comparisons do not affect manager behavior, and by extension, the discharge ratio (i.e., the traditional model identifies the null hypothesis). In contrast, if facility managers have non-financial (intrinsic) motivations and the peer comparisons operate by creating or invoking a norm that "more compliance is always better than less compliance," the peer comparisons could reduce reported discharge ratios, on average (i.e., we reject the null in favor of *H1*).[4] However, the mechanisms through which the effects of peer comparisons are achieved in the consumer context are debated (see Ferraro and Miranda, 2013). Other mechanisms yield heterogeneous treatment effects that we capture in the following two hypotheses.

Hypothesis H2 (Peer comparison effects are heterogeneous because facility managers are other-regarding). If facility managers are conformists or conditional reciprocators, peer comparisons should induce facilities with historically higher discharge ratios (i.e., lower compliance), on average, to improve the extent of their compliance by an amount greater than facilities with historically lower discharge ratios (i.e., historically higher compliance). The response of facilities with historically lower discharge ratios (i.e., higher compliance) depends on how the managers perceive their peers' behavior. Higher-compliance facilities run by conforming or conditionally cooperative managers could increase their discharge ratios, a so-called "boomerang effect." In the presence of a boomerang effect, the sign of the overall average effect of peer comparisons is ambiguous (i.e., in contrast to *Hypothesis H1*, the overall effect

4 Possible intermediate mechanisms include shame for poor performance, pride for good performance, altruism for the victims of the negative externalities, or mixes of these and other intermediate mechanisms.

could be negative if the boomerang effect is stronger than the discharge-reducing effect).

Hypothesis H3 (Peer comparison effects are heterogeneous because facility managers are boundedly rational or imperfectly informed). If facility managers are boundedly rational (e.g., inattentive), or imperfectly informed because information acquisition about optimal compliance is costly, facilities may not be optimizing their compliance activities prior to treatment assignment. In such cases, peer comparisons can affect behavior by inexpensively supplying costly-to-acquire information about private costs and benefits. As noted by Ferraro and Miranda (2013), "the 'social' comparison may actually be a 'private' signal."[5] In our context, in which facilities are over-compliant and are assumed to know their own discharges and legal limits, we would expect that peer comparisons, on average, (*H3a*) induce facility managers to reduce their extent of compliance (i.e., increase discharge ratios) and (*H3b*) induce historically higher-compliance facilities by an even greater extent. We would also expect that (*H3c*) peer comparisons generate a larger treatment effect, on average, among less sophisticated, minor facilities. In contrast to major facilities, minor facilities are, on average, smaller operations with fewer wastewater engineers. They are thus more likely to glean insight from the peer comparisons offered in our treatment letters. Although we pre-registered a uni-directional hypothesis *H3c*, we recognize an alternative perspective on facility size: the more sophisticated management of the major facilities may allow them to deploy greater resources in response to the treatment. In other words, peer comparisons could prompt a larger response from major facilities (via the information channel or the pro-social behavior/norms channel).

Of course, drawing inferences about behavioral motives from average and conditional

5 We assume that the information about own-facility compliance is not new and thus has no effect on behavior (not new because it came from the facility manager). However, presentation of own-facility compliance in the form of the discharge ratio may provide new information to a cognitively bounded manager, separate from the information in the peer comparison of ratios.

(subgroup) treatment effects is challenging if facility managers (1) are heterogeneous in their objective functions or constraints, or (2) exhibit multiple behavioral attributes. From a policy perspective, however, the sign and magnitude of the average effects are relevant, regardless of the mechanisms through which these effects arise. We do not pose a hypothesis with regards to a scrutiny channel because we believe that this channel does not operate in our context because the letters came from academics rather than regulators.

5. Modes of Inference

5.1. Estimator

In our analysis, the outcome variable (dependent variable) is the natural logarithm of the i^{th} facility's discharge ratio in time period t, C_{it}. We pre-specified a natural logarithm transformation because the 2015 and 2016 pre-treatment discharge ratio data are right-skewed and the residuals are heteroskedastic; our power simulations imply this transformation yields more precise estimates (see below). About 5 % of reported discharges are "below detectable limits"; the database codes these values as zeroes. Since this proportion is small, our pre-analysis plan did not seek to address these missing values in the log transformation of the discharge ratio.

The treatment variable, Z_{it}, equals 1 during the post-treatment time period for facilities assigned to the peer comparison, and equals 0 in the post-treatment period for facilities assigned to the control group and for all facilities during the pre-treatment periods. To estimate the average treatment effect of Z_{it} on $\ln(C_{it})$, we use the following random-effects panel data estimator:

$$\ln(C_{it}) = \alpha + \beta' Z_{it} + \theta' K_{it} + \epsilon_i + \eta_{it} , \quad (1)$$

where ϵ_i is a facility-specific error term and η_{it} is a random error term. Given randomization of

the treatment, Z_{it} is uncorrelated with ϵ_i and η_{it}. K_{it} represents control variables, which comprise quarter-by-year dummy variables and the three blocking variables used in the randomization procedure. By modeling time trends flexibly, we improve the precision of the estimated treatment effect. By including the blocking variables, we ensure that our standard errors and confidence intervals are estimated correctly (Bruhn and McKenzie, 2009). We cluster standard errors at the contact-level, which means we cluster at the facility-level for all facilities except the 28 facilities that have common contact names in the NPDES database. This clustered estimation of the variance allows for heteroskedasticity across facilities and arbitrary serial correlation within each facility (Wooldridge, 2002).

If facility manager behavior were characterized by the traditional (Becker-inspired) model in Section 3, the treatment coefficient equals zero: $\beta = 0$. In contrast, *Hypothesis H1* implies that $\beta < 0$, and hypothesis *H3a* implies $\beta > 0$ (we apply two-sided tests in all of our analyses even when the alternative hypothesis is one-sided).

In order to test *Hypotheses H2* and *H3b*, we interact the treatment variable, Z_{it}, with a binary indicator of compliance history, H_i. The indicator variable distinguishes between facilities with an average discharge ratio *above* the sample median in 2016 (pre-treatment) and facilities with an average discharge ratio *below* the sample median in 2016. The indicator variable equals 1 for all facilities with an above-median historical discharge ratio and equals 0 otherwise. We estimate an extended version of equation (1):

$$\ln(C_{it}) = \alpha + \beta' Z_{it} + \theta' K_{it} + \gamma' H_i + \sigma'(H_i \times Z_{it}) + \epsilon_i + \eta_{it}\,, \qquad (2)$$

where the coefficient β now captures the treatment response for facilities with higher historical compliance (lower discharge ratios). The sum $(\beta + \sigma)$ captures the treatment response for facilities with lower historical compliance (higher discharge ratios). The interaction coefficient,

σ, captures the difference in the treatment effects between facilities with higher and lower historical compliance.

Hypothesis H2 implies that $\sigma < 0$. Recall that, for *Hypothesis H2*, the presence of a "boomerang effect" implies that facilities with stronger compliance histories reduce their extent of compliance (increase their discharge ratio): $\beta > 0$. For *Hypothesis H3b*, which assumes facility managers are imperfectly informed or inattentive to cost-minimization, the treatment reduces all facilities' extent of compliance (i.e., raises discharge ratios), with the larger effect on facilities with stronger compliance histories; i.e., $\beta > 0$, $\sigma < 0$, and $(\beta + \sigma) > 0$.

To test *Hypothesis H3c*, we interact the treatment variable with binary indicator of whether the facility is a "major" or "minor" facility, M_{it}. The indicator variable equals 1 for major facilities and equals 0 for minor facilities. We estimate an extended version of equation (1):

$$\ln(C_{it}) = \alpha + \beta' Z_{it} + \theta' K_{it} + \gamma' M_i + \sigma'(M_i \times Z_{it}) + \epsilon_i + \eta_{it}\,, \qquad (3)$$

where the coefficient β now captures the treatment response for minor facilities. The sum $(\beta + \sigma)$ captures the treatment response for major facilities. The interaction coefficient, σ, captures the difference in the treatment effects between minor and major facilities. *Hypothesis H3b* implies that $\beta \geq 0$ and $\sigma > 0$.

In an extended specification, we add two more control variables (accidently omitted from our pre-analysis plan). Of the 328 facilities in our sample, 11 faced both monthly limits and quarterly limits at some point over the sample period (but not necessarily in the same quarter). In those 11 cases, we calculated the average of the aggregated monthly discharge ratio and quarterly discharge ratio when both are non-missing in the same quarter. We create a binary indicator variable for these facilities (=1). This classification is straightforward except when a contact person was associated with multiple facilities that had differing reporting frequencies. To address

this challenge, as described in our pre-analysis plan, we calculated the contact-level average and rounded up any positive average to one, which means that, if the contact person oversaw at least one quarterly reporting facility, we assigned him or her to the "quarterly" block. Since this assignment may not accurately capture the setting of the relevant facilities, we created an indicator variable for facilities where the facility-level reporting frequency differs from the contact-level reporting frequency (=1) and include this indicator as a control factor.

To calculate a minimum detectable effect with our sample size and our estimator (1), we conducted a power analysis simulation using the 2016 compliance data to approximate the data generating process. Setting statistical power to 80 % and the Type 1 error rate to 5 %, the design is powered to detect a 14 % (0.19 SD) difference in the mean BOD discharge ratios of the treated and control groups. Although the power of our design is better than the power of many environmental economics studies (Ferraro and Shukla, 2020), the power is inadequate if the true average effect of peer comparisons in our population were much smaller. The subgroup analyses – equations (2) and (3) – are exploratory in nature; for these analyses, we do not attempt to maintain the family-wise Type 1 error rate or control the false discovery rate.

5.2. Excludability (Exclusion Restriction)

Our design assumes that the randomization procedure has no effect on the discharge ratio except through its effect on exposure to the peer comparison letter. We did not inform facilities or other actors about the field experiment. Thus, it is unlikely that facilities were aware of the random assignment to treatment and control groups or the purpose of the randomization. Had the message come from a regulator, one might worry that receiving such a letter from the regulator could affect discharges separate from the peer comparisons, via perceptions of monitoring

scrutiny or enforcement pressure (whether that path is a violation of the exclusion restriction or simply a different mechanism is debatable). In our experiment, however, the letters came from university professors who have no regulatory authority.

5.3. Non-compliance

Control facilities had no access to our peer comparisons and would need to process NPDES data to generate their own comparisons. However, despite our best efforts to mail the treatment letter to the person who makes decisions about wastewater management, the facility manager may not have received or read our certified letter. Thus, in the presence of potential one-way non-compliance, our estimand is best interpreted as either the intent-to-treat effect of reading a peer comparison letter or the average treatment effect of having been sent such a letter.

5.4. Interference among Facilities

We assume no interference among facilities, which implies that a facility's potential pollution outcomes under treatment and control conditions do not depend on the treatment status of other facilities, i.e., Stable Unit Treatment Value Assumption, SUTVA, is satisfied. Recipients of the treatment letter could share the information in the letter with non-recipients in the control group. Should such interference exist, our estimator of the average treatment effect would be biased towards zero. Recipients of the treatment letter could also share the information with other letter recipients, which could bias our estimator either up or down.

5.5. Attrition

No municipal wastewater treatment facilities closed or lost their permits in the post-treatment period. Nevertheless, 21 quarterly-reporting facilities never report a discharge during the post-treatment period. Although this "missingness" could be viewed as a form of attrition, we do not believe it affects the internal or external validity of the design. With regard to internal validity, we believe that assuming discharge reports are missing independent of potential outcomes is reasonable (i.e., the "missingness" creates no bias). First, the count of missing reporters in the post-treatment period is similar between the treatment and control groups (11 in treated group, 10 in control group). Second, the treatment letters most likely did not induce any change in the composition of reporters in the post-treatment period. With regard to external validity, we believe that the presence of missing facilities does not require us to re-define the estimand. Consistent with this belief, the average pre-treatment discharge ratio of the 21 missing quarterly reporters is similar to the average pre-treatment discharge ratio of the 266 observable quarterly reporters (0.45 vs 0.47, a difference of 0.02, 95 % CI [-0.13, 0.08]).

6. Results

6.1. Main Estimates and Hypothesis Tests

We report the estimated coefficient β from the models, as well as the implied elasticity measure, which comes from exponentiating β and subtracting one. The elasticity is the percentage change in the discharge ratio when a facility is exposed to the treatment letter rather than not exposed. Using only Q3:2018 as the post-treatment period and the estimator that includes covariates (i.e., equation (1)), the estimated average treatment effect of peer comparisons on the discharge ratio is essentially zero (Table 2, cols. 1 and 2).

However, using the full post-treatment period and the same estimator (Table 3, col. 1), the estimated average treatment effect is a 7.9 % reduction in the discharge ratio (95% CI [-15.14 %, -0.67 %]). This estimate is similar to the estimate that one would obtain from simply taking the difference in the average post-treatment discharge ratios (untransformed) for the treatment group (0.4088) and the control group (0.4388): a 6.9 % reduction.

[**Table 2. Treatment Effect: Post-treatment Period Limited to Third Quarter 2018** here]

[**Table 3. Treatment Effect: Post-treatment Period using all Quarters and Placebo test** here]

To assess the robustness of the estimated effect in Table 3 col. 1, we conduct two additional analyses. First, to ensure that outliers are not driving this estimate, we trim (drop) the top and bottom 1 % (Table 3, col. 2) and the top and bottom 5 % (Table 3, col. 3) of discharge ratio values. The motivation for trimming is that, even after the log transformation, the data remain modestly right-skewed. After trimming, the estimated effect barely changes, while the precision of the estimate improves. The latter improvement may arise for two reasons. First, facilities with discharges near the lowest detectable values may find it difficult to reduce any further. Second, some facilities report being substantially out of compliance. For example, the top 1 % of observations have an average discharge ratio of 2.02, indicating that these facilities are 100 % above their discharge limits, on average. These grossly non-compliant facilities may struggle to address the issues underlying their non-compliance in a short time horizon.

Second, we run a placebo (falsification) test using the pre-treatment period Q1:2015 to Q2:2017. We falsely assume that the treated group was treated in Q3:2016, rather than in Q3:2017; i.e., we specify Q1:2015 to Q2:2106 as the placebo pre-treatment period and Q3:2016

to Q2:2017 as the placebo post-treatment period. We then implement the same estimator again. If, by chance, pre-existing differences exist between treated and control groups, particularly in association with the third quarter, we might falsely detect a treatment effect when none exists. The estimated placebo treatment effect is about half the size of the estimate for the actual treatment effect and does not statistically differ from zero (Table 3, col. 4). When we conduct the same trimming exercise described above, we find that the estimated effect drops towards zero (Table 3, col. 5 and 6).

To better understand the stark difference in inferences between the estimated effects in Table 2 and Table 3, we estimate quarter-by-year treatment effects, by interacting the treatment dummy variable with the quarter-by-year dummy variables. Table 4 presents the estimated treatment effects for each post-treatment quarter. The estimated treatment effect is negative and of a policy-relevant magnitude in three of the six post-treatment quarters, but not in Q3:2018. Had we instead pre-registered Q4:2018 for our "later-quarter only" specification, we would have estimated a 15 % reduction in the discharge ratio from the peer comparisons. Thus, the "later-quarter only" specification is sensitive to the choice of quarter.

[Table 4. Treatment Effect by Quarter here]

We next assess heterogeneity of the treatment effect. Table 5 displays the estimates of conditional average treatment effects for the two sub-group contrasts identified in our pre-analysis plan: (1) major versus minor facilities, and (2) facilities with below-median 2016 discharge ratios versus facilities with above-median 2016 discharge ratios. Not surprisingly, given our modest sample size, the confidence intervals on these conditional average treatment

effects are wide and the point estimates are sensitive to the trimming rules. The results provide no evidence of heterogeneous treatment effects conditional on these two facility attributes.

[Table 5. Treatment Effects for Sub-Groups based on Facility Status (Major vs Minor) and Compliance History (Above-median vs Below-Median Discharge Ratio) here**]**

6.2. Monthly and Quarterly Panel Datasets Estimated Separately

Our analysis also explores the monthly and quarterly facilities separately.[6] Table 6 displays the estimated effects of peer comparisons separately for facilities that report monthly and those that report quarterly. The estimated effect for the monthly reporters is large, irrespective of whether we trim the data or not: an estimated 11 % - 14 % reduction in the discharge ratio. Although the confidence intervals are wide, they exclude zero, and are narrower after trimming. In contrast, the estimated effect for the quarterly reporters is small, irrespective of whether we trim the data or not: an estimated 1 % - 2 % reduction in the discharge ratio.

These differences may simply be due to sampling variability, but they may also stem from facility-specific conditions that prompt agencies to require some facilities to report monthly. The NPDES Permit Writer's Manual instructs permitting agencies to require monthly reporting (as opposed to quarterly reporting) when they believe that a facility's discharges require more frequent attention from both the facility manager and the regulator.[7] More frequent reporting is encouraged when discharges are highly variable, discharges are nearing the facility's treatment design capacity, discharges are heavily polluted or highly toxic, treatment systems are not

6 Although we blocked randomized on reporting frequency, we did not include in the pre-analysis plan an analysis of monthly and quarterly facilities separately (this analysis was not specified in the pre-analysis plan). No other deviations from the pre-analysis plan occurred.
7 https://www3.epa.gov/npdes/pubs/pwm_chapt_08.pdf

achieving sufficiently high pollutant removal on a consistent basis, facilities have poor compliance histories, or the receiving waters are sensitive or near public water supplies (see MM). Given the greater attention to discharges when required reporting is more frequent, managers and staff at these facilities are likely more attentive and engaged in their operations.

[**Table 6. Treatment Effects for Sub-Groups based on Reporting Frequency (Monthly vs. Quarterly** here]

7. Discussion

Returning to our three hypotheses, the evidence in Section 6 implies that one can reject the traditional (Becker-inspired) model, described in Section 3; specifically, when using the full panel data set, we reject the null hypothesis that $\beta = 0$ at p=0.05. We reject that model in favor of the alternative model in which facility managers hold non-financial (intrinsic) motivations. The point estimates in the regressions favor *Hypothesis H1*, which posits that peer comparisons operate by creating or invoking a norm that "more compliance is always better than less compliance." This conclusion stems from three pieces of information. First, when using the full panel data set, the point estimate implies that $\beta < 0$: on average, facilities reduce their discharge ratios. Second, we find no evidence that below-median dischargers responded any differently to the treatment than above-median dischargers, which implies no evidence to support *Hypothesis H2*, which posits that facility managers act as conformists or conditional cooperators. Third, we find no evidence that the average facility or minor facilities increased their discharge ratios in response to the treatment, which implies no evidence to support *Hypothesis H3*, which posits that facility managers are imperfectly informed or acting as boundedly rational agents when choosing their privately optimal levels of compliance.

The empirical results also provide suggestive evidence that the treatment effect detected in

the full data set stems from the responses of facilities that report monthly. We hypothesize that these facilities may be more responsive because, in contrast to quarterly reporters, their permit requirements force them to pay closer attention to their operations. Thus, these facilities may be more interested in and more capable of adjusting their operations in order to reduce discharges in a relatively short period of time (six quarters).

We remind the reader that, since facilities self-report DMRs, we cannot discriminate between real changes in discharges and strategic changes in wastewater sampling or misreporting that give the impression of changes in discharges.[8]

8. Conclusions

Our study explores the potential of peer (social) comparisons to encourage wastewater treatment facilities to voluntarily reduce their pollution.[9] Based on our empirical results, we conclude that peer comparisons can reduce pollution in a relatively short period of time through the non-financial (intrinsic) motivations of facility managers. We acknowledge that, given our modest sample size and the statistical significance of the estimated treatment coefficient, our treatment effect size of a 9 % reduction in pollution may be exaggerated (Ferraro and Shukla, 2020). Nevertheless, even if the true magnitude were smaller, the low cost of the intervention makes it a cost-effective option, in comparison to other tools for prompting greater compliance with wastewater limits, e.g., inspections.

Our study has a variety of shortcomings that future studies should address. First, even

8 In Figure 1, we assume misreporting discharges means under-reporting discharges. Strategic over-reporting is also possible, but our model does not offer a reason why a facility would strategically over-report its discharge ratio; in other words, the model does not explain why a facility would misreport that it is less compliant than it is.
9 Since facilities face legal limits, a reduction to an excessively high discharge ratio is not technically "voluntary"; moreover, in the context of stochastic discharges, a reduction to a discharge ratio lying below but sufficiently close to one (exact compliance) is not technically "voluntary" either. Still we use the term "voluntary" given its use by other relevant studies.

though we had over 4,000 time-facility observations, pollution is highly variable. Thus, our design was underpowered to detect small, but policy-relevant, effect sizes. Second, our study explores only a single type of organization – a municipal facility. Third, it considers only one type of environmental medium – wastewater. Fourth, our study examines only the single U.S. state – Kansas.

Fifth, our study explores only a single type of comparison framing, namely, a negative framing using a state-wide peer group and a single dose of that framing. Other framings or more frequent messaging might differ in effectiveness, as well as in how recipients perceive its usefulness and acceptability. Sixth, our study only examines the persistence of the effect for six quarters. Later measurements may imply that the effect wanes (e.g., Bernedo, Ferraro and Price, 2014). As described in our pre-analysis plan, we plan to collect data into the future to examine persistence. The final shortcoming concerns the limits on our ability to shed light on mechanisms or decision-making within the facility. Our plan for future data collection made us hesitant to interview any more than a dozen facility managers after the endline. The inability to interview managers forced us to examine the mechanisms through which peer comparisons operate indirectly, rather than by directly measuring changes in facility processes and investments. Future studies should consider creative ways to isolate the mechanisms through which peer comparisons may operate. This said, the inability to directly measure mechanisms is a problem in behavioral economics more generally. Thus, our study is not exceptional in this regard.

Given these limitations, our study design should be replicated, preferably with a larger sample size and ideally from a broader geographic area with a broader set of polluting facilities (e.g., industrial polluters in private sector). Future research should explore additional environmental media and explore alternative ways to construct and deliver the comparisons (e.g.,

regulator vs third party as messenger).

Despite the limitations of our study design, our results are policy-relevant if the magnitude of our estimated effect reflects, even approximately, the magnitude of the true effect size. Federal and state environmental agencies could use peer comparisons as a cost-effective supplement to existing voluntary and regulatory actions. We make no claims about the social welfare implications of the use of peer comparisons, i.e., whether the benefits from reduced pollution outweigh the costs of increased pollution abatement. However, if regulators believe that current pollution levels are higher than the socially optimal levels, providing peer comparisons is cheaper than creating new voluntary programs or inspecting and implementing enforcement actions more frequently. Although we cannot speculate on whether peer comparisons could be a substitute for regulations (rather than just a complement), scholars running experiments in collaboration with regulators could explore that potential substitutability. In sum, our results imply further work on peer comparisons in the context of pollution control is warranted.

Acknowledgements: We thank John Veresh for his indispensable insight on the EPA data; Marisa Henry for assistance with the registered pre-analysis plan and for conducting the semi-structured interviews; Bill Ball, Ed Bouwer, and Stacey Lamer for their insights into wastewater treatment facility managers and operations; and Ben Balmford, Paul Feldman, Ben Meiselman, and two anonymous referees for comments that improved the manuscript.

References

Alevy, J. E., Landry, C. E. and List, J. A. (2015) 'Field experiments on the anchoring of economic valuations', *Economic Inquiry*. doi: 10.1111/ecin.12201.

Allcott, H. (2011) 'Social norms and energy conservation', *Journal of Public Economics*. doi: 10.1016/j.jpubeco.2011.03.003.

Allcott, H. and Rogers, T. (2014) 'The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation', *American Economic Review*. doi: 10.1257/aer.104.10.3003.

Andor, M. *et al.* (2018) 'Social Norms and Energy Conservation Beyond the US', *SSRN Electronic Journal*. doi: 10.2139/ssrn.3234299.

Ayres, I., Raseman, S. and Shih, A. (2013) 'Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage', *Journal of Law, Economics, and Organization*. doi: 10.1093/jleo/ews020.

Becker, R. A. (2004) 'Pollution abatement expenditure by U.S. manufacturing plants: Do community characteristics matter?', *Contributions to Economic Analysis and Policy*. doi: 10.2202/1538-0645.1231.

Benartzi, S. *et al.* (2017) 'Should Governments Invest More in Nudging?', *Psychological Science*. doi: 10.1177/0956797617702501.

Bennear, L. S. and Olmstead, S. M. (2008) 'The impacts of the "right to know": Information disclosure and the violation of drinking water standards', *Journal of Environmental Economics and Management*. doi: 10.1016/j.jeem.2008.03.002.

Bernedo, M., Ferraro, P. J. and Price, M. (2014) 'The Persistent Impacts of Norm-Based Messaging and Their Implications for Water Conservation', *Journal of Consumer Policy*.

doi: 10.1007/s10603-014-9266-0.

Bruhn, M. and McKenzie, D. (2009) 'In pursuit of balance: Randomization in practice in development field experiments', *American Economic Journal: Applied Economics*. doi: 10.1257/app.1.4.200.

Byerly, H. *et al.* (2018) 'Nudging pro-environmental behavior: evidence and opportunities', *Frontiers in Ecology and the Environment*. doi: 10.1002/fee.1777.

Chabé-Ferret, S. *et al.* (2019) 'Can we nudge farmers into saving water? Evidence from a randomised experiment', *European Review of Agricultural Economics*. doi: 10.1093/erae/jbz022.

Charness, G. and Sutter, M. (2012) 'Groups Make Better Self-Interested Decisions', *Journal of Economic Perspectives*. doi: 10.1257/jep.26.3.157.

Chong, A. *et al.* (2015) '(Ineffective) messages to encourage recycling: Evidence from a randomized evaluation in peru', *World Bank Economic Review*. doi: 10.1093/wber/lht022.

Dasgupta, S., Hettige, H. and Wheeler, D. (2000) 'What improves environmental compliance? Evidence from Mexican industry', *Journal of Environmental Economics and Management*. doi: 10.1006/jeem.1999.1090.

Delmas, M., Montes-Sancho, M. and Shimshack, J. (2007) 'Mandatory information disclosure and environmental performance in the electricity industry', in *Academy of Management 2007 Annual Meeting: Doing Well by Doing Good, AOM 2007*. doi: 10.5465/ambpp.2007.26523229.

Duflo, E. *et al.* (2013) 'Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from india', *Quarterly Journal of Economics*. doi: 10.1093/qje/qjt024.

Earnhart, D. (2004a) ‘Panel data analysis of regulatory factors shaping environmental performance’, *Review of Economics and Statistics*. doi: 10.1162/003465304323023895.

Earnhart, D. (2004b) ‘Regulatory factors shaping environmental performance at publicly-owned treatment plants’, *Journal of Environmental Economics and Management*. doi: 10.1016/j.jeem.2003.10.004.

Earnhart, D. (2009) ‘The influence of facility characteristics and permit conditions on the effectiveness of environmental regulatory deterrence’, *Journal of Regulatory Economics*. doi: 10.1007/s11149-009-9095-2.

Earnhart, D. H. and Glicksman, R. L. (2015) ‘International Review of Law and Economics Coercive vs. cooperative enforcement: Effect of enforcement approach on environmental management’, *International Review of Law and Economics*. doi: 10.1016/j.irle.2015.02.003.

Earnhart, D. and Harrington, D. R. (2014) ‘Effect of audits on the extent of compliance with wastewater discharge limits’, *Journal of Environmental Economics and Management*. doi: 10.1016/j.jeem.2014.06.004.

Earnhart, D. and Segerson, K. (2012) ‘The influence of financial status on the effectiveness of environmental enforcement’, *Journal of Public Economics*. doi: 10.1016/j.jpubeco.2012.05.002.

Ferraro, P. J. and Miranda, J. J. (2013) ‘Heterogeneous treatment effects and mechanisms in information-based environmental policies: Evidence from a large-scale field experiment’, *Resource and Energy Economics*. doi: 10.1016/j.reseneeco.2013.04.001.

Ferraro, P. J., Miranda, J. J. and Price, M. K. (2011) ‘The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment’, in *American Economic Review*. doi: 10.1257/aer.101.3.318.

Ferraro, P. J. and Price, M. K. (2013) 'Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment', *Review of Economics and Statistics*. doi: 10.1162/REST_a_00344.

Ferraro, P. J. and Shukla, P. (2020) 'Is there a Replicability Crisis on the Horizon for Environmental and Resource Economics?', *Review of Environmental Economics and Policy*, 14(2).

Festinger, L. (1954) 'A Theory of Social Comparison Processes', *Human Relations*. doi: 10.1177/001872675400700202.

Gabel, H. L. and Sinclair-Desgagné, B. (1998) 'The Firm, Its Routines, and the Environment', in Folmer, H. and Tietenberg, T. (eds) *International Yearbook of Environmental and Resource Economics 1998/1999: A Survey of Current Issues*. Cheltenham, UK: Edward Elgar.

Gabel, H. L. and Sinclair-Desgagné, B. (2001) 'The Firm, Its Procedures, and Win-win Environmental Regulations', in Folmer, H. et al. (eds) *Frontiers of Environmental Economics*. Cheltenham, UK: Edward Elgar, pp. 148–175.

Gangadharan, L. (2006) 'Environmental compliance by firms in the manufacturing sector in Mexico', *Ecological Economics*. doi: 10.1016/j.ecolecon.2005.10.023.

Henriques, I. and Sadorsky, P. (1996) 'The determinants of an environmentally responsive firm: An empirical approach', *Journal of Environmental Economics and Management*. doi: 10.1006/jeem.1996.0026.

Henry, M. L., Ferraro, P. J. and Kontoleon, A. (2019) 'The behavioural effect of electronic home energy reports: Evidence from a randomised field trial in the United States', *Energy Policy*. doi: 10.1016/j.enpol.2019.06.039.

Khanna, M., Quimio, W. R. H. and Bojilova, D. (1998) 'Toxics release information: A policy

tool for environmental protection', *Journal of Environmental Economics and Management*. doi: 10.1006/jeem.1998.1048.

Klotz, L. *et al.* (2019) *Twenty Questions about Design Behavior for Sustainability, Report of the International Expert Panel on Behavioral Science for Design.* New York, NY.

Konar, S. and Cohen, M. A. (1997) 'Information as regulation: The effect of community right to know laws on toxic emissions', *Journal of Environmental Economics and Management*. doi: 10.1006/jeem.1996.0955.

Laplante, B. and Rilstone, P. (1996) 'Environmental inspections and emissions of the pulp and paper industry in Quebec', *Journal of Environmental Economics and Management*. doi: 10.1006/jeem.1996.0029.

List, J. A. (2003) 'Does market experience eliminate market anomalies?', *Quarterly Journal of Economics*. doi: 10.1162/00335530360535144.

Maciejovsky, B. *et al.* (2013) 'Teams Make You Smarter: How Exposure to Teams Improves Individual Decisions in Probability and Reasoning Tasks', *Management Science*. doi: 10.1287/mnsc.1120.1668.

Meub, L. and Proeger, T. (2018) 'Are groups "less behavioral"? The case of anchoring', *Theory and Decision*. doi: 10.1007/s11238-017-9608-x.

Palm-Forster, L. H. *et al.* (2019) 'Behavioral and Experimental Agri-Environmental Research: Methodological Challenges, Literature Gaps, and Recommendations', *Environmental and Resource Economics*. doi: 10.1007/s10640-019-00342-x.

Pargal, S. and Wheeler, D. (1996) 'Informal regulation of industrial pollution in developing countries: Evidence from Indonesia', *Journal of Political Economy*. doi: 10.1086/262061.

Porter, M. E. (1991) 'America's green strategy', *Scientific American*.

Reeves, R. (2012) ‘Guideline, education, and peer comparison to reduce prescriptions of benzodiazepines and low-dose quetiapine in prison’, *Journal of Correctional Health Care*. doi: 10.1177/1078345811421591.

Sacarny, A. *et al.* (2016) ‘Medicare letters to curb overprescribing of controlled substances had no detectable effect on providers’, *Health Affairs*. doi: 10.1377/hlthaff.2015.1025.

Schultz, P. W. *et al.* (2018) ‘The Constructive, Destructive, and Reconstructive Power of Social Norms: Reprise’, *Perspectives on Psychological Science*. doi: 10.1177/1745691617693325.

Shimshack, J. P. and Ward, M. B. (2005) ‘Regulator reputation, enforcement, and environmental compliance’, *Journal of Environmental Economics and Management*. doi: 10.1016/j.jeem.2005.02.002.

Short, J. L. and Toffel, M. W. (2010) ‘Making self-regulation more than merely symbolic: The critical role of the legal environment’, *Administrative Science Quarterly*. doi: 10.2189/asqu.2010.55.3.361.

Smith, V. L. and Walker, J. M. (1993) ‘Monetary Rewards and Decision Cost in Experimental Economics’, *Economic Inquiry*, (2), pp. 245–261.

Stafford, S. L. (2002) ‘The effect of punishment on firm compliance with hazardous waste regulations’, *Journal of Environmental Economics and Management*. doi: 10.1006/jeem.2001.1204.

Telle, K. (2009) ‘The threat of regulatory environmental inspection: Impact on plant performance’, *Journal of Regulatory Economics*. doi: 10.1007/s11149-008-9074-z.

Telle, K. (2013) ‘Monitoring and enforcement of environmental regulations. Lessons from a natural field experiment in Norway’, *Journal of Public Economics*. doi: 10.1016/j.jpubeco.2013.01.001.

Thaler, R. H. and Sunstein, C. R. (2008) *Nudge: Improving decisions about health, wealth, and happiness, Nudge: Improving Decisions about Health, Wealth, and Happiness*. doi: 10.1016/s1477-3880(15)30073-6.

Wallander, S., Ferraro, P. and Higgins, N. (2017) 'Addressing participant inattention in federal programs: A field experiment with the conservation reserve program', *American Journal of Agricultural Economics*. doi: 10.1093/ajae/aax023.

Wasserman, C. (1984) *Improving the Efficiency and Effectiveness of Compliance Monitoring and Enforcement of Environmental Policies, United States: A National Review*.

Winter, S. C. and May, P. J. (2001) 'Motivation for compliance with environmental regulations', *Journal of Policy Analysis and Management*. doi: 10.1002/pam.1023.

Wooldridge, J. M. (2002) 'Econometric Analysis of Cross Section and Panel Data', *Booksgooglecom*. doi: 10.1515/humr.2003.021.

FIGURES

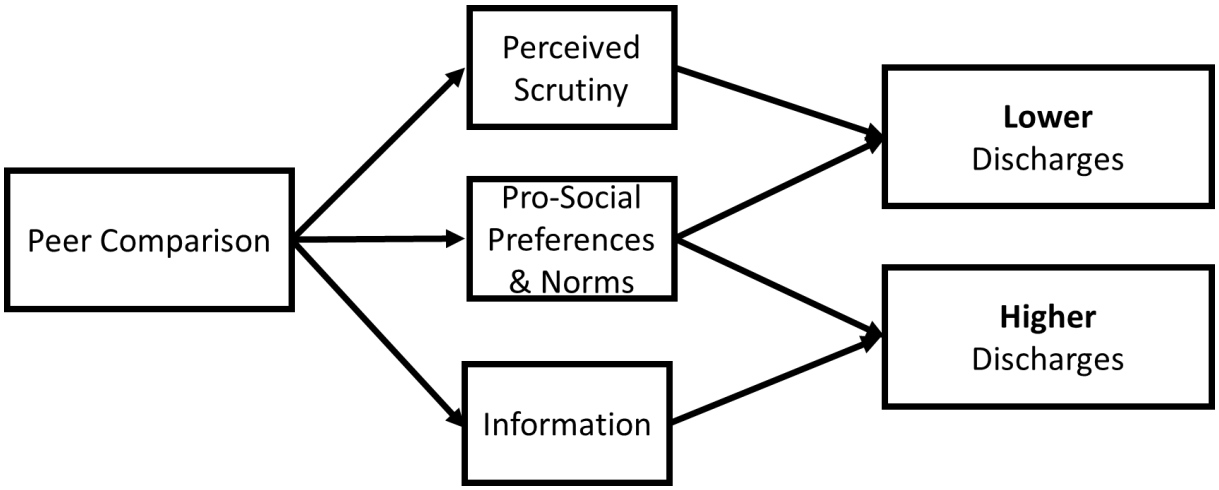

Figure 1. Potential Mechanisms through which Peer Comparisons could Affect Pollution Discharges. (a) The pro-social preferences and norms channel is typically assumed to induce a decrease in discharges among all managers (or increase the strategic under-reporting of discharges). However, peer comparisons could induce an increase in discharges among higher-compliance managers if those managers are conformists or conditional cooperators; (b) Given that abatement is costly, the information channel is assumed to signal to self-interested, over-compliers that such over-compliance may be sub-optimal; (c) If managers of regulated facilities perceive greater scrutiny of their actions by regulators or civic actors via delivery of peer comparisons, the managers may be induced to reduce their pollution (or increase the strategic under-reporting of discharges).

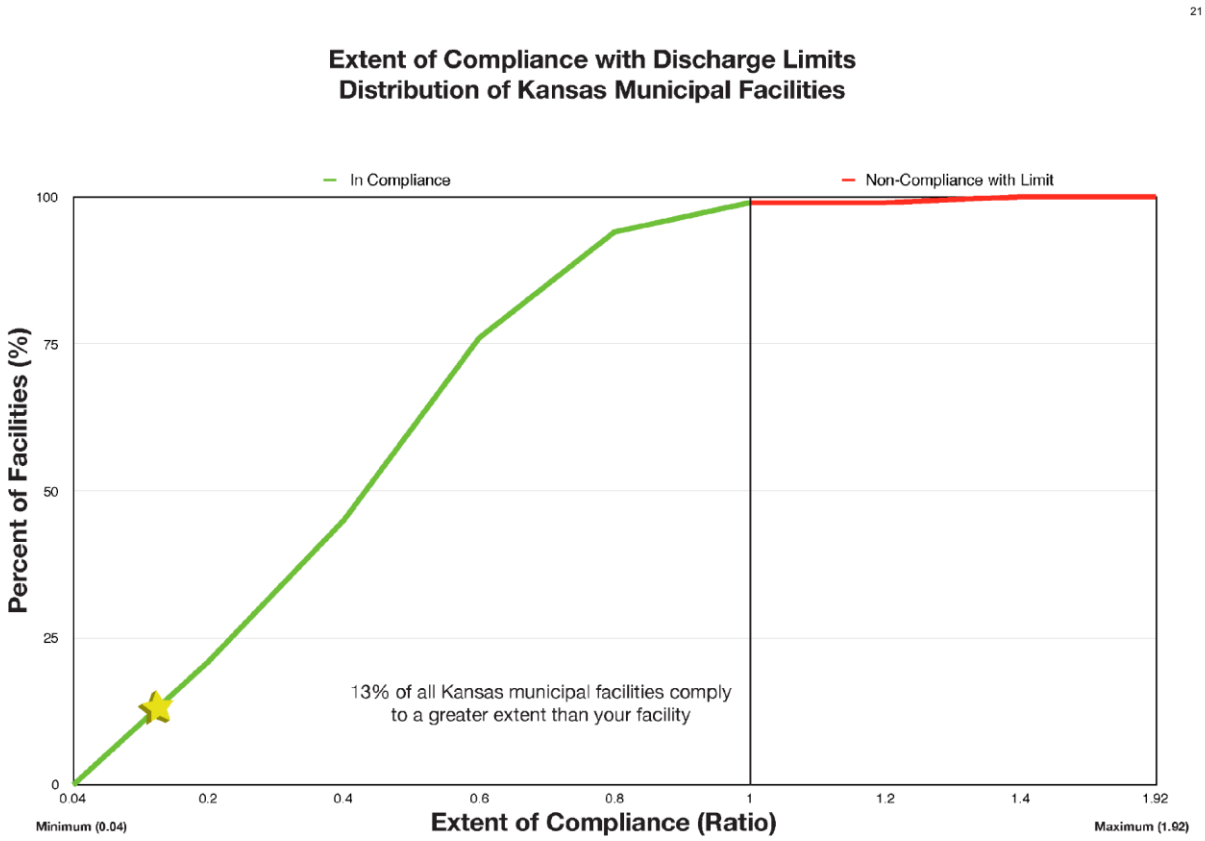

Figure 2. Peer Comparison Message (example). Recipients received a full-page graphic showing the distribution of discharge ratios among wastewater treatment facilities in Kansas and where the recipient's facility falls on the distribution. The graphic was accompanied by a two-sided letter, which explained the discharge ratio metric, the peer comparison, and the motivation for the mailing (see MM).

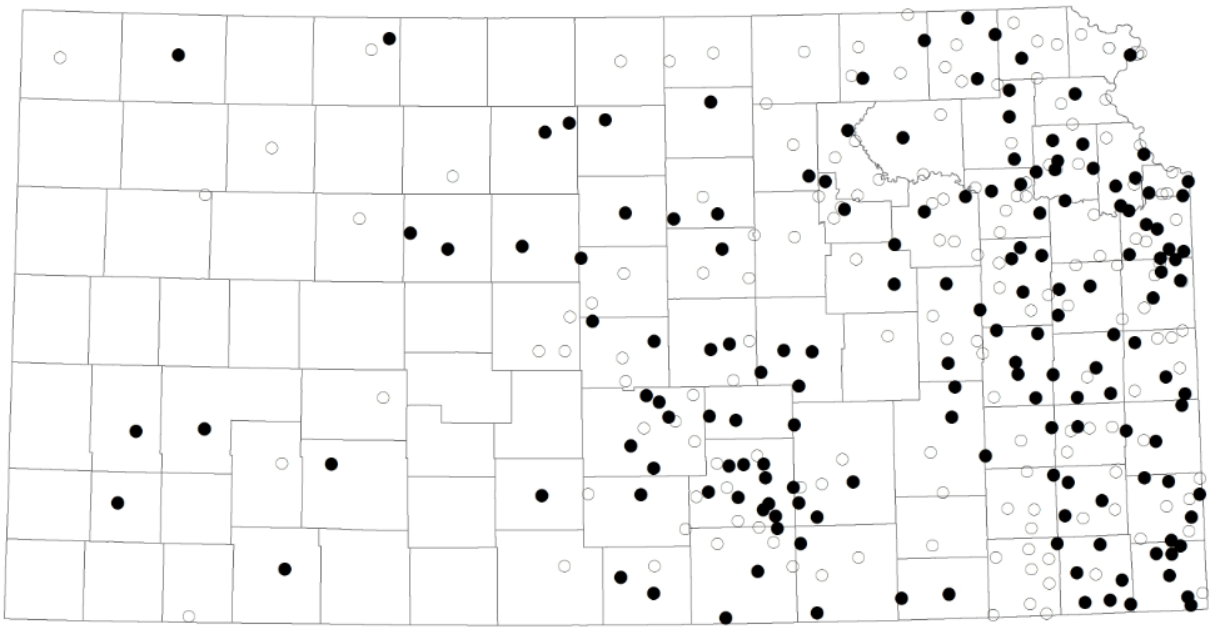

Figure 3. Kansas Municipal Wastewater Treatment Facilities in Experiment. Gray lines represent county boundaries. ● = treated facilities; ○ = control facilities

TABLES

Table 1
Descriptive Statistics

	Overall Sample (N=328)	**Treated Group** (N=164)	**Control Group** (N=164)
Mean Discharge Ratio (2016) [a,b]	0.438 (0.253)	0.436 (0.247)	0.439 (0.260)
Proportion in Top Quartile of Discharge Ratio (2016) [a]	0.247	0.244	0.250
Proportion in Third Quartile of Discharge Ratio (2016) [a]	0.250	0.250	0.250
Proportion in Second Quartile of Discharge Ratio (2016) [a]	0.250	0.244	0.256
Proportion Major Facilities	0.125	0.122	0.128
Proportion Quarterly Reporting	0.610	0.616	0.604
Proportion Minor Facilities Reporting Monthly	0.265	0.262	0.268

[a] Facility discharge ratios from 2016 were used to form the peer comparisons.
[b] Standard deviation shown in parentheses.

Table 2
Treatment Effect: Post-treatment Period Limited to Third Quarter 2018 [a]

	Specification 1	**Specification 2**
Estimated Coefficient	0.0034	-0.0005
Standard Error [b]	0.0677	0.0678
95 % Confidence Interval	[-0.1292, 0.1360]	[-0.1334, 0.1325]
Elasticity	0.34 %	-0.05 %
95 % Confidence Interval	[-12.97 %, 13.65 %]	[-13.34 %, 13.24 %]
Sample Attributes		
Number of Observations	2,961	2,961
Number of Facilities	328	328
Number of Contact Names	313	313
Control Variables		
Blocking Covariates	Y	Y
Time Covariates	Y	Y
Facility Covariates		Y

[a] The post-treatment period is restricted to only the 3rd quarter of 2018. The pre-treatment period runs from the 1st quarter of 2015 through the 2nd quarter of 2017.
[b] Standard errors are clustered at the level of the contact name.

Table 3
Treatment Effect:
Post-Treatment Period using all Quarters and Placebo Test [a]

	Post-Treatment Period (all quarters) [b]			**Placebo Test** [c]		
	Full	**Trimmed 1%** [d]	**Trimmed 5%** [d]	**Full**	**Trimmed 1%** [d]	**Trimmed 5%** [d]
Estimated Coefficient	-0.0823	-0.0870	-0.0812	-0.0441	-0.0168	-0.0015
Standard Error [e]	0.0401	0.0374	0.0340	0.0396	0.0357	0.0305
95 % Confidence Interval	[-0.1609, -0.0038]	[-0.1603, -0.0137]	[-0.1479, -0.0145]	[-0.1216, 0.0335]	[-0.0867, 0.0531]	[-0.0613, 0.0583]
Elasticity	-7.90 %	-8.33%	-7.80%	-4.31 %	-1.66 %	-0.15 %
95 % Confidence Interval	[-15.14 %, -0.67 %]	[-15.05 %, -1.61 %]	[-13.95 %, -1.65 %]	[-11.73 %, 3.11 %]	[-8.54 % , 5.21 %]	[-6.12 % , 5.83 %]
Sample Attributes						
Number of Observations	4,151	4,070	3,743	2,685	2,632	2,410
Number of Facilities	328	328	326	328	328	325
Number of Contact Names	313	313	311	313	313	310

[a] All regressions include controls for the blocking, time, and facility covariates.
[b] The full post-treatment period runs from the 3rd quarter of 2017 through the 4th quarter of 2018. The pre-treatment period runs from the 1st quarter of 2015 through the 2nd quarter of 2017.
[c] The full post-placebo period runs from the 3rd quarter of 2016 through the 2nd quarter of 2017. The pre-treatment period in this regression runs from the 1st quarter of 2015 through the 2nd quarter of 2016.
[d] Analysis trims the full sample by eliminating the top 1 % or 5 % and bottom 1 % or 5 % of the discharge ratio distribution.
[e] Standard errors are clustered at the level of the contact name.

Table 4
Treatment Effect by Quarter [a,b]

	Q3:2017	**Q4:2017**	**Q1:2018**	**Q2:2018**	**Q3:2018**	**Q4:2018**
Difference in Coefficients: Treatment vs. Control Groups	-0.1423	-0.1501	-0.0416	0.0057	0.0093	-0.1633
Standard Error [c]	0.0755	0.0759	0.0650	0.0764	0.0748	0.0704
95 % Confidence Interval	[-0.2903, 0.0057]	[-0.2989, -0.0014]	[-0.1689, 0.0858]	[-0.1441, 0.1555]	[-0.1374, 0.1559]	[-0.3013, -0.0253]
Elasticity	-13.26 %	-13.94 %	-4.07%	0.57 %	0.93 %	-15.07 %
95 % Confidence Interval	[-26.10 %, -0.43 %]	[-26.74 %, -1.14 %]	[-16.29 %, 8.15 %]	[-14.49 %, 15.63 %]	[-13.87%, 15.73 %]	[-26.79 %, -3.35 %]

[a] The regression includes controls for the blocking, time, and facility covariates.
[b] Sample comprises 4,151 observations on 328 facilities and 313 contact names.
[c] Standard errors are clustered at the level of the contact name.

Table 5
Treatment Effects for Sub-Groups based on Facility Status (Major vs Minor) and Compliance History (Above-Median vs Below-Median Discharge Ratio) [a,b,c]

	Facility Status				Compliance History [d]			
	Major Facilities	**Minor Facilities**	**Difference**	**Difference (trimmed sample)** [e]	**Above-median Discharge Ratio**	**Below-median Discharge Ratio**	**Difference**	**Difference (trimmed sample)** [e]
Estimated Coefficient	-0.1507	-0.0682	-0.0825	-0.0063	-0.0683	-0.0736	0.0053	-0.0568
Standard Error [f]	0.0744	0.0422	0.0766	0.0615	0.0459	0.0530	0.0583	0.0505
95 % Confidence Interval	[-0.2966, -0.0049]	[-0.1509, 0.0145]	[-0.2326, 0.0676]	[-0.1270, 0.1143]	[-0.1583, 0.0216]	[-0.1776, 0.0303]	[-0.1090, 0.1196]	[-0.1558, 0.0423]
Elasticity	-13.99 %	-6.59%	-7.40 %	-0.58 %	-6.60%	-7.10 %	-0.49 %	-5.27 %
95 % Confidence Interval	[-26.54 %, -1.45 %]	[-14.32 %, 1.13 %]	[-20.52 %, 5.72 %]	[-11.66 %, 10.49 %]	[-15.00 %, 1.80 %]	[-16.75 %, 2.56 %]	[-10.15 %, 11.14 %]	[-14.54 %, 4.00 %]

[a] All regressions include controls for the blocking, time, and facility covariates.
[b] The post-treatment period runs from 3rd quarter of 2017 through 4th quarter of 2018. The pre-treatment period runs from 1st quarter of 2015 through 2nd quarter of 2017.
[c] Sample comprises 4,151 observations on 328 facilities and 313 contact names except as noted.
[d] Pre-treatment compliance history is based on calendar year 2016 discharge ratios.
[e] Analysis trims the sample by eliminating the top and bottom 5% of the discharge ratio distribution; the resulting sample comprises 3,743 observations on 326 facilities and 311 contact names.
[f] Standard errors are clustered at the level of the contact name.

Table 6
Treatment Effects for Sub-Groups based on Reporting Frequency (Monthly vs. Quarterly)[a]

Treatment Effect	Monthly Reporters			Quarterly Reporters		
	Full [b]	**Trimmed 1%** [c]	**Trimmed 5%** [c]	**Full** [b]	**Trimmed 1%** [c]	**Trimmed 5%** [c]
Estimated Coefficient	-0.1202	-0.1218	-0.1468	-0.0169	-0.0117	0.0093
Standard Error [d]	0.0600	0.0552	0.0548	0.0478	0.0425	0.0322
95 % Confidence Interval	[-0.2378, -0.0026]	[-0.2300, -0.0136]	[-0.2543, -0.0393]	[-0.1106, 0.0768]	[-0.0951, 0.0716]	[-0.0538, 0.0724]
Elasticity	-11.33 %	-11.47 %	-13.65%	-1.67 %	-1.17 %	0.94 %
95 % Confidence Interval	[-21.75 %, -0.90 %]	[-21.05 %, -1.89 %]	[-22.93 %, -4.37 %]	[-10.88 %, 7.54 %]	[-9.40, 7.07 %]	[-5.44 %, 7.31 %]
Number of Observations	4,432	4,349	3,980	2,421	2,371	2,084
Number of Facilities	133	133	132	206	206	205
Number of Contact Names	129	129	128	200	200	199

[a] All regressions include controls for the blocking, time, and facility covariates.
[b] The full post-treatment period runs from the 3rd quarter of 2017 through the 4th quarter of 2018. The pre-treatment period runs from the 1st quarter of 2015 through the 2nd quarter of 2017.
[c] Analysis trims the full sample by eliminating the top 1 % or 5 % and bottom 1 % or 5 % of the discharge ratio distribution.
[d] Standard errors are clustered at the level of the contact name.