Title

Socioeconomic Disparities in Privatized Pollution Remediation: Evidence from Toxic Chemical Spills

Journal

American Economic Journal Applied Economics, 16(3)

ISSN

1945-7782

Authors

Marion, Justin

West, Jeremy

Publication Date

2024-07-01

DOI

10.1257/app.20220295

Socioeconomic Disparities in Privatized Pollution Remediation: Evidence from Toxic Chemical Spills

Justin Marion * Jeremy West

University of California, Santa Cruz

Manuscript accepted September 2023 at
American Economic Journal: Applied Economics

Abstract

Governments often privatize the administration of regulations to third-party specialists paid for by the regulated parties. We study how the resulting conflict of interest can have unintended consequences for the distributional impacts of regulation. In Massachusetts, the party responsible for hazardous waste contamination must hire a licensed contractor to quantify the environmental severity. We find that contractors' evaluations favor their clients, exhibiting substantial score bunching just below thresholds that determine government oversight of the remediation. Client favoritism is more pronounced in socioeconomically disadvantaged neighborhoods and is associated with inferior remediation quality, highlighting a novel channel for inequities in pollution exposure.

JEL: L51, Q53, D63, J15, K32

Keywords: privatized assessments, socioeconomic disparities, environmental remediation

*Marion (corresponding author): marion@ucsc.edu. West: westj@ucsc.edu. This work benefited from valuable discussion with several retired Licensed Site Professionals and from helpful comments by the editor and several anonymous reviewers, and suggestions from Spencer Banzhaf, James Bushnell, Peter Christensen, Kenneth Gillingham, Joshua Graff Zivin, Wayne Gray, Alex Hollingsworth, Koichiro Ito, David Keiser, Ashley Langer, Matthew Neidell, Paulina Oliva, Ivan Rudik, Nicholas Ryan, Edson Severnini, Joseph Shapiro, Richard Sweeney, Christopher Timmins, Arthur van Benthem, John Voorheis, Jessica Wolpaw Reyes, and seminar participants at the 2020 NBER Energy and Environmental Economics Conference, the 2021 European Urban Economics Conference, the 2021 International Industrial Organization Conference, the 2022 Southern California Conference on Applied Micro, the UC-Environmental Economics Workshop, Santa Clara University, UAB, UC-Santa Cruz, and Yale University. Any errors, opinions and conclusions in this study are our own and do not necessarily represent the position of the Massachusetts Department of Environmental Protection or the Massachusetts Licensed Site Professional Association. This research did not receive any specific grant from funding agencies.

1 Introduction

How the enforcement of environmental regulation can contribute to differential pollution exposure by socioeconomic status is an unsettled question in the environmental literature. In many settings, an important role in regulatory enforcement and information provision is played by private third-party agents hired by the very parties subject to the regulations. Examples of such arrangements include credit ratings, emissions monitoring, and food safety inspections, to name a few (White, 2010; Duflo et al., 2013a,b; Lytton and McAllister, 2014; Oliva, 2015). This delegation of administrative duties to the private sector is attractive because it leverages the expertise of firms and their cost-containment motive, while also shifting some of the fiscal burden off of government budgets. However, a principal-agent problem arises. The evaluator's assessment may be driven by the client's interests and not necessarily the interests of society, leading to an inefficient provision of the regulated quantity (pollution, food safety, etc.). The biased assessments that may result from this conflict of interest have been established in several empirical studies, including pollution abatement (Duflo et al., 2013b). Less well studied are the potential distributional consequences of these conflicts of interest, which could arise either from the subjective biases of the agent or if the agents' incentives lead to inequitable outcomes of enforcement.

In this study, we provide evidence regarding the heterogeneous effects of privatizing the administration of regulation in the context of hazardous waste site remediation in Massachusetts. Although the state provides umbrella regulatory enforcement, the party responsible for the environmental contamination is legally required to hire a private firm, called a Licensed Site Professional (LSP), to assess the site's severity. The state then relies upon these evaluations in order to target government oversight of site remediation towards the most serious spills. A conflict of interest thus arises: the state requires accurate assessments to be able to efficiently monitor site cleanups, whereas responsible parties may prefer discounted assessments in order to reduce their private costs of remediation. Our empirical results examine whether the principal-agent problem is more pronounced—and the remediation outcomes worse—in socioeconomically disadvantaged neighborhoods.

We begin by presenting a model of the incentives for misreporting site severity that highlights the channels through which inequities in pollution exposure can arise from regulatory enforcement. The LSP could suffer less reputation or psychic cost by providing lax assessments in lower income or greater minority areas. Alternatively, the bias in assessments can also be driven by the willingness-to-pay for environmental amenities in a neighborhood, resulting in worse cleanup outcomes and greater pollution exposure in poorer neighborhoods.

In such neighborhoods, the property value gain to cleanup is smaller, and as a consequence the polluter may wish to put inefficiently low effort into site cleanup and will demand an inaccurate assessment of site severity to obtain less government oversight. This particular mechanism ties together two distinct mechanisms presented in the environmental justice literature: that environmental regulations may be differentially applied depending on race and income, and that market forces lead people of color and the poor to experience more pollution because of a lower environmental willingness-to-pay.

Guided by this model of the theoretical incentives for misreporting site severity assessments, we present three sets of empirical evidence from this setting. First, we demonstrate that LSPs provide favoritism to their clients, which is enabled in part by the discretion that they have in conducting their evaluations. Second, we find that this client favoritism is associated with adverse environmental consequences including lower quality site cleanups and reduced government oversight of comparatively more serious sites. Finally, we show that this client favoritism has adverse equity consequences. The principal-agent problem is most pronounced for sites located in neighborhoods with lower income, lower property values, lower education, and a greater share of population of color.

To arrive at these findings, we study discontinuities in the scoring criteria that the government required to categorize sites according to their severity. Using a government-specified scoresheet called the Numerical Ranking System (NRS), LSPs assigned each contamination site a quantitative score that denotes the site's potential impact on human and ecological populations. Based almost exclusively on this NRS score, each site was then classified into one of four distinct severity categories called tiers. More hazardous spills (with more serious tier classifications) receive greater scrutiny and oversight throughout the site cleanup by the government.

We exploit the discontinuous regulatory process in several ways. By examining the the distribution of NRS scores, we find substantial bunching just below the tier thresholds, indicating that LSPs manipulate site severity evaluations in favor of their clients. Although LSPs potentially face legal, reputation, or psychic costs from misreporting, they have an incentive to report downgraded scores if responsible parties share some of the associated (cleanup cost-savings) surplus with LSPs.[1] Altogether, this score bunching has a significant impact on the composition of tier classifications. The most prominent tier cutoff is between

[1]Responsible parties can share this surplus with LSPs either explicitly or implicitly via repeated business. We provide further discussion and examples pertaining to surplus sharing in Section 2.

Tier II (less severe) and Tier I (more severe), due to its location within the NRS distribution.[2] Using our estimated counterfactual density, we find that if the score distribution were instead smooth across this threshold—as would be expected absent manipulation—then the total number of sites receiving the more involved Tier I government oversight of remediation would increase by more than 20 percent. Restricted to the "manipulation region" in which we estimate that site scores can feasibly be manipulated, we find that 40 percent of scores are downgraded to fall below the Tier II threshold.

We further show that discretion afforded to LSPs in conducting their site evaluations appears to directly facilitate this NRS score manipulation. We examine LSPs' use of a NRS sub-score component that allows for score adjustments based entirely on the subjective judgment of the LSP. Empirically, these adjustments are rarely used, except for marginal sites that would otherwise be classified into a more severe tier. Holding other NRS components constant, setting these subjective adjustments to zero would alone increase the number of Tier I sites by 13 percent, almost two-thirds of the total incidence of score manipulation.[3]

Next, we explore how site characteristics vary discontinuously across tier thresholds to provide evidence of the environmental and equity consequences of assessment favoritism. As predicted by the model, we find that sites just barely receiving a Tier II classification are substantially less likely to be cleaned to a permanent solution that involves "no significant risk" and are more likely to achieve remediation resolution through land use restrictions as opposed to a complete removal of the hazardous material. This evidence supports that the conflict of interest leads to a lower quality cleanup of more severe spills, which amplifies the welfare consequences of favoritism in LSPs' site evaluations.

We then consider the heterogeneous consequences of client favoritism by examining how the likelihood of score manipulation varies across sites located in Census Tracts with differing socioeconomic characteristics. We do so using two methods. First, we estimate how predetermined characteristics of site Census Tracts vary across tier thresholds, finding that income, property values, education, and the white population share all increase discontinuously at the Tier I/II threshold. Second, we estimate the counterfactual density and manipulated density for subsamples of the data, finding that manipulation is much less likely to occur in Census Tracts with higher income, property values, education, and white population share. Both approaches demonstrate that sites located in neighborhoods that are socioeconomically

[2]As Section 3 describes in more detail, the Tier I category is subdivided in order of decreasing severity into Tiers IA, IB, and IC. Along with Tier II, these serve as the four distinct tier classifications in the NRS.

[3]This is an upper bound of the effect of eliminating this subjective criterion entirely, as LSPs might (further) adjust other NRS sub-score components in lieu of an explicitly discretionary factor.

disadvantaged are more likely to be manipulated into the less severe Tier II classification, and therefore the adverse impacts of NRS manipulation are concentrated among populations that are poorer, less educated, and of color.

Finally, we examine a 2014 reform that eliminated the role of site scoring in tier classification, thereby significantly limiting the role of subjective agent assessment. Not only did the share of sites classified in the most favorable tier drop substantially, but the socioeconomic gap between Tier I and Tier II sites subsequently narrowed, lending further support to our finding that manipulation of regulations pertaining to hazard site remediation has disparate effects depending on local socioeconomic characteristics.

Our study has several important policy implications and contributes to multiple strands of the literature. Most broadly, we add to a growing literature on the incentives and consequences of agents hired to serve in public policy administration capacities (e.g. Oliva, 2015; Fisman and Wang, 2017; Blonz, 2018; Jin and Lee, 2018; Dee et al., 2019; Gillingham et al., 2019; Reynaert and Sallee, 2019). The potential conflicts of interest that may arise from third-party assessments are shown by Duflo et al. (2013a,b), who study the monitoring of emissions for industrial plants in India.[4] As in our setting, privatized evaluators tend to report emissions levels that are just below regulatory thresholds, and a field experiment shows that truth-telling incentives reduce both scoring manipulation by evaluators and pollution emissions by firms. Although we document similar patterns of behavior in a related context, a key difference between their study and ours is that we focus on heterogeneity in agents' use of scoring manipulation. Affording third-party auditors some discretion in forming their assessments has the potential to be beneficial, while at the same time exacerbating the incentives for misbehavior. Our study provides novel evidence that the use of discretion can also be disparate, potentially amplifying socioeconomic inequities in pollution exposure.

We also contribute to the literature examining hazardous waste sites and their remediation. This literature has generally (though not always) estimated beneficial impacts of site cleanup on surrounding communities. Whereas Greenstone and Gallagher (2008) find little effect of Superfund status on nearby housing values, other studies find significant price appreciation upon waste site cleanup, with benefits concentrated in areas with low property values (e.g. Gamper-Rabindran and Timmins, 2013; Haninger et al., 2017). The literature also shows beneficial effects of waste site remediation for health outcomes and cognitive development (Currie et al., 2011; Persico et al., 2020). Prior work on hazard sites demonstrates

[4]See Shimshack (2014) for a broader discussion of the literature on environmental compliance monitoring. Studies in other contexts also show that increased oversight can improve the behavior of government agents (e.g. Borcan et al., 2017; West, 2018; Calvo et al., 2019).

that spill likelihood is affected by the financial status of the site owner (Cohn and Deryugina, 2018). Our findings highlight that, even following a spill, there is substantial heterogeneity in site remediation quality depending on site-specific factors.

In doing so, we also join a significant environmental justice literature that considers differences in exposure to pollution by race or income. In a detailed review of this literature, Banzhaf et al. (2019) suggest several mechanisms through which differential exposure can arise, including the initial siting of pollution, from household sorting by willingness-to-pay for environmental amenities, or by disparities in the enforcement of regulation. Hausman and Stolper (2021) provide a theoretical framework highlighting how disparities in the information about pollution can contribute to differential exposure. Our study provides new evidence related to the differential enforcement of regulation, for which the existing evidence is mixed. Whereas Lavelle and Coyle (1992) find that court-assessed penalties for violating environmental regulations are lower in high-minority areas, other studies find no or minimal disparities in pollution regulation enforcement by the local racial or income composition (Gupta et al., 1996; Viscusi and Hamilton, 1999; Gray and Shadbegian, 2004; Shadbegian and Gray, 2012). We present some of the only evidence of clearly intentional differences in the implementation of pollution regulations across areas of differing socioeconomic status, a previously under-explored dimension of environmental justice.

Finally, our results relate to the literature on willingness-to-pay for environmental amenities. Numerous studies show that heterogeneous household willingness-to-pay leads to socioeconomic differences in pollution exposure through residential sorting (e.g. Banzhaf and Walsh, 2008; Crowder and Downey, 2010; Gamper-Rabindran and Timmins, 2011; Depro et al., 2015). The evidence we present in this paper is consistent with these findings through an analogous mechanism. Polluters could seek lighter regulation of waste remediation and reduce remediation quality in socioeconomically disadvantaged neighborhoods if these communities have a comparatively lower willingness-to-pay (or ability-to-pay) for reductions in local pollution.

The remainder of this paper is organized as follows. In Section 2, we construct a model that illustrates the theoretical framework for scoring manipulation. In Section 3, we provide background institutional details on the Massachusetts hazardous waste site remediation program and describe the data we use in our empirical study. In Section 4, we present our empirical findings. Section 5 concludes.

2 Theoretical framework

In this section, we present a model of the principal-agent problem in the context of hazard site evaluation. The model characterizes the incentives for the evaluator to provide an inaccurate assessment to the government that is favorable to the evaluator's client, the responsible party, and it suggests several empirical implications that can be tested in the data.

When a hazardous spill occurs, the responsible party must hire a third-party specialist (hereafter agent), who assesses the environmental contamination at the site. This assessment uncovers the true site severity, z^*, observed only by the agent, who then reports a potentially discounted severity score of $z \leq z^*$ to the government.[5] A threshold score z_0 determines Tier assignment and thereby the stringency of regulation, with $z \leq z_0$ resulting in a less-serious Tier II categorization and $z > z_0$ the more serious Tier I. The agent incurs a cost of misreporting given by $\phi(z^* - z)$, continuous and differentiable in z, with $\phi'(z^* - z) > 0$ and $\phi(0) = 0$. This cost can represent loss of credibility, legal penalties, or a disutility of dishonesty.

The responsible party faces a cost of remediation, $c(e, x, z^*)$ and a benefit, $v(e, x, z^*)$, reflecting the capitalization of site cleanup into the property's value. Both the costs and benefits depend on the true z^*, other site characteristics x, and an endogenously chosen choice of cleanup effort $e \in \{0, 1\}$. Both $c(\cdot)$ and $v(\cdot)$ are increasing in e and differentiable in z.

Regulation places a constraint on e. If the site is classified as Tier I, then $e = 1$ must be provided. For some sites with a true severity of $z^* > z_0$, the responsible party's desired effort is $e = 0$ and the regulation binds. It is this subset of sites—those with severity greater than the threshold but with a low desired effort—where a conflict of interest arises for the agent. Misreporting the severity score to be $z = z_0$ increases the client's surplus by $w(x, z^*) = \Delta c - \Delta v$. If the agent receives a share of this surplus, $\lambda \in (0, 1]$, the score is misreported for sites where

$$\lambda w(x, z^*) > \phi(x, z^* - z_0). \tag{1}$$

A unique $\overline{z}(x)$ exists, representing the largest z^* that is manipulated downward to z_0, if

[5] We assume that the agent has no incentive to overstate the severity score, an assumption we discuss further in section 2.2.

$\phi(x, z^* - z_0)/\lambda w(x, z^*)$ is increasing in z^*.[6] This is a standard single crossing condition that holds if $\partial\phi/\partial z^* > \lambda \partial w/\partial z^*$ is satisfied, or in other words the cost of manipulation rises more steeply in site severity than the agent's benefit.[7] This assumption is likely to hold if credibly scoring sites below the threshold becomes increasingly infeasible as z^* rises. The misreporting region depends on site characteristics x that are related to Δv, Δc, or the misreporting cost ϕ. For instance, if the reputation loss is lower for agents manipulating scores in socioeconomically disadvantaged neighborhoods, then the manipulation region is wider for sites in those areas.

Two institutional details reveal how surplus may be transferred to the agent in exchange for a more favorable severity score. First, Seifter (2006) finds through interviews with market participants that the responsible party often hires the LSP to both score and remediate a site under a fixed price contract. This makes the LSP the residual claimant on savings achieved in cleanup cost through score manipulation, who would choose to manipulate if $\Delta c > \phi$, maximizing the joint surplus of the responsible party and LSP if the capitalization of cleanup effort into the property value is low ($\Delta v = 0$).[8] Second, reputation is a consideration for the LSP. We found evidence that LSPs advertise past occasions where Tier II status was achieved on sites they had scored, boasting that "successfully" classifying sites as Tier II reduces clients' remediation costs.

Using the above framework, we now illustrate how sorting into regulatory status based on x can arise. The model is deterministic and scores for all sites with true severity within the manipulation region $z^* \in (z_0, \overline{z}(x)]$ are manipulated to below the threshold. Suppose that the net benefit from manipulation is monotone in x, with $\lambda w(x, z^*) - \phi(x, z^* - z_0)$ decreasing in x. Then the width of the manipulation region $(z_0, \overline{z}(x)]$ shrinks as x increases. Sites observed just above the threshold must have a sufficiently high value of x such that even a small degree of manipulation is undesirable. Conversely, sites with lower values of x have wider manipulation regions, and relatively more scores are manipulated down to the tier

[6] If $w(x, z^*) \leq 0$, the solution to Equation (1) is trivial as the agent never misreports. We implicitly assume that there are at least some values of x such that $w(x, z^*) > 0$ and focus on these sites where there is an incentive to misreport.

[7] The cost of remediation and the property value both depend on the hidden z^*, and these enter non-separably in Equation (1). Therefore, restrictions on either $c(\cdot)$ or $v(\cdot)$ on their own are not sufficient to guarantee single crossing.

[8] The official NRS manual discusses that, "The Numerical Ranking System serves as the basis ... to classify a disposal site as either Tier I (requiring a Permit and some level of [Department of Environmental Protection] DEP oversight during remediation) or Tier II (requiring no permit or direct DEP oversight)." If a LSP is paid using a fixed price contract, then permit costs and (indirect) costs from government oversight would erode the LSP's profits.

threshold. Thus, the composition of sites with reported scores at the threshold is comprised to a greater extent by lower values of x, whereas the composition of sites with reported scores just above the threshold is comprised to a greater extent by higher values of x.

This relates closely to the literature that evaluates manipulation of the running variable for regression discontinuity designs. As described by DiNardo and Lee (2004) and, in the absence of manipulation, predetermined characteristics should be smooth across categorical thresholds. The socioeconomic characteristics of the neighborhood, x^{SE}, are determined prior to the decision to misreport the score. Without manipulation, the expected value of x^{SE} is approximately the same whether approaching the threshold from below or above. However, if scores are manipulated, then selection generates discontinuities in the expected value of x^{SE} at the threshold, with the sign of this discontinuity depending on the relationship between x^{SE} and the terms Δv and Δc. Likewise, Δc and Δv vary discontinuously at the threshold, though these objects are unobserved. Indeed, any variable that influences these objects, including unobserved site severity, can vary discontinuously at the threshold due to selective score manipulation.

Taken together, the model has three implications. First, a manipulated score distribution exhibits excess mass at the tier threshold.[9] Second, the remediation quality increases discontinuously at the threshold. Third, the mean values of site characteristics that are negatively related to the net benefits of manipulation (positively related to the cost of manipulation or the property value of cleanup, or negatively related to remediation costs) increase discontinuously at the threshold.

2.1 Welfare discussion

We next consider the welfare impacts of score misreporting and the tier classification system. The model above predicts score bunching at regulatory thresholds, but the bunching is not itself a source of inefficiency. Rather, bunching allows for the identification of score misreporting by agents. Inefficiencies arise from the coarseness of the Tier regime and from score manipulation. Tiering targets government resources toward the most severe sites, but the step function assigning tiers to site severity is only an approximation of the optimal regulation. The discreteness of tier assignment is inefficient relative to a smooth corrective taxation function. Furthermore, manipulation of site severity scores can lead to inefficiencies

[9] In our model, the agent has precise control over the reported z, but as we discuss below in Section 3.1, in practice this does not hold and the excess mass in the empirical distribution of z may not be concentrated exactly at the threshold.

under either tier classification or Pigouvian taxation. In principle, a corrective tax can be efficient if it is chosen in anticipation of score misreporting. However, it is not possible to design a tax that corrects for heterogeneity in misreporting due to unobserved costs and benefits, a consideration formalized by Banerjee et al. (2022) in the context of means-tested benefits. Therefore, the inefficiency of manipulation would persist even if a tier system is replaced by a more flexible Pigouvian tax.

Suppose that the total external cost of the spill is given by ηz^*, where η is the marginal external damage per severity unit. If $e = 1$, then severity is reduced to $z^* = 0$, whereas if $e = 0$ then severity remains at z^*. Social surplus is therefore given by $\Omega(e) \equiv v(e) - c(e) - (1 - e)\eta z^*$. It is socially efficient to exert cleanup effort if $\Omega(1) \geq \Omega(0)$:

$$\Delta v - \Delta c + \eta z^* \geq 0. \tag{2}$$

Absent manipulation of the evaluator's assessment of severity, a Pigouvian tax of $t = \eta$ per severity unit would internalize the pollution externality. The Tier classification system is inefficient by comparison, replacing the smooth Pigouvian tax schedule with a discontinuous step function in severity. The inefficiency arises because $e = 1$ is required of all sites above z_0 even when $e = 0$ is efficient, and similarly $e = 0$ is allowed for all sites below z_0 even though $e = 1$ may be socially efficient for some of these sites.

Manipulation of the severity assessment introduces further inefficiencies under both the tier classification and corrective taxation policies. Under tier classification, some sites above the threshold could satisfy Equation (2), but not $\Delta v - \Delta c \geq \phi(z_0)$. The severity scores for these sites are manipulated downward to $z = z_0$ and effort is inefficiently set at $e = 0$.

With taxation, efficient remediation can be attained, but only if there is homogeneity across sites in the cost of manipulation. Manipulation drives an additional wedge between the private net benefit of exerting high- versus low-effort, which the government can anticipate and undo through the choice of the tax rate if sites are homogeneous. To illustrate this, let $\lambda = 1$ and $T(z)$ be a potentially nonlinear tax schedule.[10] If $e = 0$, then the firm's reported $\hat{z} = \text{argmax}_z\{-T(z) - \phi(z^*, z)\}$. Then, after substituting $\hat{z}$ into the payoffs under no effort, $e = 1$ is chosen if $\Delta v - \Delta c - T(\hat{z}) - \phi(z^*, \hat{z}) \geq 0$. If the tax schedule $T(z)$ is set such that $T(\hat{z}) + \phi(z^*, \hat{z}) = \eta z^*$, then the social optimality condition in Equation (2) is satisfied despite the potential score manipulation. In contrast, if the cost of manipulation is heterogeneous across sites, then optimal corrective taxation is not possible and will distort the choice of e. Furthermore, depending on the nature of the heterogeneity, adverse distributional

[10] A nonlinear tax schedule is required if the level of manipulation depends on the true severity.

consequences can arise, as with the tier classification regime.

2.2 Welfare enhancing manipulation

If the objective function of the evaluator happens to align with society, social welfare could be enhanced by the manipulation of the site severity score. Two types of errors could arise that would inefficiently lead to a Tier mis-classification, which a well-intentioned evaluator could undo through manipulation in favor of society. First, there could be idiosyncratic site-specific conditions that are not well captured by z. Put differently, the severity score might contain *measurement error*. Second, the scoring criteria, or equivalently the scoring threshold, could be on average too strict (or not strict enough), so that the score has *bias*. These two scenarios have different empirical implications. An agent acting in the public interest would undo mean-zero measurement error by boosting some scores above the threshold while manipulating the scores of other sites below the threshold. With a scoring system that is biased upward, an agent acting in the public interest would downgrade site scores more often—or there would be more Tier I sites than socially optimal, or both.

Our empirical results provide evidence against these two sources of inefficiency. As a preview, we find that manipulation is almost always downward, in favor of the site receiving the less-severe classification. This is inconsistent with mean-zero error in the scoring criteria. Furthermore, the state revised its scoring criteria in 2014, and subsequently the number of sites receiving the more serious Tier I classification increased substantially. The outcome of this reform revealed a preference for more sites being classified as Tier I rather than less, casting serious doubt that the pre-reform scoring manipulation is welfare enhancing.

3 Empirical setting

3.1 The Massachusetts waste site cleanup program

Historically, Massachusetts provided "virtually no environmental regulation" of industrial activity and thousands of properties became contaminated with oil and hazardous material (Massachusetts Department of Environmental Protection, 2007). In 1983, the state began comprehensively regulating releases of hazardous substances, with MassDEP initially conducting site remediation and recovering cleanup costs from the responsible parties. However, MassDEP lacked sufficient resources to remedy pre-existing and new spills, and "the agency became backlogged to the point of ineffectiveness" (Seifter, 2006). Furthermore, cleanup

efforts often were not targeted to the most serious sites that pose the greatest threat. To address these shortcomings, in 1993 the state privatized much of the responsibilities for site assessment and cleanup. While specifics of the regulations have been revised numerous times over the past three decades, this privatized cleanup program remains in place.

Under this privatized process, the responsible party must notify MassDEP upon discovery of a hazardous spill.[11] In addition, the responsible party must hire a Licensed Site Professional (LSP) within one year to formally assess the severity of the site and report to MassDEP. The Tier Classification Opinion submitted by the LSP then ultimately determines the regulatory treatment of the site remediation. From the initial program privatization in late 1993 through early 2014, the core of this evaluation was the Numerical Ranking System (NRS), a worksheet completed by the LSP that quantitatively evaluates the spill's likely impact on local human and ecological populations. In April 2014, the NRS was replaced with a simplified tier classification process involving several binary criteria pertaining to the site.

As shown in Appendix Table A1, the NRS contains five components which are summed to form an overall score ranging from 18 to 1320 points. Four of the components respectively describe the potential exposure pathways, the volume and toxicity of the spilled substances, the potential impacts on nearby human populations and water supplies, and the potential impacts on nearby ecology. Of note, nothing about local income, property values, or potential economic impact of the contamination enters the scoring formula. Appendix Figure A2 shows the empirical contribution of each of the components to the total NRS score. Additionally, there is a component allowing for ad hoc adjustments of $\pm$ 0-50 points for "mitigating site-specific conditions," determined at the discretion of the LSP.[12]

Each site is assigned a tier classification based on its total NRS score. If the total score is below 350, the site is determined to be Tier II. Eighty-seven percent of sites are scored below 350.[13] Sites scored 350 or above are more serious and obtain a classification of Tier I.

[11] Notification is required within 2 hours, 72 hours, or 120 days, depending on the severity of the spill. Sources of spills may be stationary (e.g. an underground storage tank) or mobile (e.g. a fuel tanker truck). Petroleum products are by far the most frequently released chemicals, followed by aromatic hydrocarbons (like benzine, used to make lubricants and dyes), hydraulic fluids, and arsenic. Compared to Superfund sites, amounts released are fairly small. For instance, a typical spill of number 2 fuel oil is about 300 gallons. Appendix Figure A1 shows the locations of sites scored using the Numerical Ranking System.

[12] The mitigating site specific score is meant to address some particular sub-component(s) that the LSP determines is inaccurately measured for that site. The points allocated to this mismeasured sub-component, and the scoring criteria for that sub-component, constrains the size of the adjustment.

[13] A site with a NRS score below 350 may still be classified as Tier I if there is an "imminent hazard" associated with the site. Less than one percent of sites with scores below 350 have imminent hazards.

This tier is further subdivided into Tier IC (350-449, 8.1 percent of sites), Tier IB (450-549, 4 percent of sites), and Tier IA ($\geq$ 550, 0.9 percent of sites).

After a site is assigned its tier classification, a LSP (potentially the same one) must conduct the remediation, with the state providing direct oversight only for the most serious sites. Generally speaking, there are two ways for remediation to be considered as resolved. One option is to reduce site contamination to a level that poses "no significant risk," which is formally designated as a permanent solution of quality A1 or A2. Alternatively, if the pollution still poses some risk, the responsible party may be allowed to place statutory limitations on the use of the land, termed an Activity and Use Limitation (AUL).[14]

Throughout the cleanup process, a site's tier classification affects remediation costs in various ways. The most burdensome classification is Tier IA, and MassDEP may take lead of the remediation for these sites. Distinctions between Tiers II, IC, and IB are less stark, but there are numerous advantages of a Tier II classification. For one, mandatory site cleanup permits are less expensive for Tier II sites. In addition, responsible parties must notify local communities about waste sites, and public involvement activities are much less likely for Tier II sites. Most importantly, Tier II site cleanups receive less government scrutiny, both directly and indirectly via MassDEP audits.[15]

This Massachusetts setting exemplifies the tension of privatizing regulatory enforcement. Prior to privatization, the pace of hazardous waste site cleanups was slow and poorly targeted. Following privatization, the pace of cleanups rapidly improved: 3200 sites achieved a permanent solution within two years, including 700 sites that had "languished under the old rules with no clear way out of the cleanup process" (Massachusetts Department of Environmental Protection, 2007). However, while the pace of cleanups is dramatically better under privatization, the program has drawn criticism for the conflicts of interest that it creates (Seifter, 2006). Below, we provide empirical evidence that LSPs have tended to score sites in a way that favors their responsible party clients' interests rather than those of the public.

[14]For example, an Activity and Use Limitation might require that the property cannot be used for residential, daycare, schooling, or agricultural purposes, and prohibit any renovation involving subsurface excavation.

[15]We observe whether a site was audited by MassDEP, but the impact of tier status on audit likelihood is challenging to causally identify. Audit likelihood does discontinuously increase at the Tier I threshold. However, Tier I sites also take longer to remedy, which mechanically increases their cumulative likelihood of being audited. Furthermore, as we will show, the share of Tier I sites declines over time, so the average Tier I site is comparatively older and has had a longer period to be audited. Thus, we do not attempt to draw strong conclusions about the relationship between tier classification and audit likelihood.

3.2 Data

We compiled data from MassDEP (Massachusetts Department of Environmental Protection, 2019) on the universe of hazardous contamination sites in Massachusetts for spills that occurred during 1984 to June 2019. These data include details on each site location, the chemical(s) that were spilled, and the history of official actions taken throughout the remediation process such as the tier classification. For sites in this database that are scored using the Numerical Ranking System, we augmented the data by obtaining the NRS component scores directly from the websites that MassDEP hosts for each site. We additionally geocoded site locations and spatially joined these coordinates to Census Bureau shapefiles to obtain Census Tract-level characteristics for each site. As the privatized program began in 1993, most of our analyses use the 1990 Decennial Census as a consistent source of predetermined neighborhood characteristics (United States Census Bureau, 1990). Where noted, we also use data from the 2010 Census and American Community Survey (United States Census Bureau, 2010a,b). We use MassDEP regions and Massachusetts municipalities from Massachusetts Department of Geographic Information (2006, 2020).

Table 1 presents summary statistics for the population of 11,347 sites included in the NRS. In Panel [A], we show details of site scoring and measures of cleanup quality. The average NRS score is 250, with a standard deviation of 104. Recall that 350 points is the threshold separating Tier II from Tier I classification, and only 13 percent of sites are scored above 350 (Tier IA, IB, or IC). Across all sites, the average ad hoc adjustment via the Component VI sub-score is -0.46 points, and only 5.4 percent of all sites exhibit a negative discretionary adjustment. Per the NRS user manual, MassDEP "anticipates that a limited percentage of NRS classifications will require use of Section VI," and this is indeed the case; however, as we show below, the use of Component VI adjustments is far from uniform across the NRS score distribution.[16] Turning to cleanup quality, 58.3 percent of sites that have reached a permanent solution were cleaned to the highest quality (of A1 or A2), while 21.3 percent of permanent solutions involve an Activity and Use Limitation (AUL).

In Panel [B], we present statistics on Census attributes of the neighborhoods containing each site. The average site is located in a 1990 Tract that had average household earned income of $34,501; had a median home value of $167,862; was demographically 87.5 percent white; and had 48.7 percent of adult (25^{+}) population with any college education. As a point of reference (not shown in the table), these values respectively correspond to about

[16]The official site scoring manual states the discretionary component is used to "alter a site score to reflect unusual site conditions that may not be accurately assessed by the NRS."

the 57th, 61st, 29th, and 52nd percentiles across all Tracts statewide (unweighted).

4 Results

The model and predictions derived in Section 2 guide our empirical work. We examine bunching in the distribution of scores at the NRS Tier I threshold and estimate discontinuities for measures of site cleanup quality and for predetermined neighborhood characteristics. First, we document that LSPs intentionally manipulate site severity scores in favor of their clients. Next, we show that this score manipulation facilitates lower-quality remediation of sites. Finally, we find that the prevalence of score manipulation varies across neighborhoods and is more pronounced in Census Tracts with lower household incomes, lower home values, lower adult educational attainment, and larger populations of color.

4.1 Evidence of NRS score manipulation

To document evidence of manipulated site severity reporting, we begin by examining the distribution of NRS scores. We observe the official score reported by the LSP, z_i, which might differ from the true severity score that would be observed absent manipulation, z_i^*. Under the assumption that the distribution of z_i^* is continuous at the cutoff for tier classification, then any excess bunching in the distribution of z_i below the Tier I threshold is indicative of score manipulation. In the model presented in Section 2, the cost of misreporting leads LSPs to report manipulated scores that are barely below the tier threshold. However, optimization frictions may prevent precise control, especially given that some of the scoring criteria are large and discrete.[17] Because of this scoring discreteness, manipulation can lead to excess mass in the score distribution even inframarginal to the tier cutoff.[18]

In Figure 1, we plot the full distribution of the observed site severity scores. The discontinuity in the empirical distribution at the Tier I threshold is both visibly obvious and extremely unlikely to have arisen by chance. The McCrary (2008) log-density test statistic is -1.144 (se = 0.074), which is interpreted as the density at the threshold being more than three times as large approaching the Tier I cutoff from the left compared to the right. Our

[17] For instance, the possible assessments pertaining to a groundwater exposure pathway in NRS Component II include "None," "Evidence of contamination," "Potential exposure pathway," or "Likely or confirmed exposure pathway," with point values corresponding to these responses of 0, 20, 100, and 150.

[18] Even if LSPs find it preferable to misreport one of the more discrete criteria, perhaps due to ambiguity, there is no particular reason for the distribution of z_i^* to be discontinuous around tier classification thresholds, and the empirical distribution of scores is smooth away from the tier thresholds.

focus below is only on this tier threshold, but we note here that the other tier thresholds also exhibit significant discontinuities in distributional mass. In Appendix Figure A3, we zoom in to show the bunching at the higher tier cutoffs. The McCrary test statistic at the Tier IA/IB threshold is even larger at -1.689 (se = 0.273), which is of particular relevance as MassDEP provides direct oversight of Tier IA sites.

To quantify the magnitude of scoring manipulation, we use a bunching estimator adapted from Diamond and Persson (2016), Kleven (2016), and Chen et al. (2021). In brief, the methodology uses k-fold cross-validation with a grid search over possible widths of the manipulation region. For each fold of training data and guess of the manipulation region, we estimate parameters that best fit a log-normal distribution to the density *outside* of the manipulation region, with the quality-of-fit determined by the sum of squared errors. Then, using the estimated counterfactual density function and manipulation region, we calculate the mean squared error (MSE) for the hold-out testing sample. We then select the manipulation region that yields the smallest out-of-sample MSE, conditional on passing a statistical test for equivalence of the predicted and total density within the manipulation region, i.e. total excess mass equals total missing mass. Finally, we use the full sample of data outside of the chosen manipulation region to estimate the counterfactual log-normal distribution. Additional methodological details are provided in Appendix B.

The results of this bunching estimator are shown in Figure 2. We estimate that scoring manipulation occurs for sites with true scores between 350 and 399 points, with the manipulated scores downgraded to fall between 325 and 349 points. This manipulation region is indicated by the dashed vertical lines in the figure. The solid red curve shows a log-normal density function that is fit to the data excluding this region. Visually, this counterfactual density function closely fits the data for scores that are far from the Tier I/II cutoff. The estimated counterfactual density shows that 6.69 percent of mass would fall within the 350 to 399 score range. By comparing the data to the estimated counterfactual density, we find that 2.65 percent of mass is manipulated to be below 350, with a bootstrapped standard error of 0.18 percent. Quantitatively, this means that 39.65 percent (se = 2.72) of true scores in the above-350 manipulation region are downgraded to have a Tier II status.

Next, we provide evidence that the excess bunching in the NRS score distribution is intentional, rather than a statistical artifact. To do so, we examine the sub-score recorded by the LSP in Component VI for "mitigating site-specific conditions." As discussed in Section 3, this score component is an ad hoc adjustment at the discretion of the LSP. The maximum size of the adjustment is ± 0-50 points, and otherwise is only limited by the scoring rubric

for the sub-components being adjusted.[19] Figure 3(a) plots local averages of this sub-score against the overall site severity score in bins of ten points. In addition, we graph LOESS curves fit to the data. The local averages remain close to zero for scores up to 300 (50 points below the threshold), which is notable in light of the possible score adjustment range. As the total score approaches the tier threshold from the left, component VI becomes more and more negative, until there is a very noticeable discontinuity at the Tier I threshold. This pattern strongly supports that this component is used to push scores below the tier threshold.

Unsurprisingly, Figure 3(b) shows that this discontinuity is driven by downward adjustments. The prevalence of negative Component VI scores overall is fairly rare, with only 5.4 percent of all site scores exhibiting a downward adjustment. This is especially true for sites more than 50 points below the Tier I threshold. Even for sites scored between 300 and 329, only 10.7 percent are downward adjusted using Component VI. For sites scored between 330 and 349, nearly one-quarter are downward-adjusted. In contrast, not one of the 90 (Tier I) sites scored between 350 and 359 has a downward adjustment.

In Table 2, we provide the regression estimates corresponding to Figure 3, obtained using kernel-based local linear regression. In this and the following RD results tables, Column (1) shows the unconditional RD estimates and Columns (2) and (3) subsequently add year and MassDEP region fixed effects, while the specification shown in Column (4) includes county fixed effects.[20] These four columns use optimal bandwidths calculated using the methods of Calonico et al. (2014), while Column (5) shows results from a fixed bandwidth of 50 points. Standard errors for all specifications are heteroskedasticity-robust and bias-corrected, also using methods from Calonico et al. (2014). In Panel [A] of Table 2, we show the estimated discontinuity in the average Component VI sub-score at the Tier I threshold. These scores are 8.9 points (se = 1.46) higher just above the threshold compared to just below. This point estimate and its statistical significance remain very stable across the specifications. In Panel [B], we consider the likelihood that a site experienced a downward adjustment. The discontinuity is -0.266 (se = 0.03) at the tier threshold, and again the estimate and significance change little across specifications.

The evidence shown in Figure 3 and Table 2 provides a clear indication that the excess bunching of the site score distribution is intentional and that Component VI is a substantial

[19] For instance, being located within 500 feet of a private drinking well increases the site score by 25 points, and the LSP could determine that the contamination will not reach the well. In that case, the mitigating site-specific component would reduce the site score by 25 points.

[20] MassDEP is divided into four regional offices of Central, Northeast, Southeast, and West Massachusetts.

factor. Setting this component to zero and holding the other components constant would increase the share sites scored above 350 by 12.86 percent and the share of sites scored within 350-399 by 41.48 percent. Therefore, this discretionary component alone explains almost two-thirds of the total excess bunching.[21]

4.2 Evidence of reduced site cleanup quality

Having established that LSPs manipulate site severity scores to obtain more favorable regulatory treatment, we next evaluate whether responsible parties take a different approach to cleanup for these sites. Consistent with the model in Section 2, remediation quality is discontinuously inferior for Tier II sites, which likely leads to worse outcomes for manipulated sites than would be the case had they been correctly classified as Tier I. We examine two of the possible permanent solutions for a hazardous waste site. As described in Section 3, one official solution is to reduce contamination to a level which poses no significant risk to human or ecological populations. Another possible outcome is to impose an Activity and Use Limitation (AUL) on the site property, which limits the adverse impact of substances left in place by restricting the allowed uses of the land. Seeking an AUL and exerting cleanup effort are substitutes. The choice of which approach to use in remedying the site will vary discontinuously at the tier threshold if tier classification affects the effort expended on site cleanup.

Figure 4 shows utilization of these two types of permanent solution. In Panel (a), we plot how the likelihood of remediation to a level of "no significant risk" varies discontinuously at the Tier I/II threshold. Sites just barely qualifying as the less serious Tier II classification are substantially less likely to achieve this highest cleanup quality. Notably, there is little relationship between site severity and the likelihood of this permanent solution for sites scored well below the tier threshold. Only near the threshold is the likelihood of "no significant risk" noticeably reduced. In Panel (b), we plot an analogous pattern for the likelihood of an AUL as part of the permanent solution. As the figure shows, land use restrictions are much more prevalent for sites scored just below the tier threshold compared to those just above.

In Table 3, we present regression estimates that correspond to the evidence in Figure 4. As described above, all RD estimates use kernel-based local linear regression. Most specifications use the optimal bandwidth for that specification, while Column (5) uses a constant bandwidth of 50 points across all outcomes. Panel [A] shows estimates for the

[21] We do not further dissect the sub-components that lead to bunching because there is mechanical correlation between the various sub-score components, conditional on a total score.

discontinuity in the likelihood of sites' permanent solutions entailing "no significant risk." Consistent with the figure, we find that barely-Tier I sites are 31.1 percent more likely to achieve this highest quality of permanent solution (se = 8.1 percent). This finding is robust to the inclusion of year, region, and county fixed effects. When a fixed bandwidth of 50 is used in Column (5), the RD estimate increases somewhat, to 36.5 percent. Panel [B] of Table 3 presents similar estimates for land use limitations, with results that mirror those shown in Panel [A]. We find that the likelihood of an AUL decreases discontinuously at the Tier I/II threshold by 20.2 percent (se = 6.1 percent). Again, the estimated discontinuity is stable as we include year, region, and county fixed effects, and when specifying a fixed bandwidth of 50 points.

Given that only one-fifth of all site permanent solutions involve an AUL, these estimated differences in site remediation quality are substantial. Because these measures of cleanup effort are also observable by MassDEP, we do not view these discontinuities in remediation quality as evidence of shirking in the classic principal-agent sense (in which the agent's effort is unobserved by the principal). Rather, this evidence indicates that hazardous waste cleanup is approached differently depending on the intensity of government oversight.

4.3 Evidence of unequal treatment of neighborhoods

Our third set of results considers how scoring favoritism differs by the neighborhood (Census Tract) containing the hazardous waste site. The model in Section 2 supports three potential mechanisms for spatial heterogeneity in score manipulation. First, neighborhoods with higher willingness-to-pay (or ability-to-pay) for environmental amenities provide larger property value benefits to site owners for conducting a thorough cleanup; score manipulation should be less frequent in such neighborhoods. Second, neighborhoods with lower cleanup effort costs should also see less prevalent score manipulation. Finally, the reputation or psychic cost to LSPs of manipulation could be relatively higher in some areas.

We empirically identify how neighborhoods influence score manipulation using two approaches. First, we examine how predetermined socioeconomic characteristics vary across the Tier I/II threshold. If a neighborhood characteristic discontinuously increases across this threshold, this indicates that it is negatively associated with the likelihood of manipulation. That characteristic is thereby either positively related to environmental WTP, negatively related to cleanup effort cost, or it increases LSPs' manipulation cost. We evaluate four Census Tract-level covariates: average household earned income, median home values, the white population share, and the share of the adult (25 or older) population with any college.

Second, we estimate the counterfactual score density and manipulated density for subsamples of the data, e.g. sites in Census Tracts with above-median income. If manipulation is less prevalent in the above-median portion of sites for a given neighborhood characteristic, this likewise indicates the characteristic is negatively associated with manipulation.

The regression discontinuity results for these four neighborhood characteristics are presented visually in Figures 5 and 6, which maintain the same score range and ten-point local average bins as shown in the previous figures. The graphs for all four Census attributes show clearly-evident discontinuities at the tier threshold. Barely-Tier I sites are located in neighborhoods with visibly higher income, higher home values, higher white population share, and higher educational attainment.

To more formally quantify these discontinuities, Table 4 presents the corresponding RD estimates, again using the kernel-based local linear regression procedure and specifications described above. To more readily compare the magnitude of the coefficients across the different outcomes, we use the percentile of the Census Tract neighborhood characteristic in the distribution of all Massachusetts Census Tracts.[22] Panel [A] shows that the discontinuity in average annual household earned income is 13.1 percentiles (se = 3.3 percentiles). In other words, the average site scored just above the Tier I threshold is located in a Census Tract 13 percentiles higher in the Tract income distribution than the average site scored just below the threshold. This estimate remains large in magnitude and statistically significant with the inclusion of year, region, and county fixed effects. A similar pattern is shown for home values in Panel [B]. We find a discontinuity of 7.6 percentiles (se = 2.4), which changes little with the inclusion of year, region, and county fixed effects.

The latter two panels of Table 4 also show large and significant discontinuities in Census characteristics at the Tier I/II threshold. The white population share in Panel [C] increases by 17.4 percentiles at the tier threshold (se = 3.2). This point estimate is largely unaffected by the inclusion of year effects, but is somewhat attenuated to 9.9 percentiles (se = 2.7) when conditioning on region effects and 7.9 (se = 2.5) with county effects. This attenuation is perhaps not that surprising, given the geographic concentration of the non-white population in Massachusetts, which the spatial fixed effects control for much of. Panel [D] shows that the college share rises by an estimated 13.0 percentiles (se = 2.7) at the threshold. This estimate barely changes with the inclusion of year, region, and county effects, or with using the common bandwidth of 50 points as shown in Column (5). As with the other three Census outcomes, these estimated discontinuities are large and economically significant.

[22] Appendix Table A2 provides RD estimates for Census Tract neighborhood characteristic in levels.

As discussed in Section 2, evidence of discontinuous neighborhood characteristics at the tier cutoff indicates manipulation of the running variable in a regression discontinuity design. This analysis shows descriptively how the marginal neighborhood changes around the threshold due to heterogeneity in LSPs' propensity to manipulate the score depending on a site's location. To more directly demonstrate this heterogeneity, we use the bunching estimator described in Section 4.1 to estimate the counterfactual density and manipulated density separately for subsamples of the data. Specifically, we quantify the extent of manipulation for the eight subsamples of sites located in below median or above median Census Tracts based on household income, home values, white population share, and college education.

Table 5 presents the bunching estimates.[23] The first panel shows results using the full sample. As discussed above, Column (1) shows that the estimated manipulation region is 350 to 399 points and Column (2) shows that the estimated counterfactual density has 6.69 percent of sites scored within this region. Column (3) compares the data to this counterfactual, estimating that 2.65 percent of the counterfactual mass is "missing" in the manipulation region and shifted to be below 350, with a bootstrapped standard error of 0.18 percent. Although total manipulated mass is one way to quantify the extent of manipulation, it is less useful as a comparison across subsamples of sites, because the estimated widths and counterfactual densities of the manipulation regions vary across subsamples. As a more useful comparison, Column (4) shows the fraction of scores in the above-350 manipulation region that is manipulated downward, which is 39.65 percent (se = 2.72). This estimate can also be interpreted as the probability of manipulation for a score in the manipulation region.

The second panel of Table 5 presents results for the subsample of sites located in Census Tracts with below median household income. The third panel shows analogous results for sites with above median income. In Column (3), we find that the manipulated density is 2.99 percent (se = 0.22) in the lower-income portion of the state and 1.54 percent (se = 0.25) for sites with above median income. As shown in Column (4), the probability that a score in the manipulation region is manipulated is 54.06 percent for sites with below median income, compared to 27.71 percent for sites with above median income. That is, site scores in economically disadvantaged neighborhoods are almost twice as likely to be manipulated to fall below the Tier II threshold. We find generally similar results for the other three neighborhood characteristics. The probability of manipulation is 46.54 percent for sites with below median home value, compared to 30.08 percent for above median sites. Using white population share, the probabilities are respectively 53.64 and 28.19 percent. For college

[23]Plots for each subsample, akin to Figure 2, are provided in Appendix Figures A4-A7.

education, they are 55.16 and 24.06 percent. All estimates are economically and statistically significant, and consistently the pattern is that manipulated density and probability of manipulation are much greater in the socioeconomically disadvantaged neighborhoods.

Ultimately, these Census attributes capture spatial variation, and the four characteristics that we consider are (strongly positively) correlated with one another. A discontinuity in one measure might simply be due to a scoring choice that is influenced by another neighborhood characteristic. As an attempt to evaluate each socioeconomic attribute's marginal contribution to score manipulation, we estimate a specification that conditions on each of the socioeconomic explanatory variables in the same regression.[24] To do so, first we consider sites within 50 points of the Tier I/II threshold. This score region is both the range in which we predominantly find manipulated mass in the score distribution and is the scope for manipulation via the explicitly discretionary NRS Component VI. For sites with a total score between 300 and 400, we estimate how the four socioeconomic terms predict the likelihood that the site was scored above the Tier I threshold using the following regression:

$$\mathbb{1}\{z_{ijt} \geq 350\} = \beta_0 + B'\tilde{X}_i^{SE} + \rho_j + \gamma_t + \epsilon_{ijt}$$

where i, j, and t index site, location (alternatively MassDEP region or county), and year of tier assignment. As in Table 4, each socioeconomic measure is converted into the Tract's percentile across all Tracts in the state so that the coefficients of interest in the vector B will be comparable in magnitude. These percentile socioeconomic measures are captured in the vector $\tilde{X}_i^{SE}$. The specification also includes region or county fixed effects, ρ_j, and year fixed effects, γ_t.

Table 6 presents the results of this estimation. The first four columns show the univariate regressions for comparison. In the specification shown in Column (5), we include only the four socioeconomic measures as regressors. Specifications (6), (7), and (8) respectively add year, region, and county fixed effects. We find that both white population share and the college education share have a positive and significant relationship with the site being scored Tier I. Conversely, household earned income and median home values do not. For each ten percentile increase in the Tract's white population share, the likelihood that a site is scored above 350 increases by 1.4 percentage points (se = 0.36). Similarly, each ten percentile

[24] We have chosen to be parsimonious with the inclusion of socioeconomic covariates, because these measures are highly correlated and adding additional unnecessary controls can lead to unexpected sources of bias. Even with just a few such covariates, the results should be interpreted with caution. As an example, if regressing manipulation on income, adding a covariate such as housing values can introduce collider bias, or can indirectly control for a mediator if that mediator affects house prices.

increase in the college population share raises the likelihood of a site score being above 350 by 1.66 percentage points (se = 0.52). The size of the estimated effects of race and education effect are meaningful relative to the mean of the dependent variable, as only 21 percent of the sites in the 300-400 point score region are Tier I. The estimates are robust to the inclusion of year effects, but attenuate somewhat with the inclusion of region and county fixed effects. Upon inclusion of the latter, the estimated effect of white population share is 0.079 (se = 0.043) and 0.106 (se = 0.058) for college share.

As an additional approach, we regress the estimated excess (or missing) density at each site's NRS score on the neighborhood socioeconomic characteristics, using the same eight specifications described just above. Results for these estimations are shown in Appendix Table A3. The four univariate regressions show that each of the socioeconomic terms is strongly negatively associated with excess score density, consistent with the regression discontinuity estimates in Table 4 and the bunching estimates in Table 5. The multivariate regressions show that white population share has the strongest relationship with excess density, a result that remains generally robust to the inclusion of the temporal and spatial fixed effects.

On the whole, these results show that NRS score manipulation is less likely in neighborhoods that have higher educational attainment and a greater white population share, even conditional on local income and property values.[25] In the context of the model, this could operate through the environmental willingness-to-pay mechanism. College education reflects WTP if it increases knowledge about the health effects of pollution, or if college-educated residents are more informed about pollution siting. Alternatively, score manipulation might offer less scope for reduced cleanup effort in these areas, if a better-educated populace provides more community scrutiny of site cleanup quality.[26] Finally, LSPs' personal loss function for manipulation might be steeper in such areas, though we are unable to directly examine the possibility of racial discrimination or similar-to-me bias. Because the four Census Tract characteristics are highly and positively correlated with each other, we caution readers not to draw strong conclusions about mechanisms from the multivariate regressions.

Altogether, the relationships between Census attributes and site score manipulation indicate that the principal-agent problem we document above has a much more pronounced

[25]It is noteworthy that neighborhood income and home values do not predict score manipulation after controlling for race and education. One possibility is that real estate markets only loosely map to Tract boundaries, so that these measures poorly capture the property value boost from high-effort remediation.

[26]In support of this hypothesis, we examined formal community involvement in site remediation through Public Involvement Plans (PIP). The LSP for a PIP site must lead community meetings and present plans for site cleanup. While relatively few sites have a PIP, these activities are more common in higher-education neighborhoods, and our conversations with LSPs indicate that responsible parties fear having a site PIP.

impact on socioeconomically disadvantaged neighborhoods. Coupled with the results shown in Section 4.2 of inferior cleanup quality for barely-Tier II sites, the implication is that socioeconomically disadvantaged neighborhoods receive increased exposure to pollution through LSPs' disparate choices in scoring hazard sites.

4.4 Evidence from reform of tier classification process

This final results section evaluates the 2014 reform that MassDEP made to the tier classification procedure. As discussed in Section 3, this reform greatly simplifies the process by replacing the NRS scoresheet with a short set of binary criteria (among other changes). If the LSP indicates that any of the criteria are present, then the site is classified as Tier I. This overhaul was supported by the LSP Association as providing increased transparency and reduced paperwork. It also presumably reduces the degree of subjectivity available to the LSP in making his or her assessment.[27]

We utilize this reform to provide additional evidence supporting the disparate impact of score manipulation on socioeconomically disadvantaged neighborhoods. In Section 4.3 above, we show that household income, home values, racial composition, and college education levels for site neighborhoods all change discontinuously at the Tier I threshold. By removing some of the tier classification discretion from LSPs, the reform should lead to a narrowing of the socioeconomic differences between Tier I and Tier II sites.

First, we document that the reform substantially increases the likelihood of a site being classified as Tier I. In Figure 7, we show the share of sites receiving a Tier I classification by year. Between 1995 and 2005, the share of Tier I sites was 14.0 percent and fairly stable across years. After experiencing a slight uptick in 2006 and 2007, the Tier I share rapidly declined over the subsequent six years, reaching a low point in 2011 at 5.9 percent of sites. This is consistent with evidence from examining excess bunching in the NRS score distribution, which grew substantially over this time. In the year prior to the reform, only 11.2 percent of sites were Tier I. Then, post-reform the Tier I likelihood jumps substantially

[27]The criteria are: (i) Groundwater contamination that could affect sources of drinking water, where the concentrations of the hazardous materials exceed substance-specific thresholds. (ii) The contamination is an imminent hazard, which means that vapors exceed a quantitative threshold for the danger of an explosion, the release is on a roadway and endangers safety, or it is a risk to human health if present for even a short amount of time. (iii) Immediate remedial action (IRA) is required. An IRA can be triggered by any one of a number of situations, largely evaluated by objective criteria. Just to provide one example, an IRA is required if the released liquid "is detected in soil or groundwater during an underground storage tank (UST) removal or closure, at concentrations equal to or greater than 100 parts per million by volume, referenced to benzene, using a headspace screening methodology, and the sample was obtained within ten feet of the UST and more than two feet below the ground surface."

to 25.0 percent of sites, a proportion that has generally held since.

This increase in the Tier I share is not directly informative about LSPs' choices, as the reform changed the tier classification criteria in addition to reducing discretion. Instead, we use the reform to examine how the characteristics of site neighborhoods change as the classification process becomes more objective. The reduced subjectivity blunts the ability of LSPs to act on incentives for manipulation of tier classifications, and the socioeconomic gap between Tier I and Tier II sites should narrow as a result. Note that, because the reform increased the share of sites categorized as Tier I and the socioeconomic characteristics varied between Tier I and Tier II sites prior to the reform, the socioeconomic differences between tiers could also shrink mechanically. We cannot separately identify how much of a narrowing socioeconomic gap is mechanical from that due to a change in the LSPs' subjectivity. Rather, we interpret the evidence more generically as showing whether it is possible for a reform to make the treatment of sites more equitable.

Our evaluation uses difference-in-differences specifications of the form:

$$y_{it} = \alpha_1 \text{I}\{\text{Tier I}\}_i + \alpha_2 \text{I}\{\text{Post-reform}\}_i + \alpha_3 \text{I}\{\text{Tier I}\}_i \cdot \text{I}\{\text{Post-reform}\}_i + \gamma_t + \epsilon_{it}.$$

The dependent variables are the four socioeconomic measures examined earlier—the tract percentile in the state for average household income, median housing values, white population share, and the share of the adult population that has at least some college. $\text{I}\{\text{Tier I}\}_i$ is an indicator for whether site i is classified as Tier I. $\text{I}\{\text{Post-reform}\}_i$ indicates whether the site was tier-classified during the post-reform period. The coefficient of interest is α_3, which is interpreted as the change in the average value of y for Tier I sites compared to Tier II sites. If the reform closes socioeconomic gaps in tier classification, as we hypothesize, then the sign of α_3 should be opposite that of α_1. In other words, differences in neighborhood characteristics between Tier I and II sites should shrink in the post-reform period.

Table 7 presents these estimates for 2010 Census Tract-level attributes, using a sample period spanning 2010-2019. Prior to the reform, the estimated α_1 coefficients for the four measures all indicate generally similar socioeconomic differences as those shown above for the local averages near the tier threshold. Turning to the difference-in-differences coefficients of interest, three of the four coefficients indicate some reversal of the pre-reform disparities, with two showing that gaps are fully eliminated after the reform. The only exception is the white population share, for which the Tier I-II gap is large pre-reform (12.6 percentiles higher in Tier I sites) and does not appear to have narrowed after the reform. On the whole, however, this supplemental evidence from the tier reform corroborates our primary

analyses above and further supports that LSPs' score manipulation choices differ based on local neighborhood characteristics.

5 Conclusions

Public policy makers regularly turn to the private sector to assist with the administration of regulation. Privatizing compliance monitoring can ease fiscal burden and leverage firms' expertise, but it also introduces conflicts of interest: third-party evaluators may favor their regulated clients' objectives over those of the public. Thus, privatization can result in unintended consequences for the efficiency and equity of regulations.

Our paper examines this agency concern in the context of hazardous waste site remediation in Massachusetts. Following a spill, the responsible party must hire a private Licensed Site Professional (LSP) to assess and remedy the environmental contamination. While the state seeks an accurate evaluation of the hazard site, the responsible party may prefer a duplicitous reporting in order to reduce cleanup costs and minimize regulatory oversight.

By exploiting discontinuities in the mapping of LSPs' quantitative site evaluations into tiers of remediation regulations, we document three patterns of behavior in this setting. First, we show that LSPs' site assessments significantly favor their responsible party clients, a choice that is facilitated in part by the discretion given in the evaluation process. Second, we demonstrate that this client favoritism is associated with inferior cleanup quality, such as achieving remediation resolution through land use restrictions rather than by complete removal of the hazardous material. Finally, we find that these principal-agent problems are most pronounced for sites located in neighborhoods with lower income, lower property values, lower education, and a greater share of population of color.

Our study makes several contributions. Prior research typically finds beneficial effects of hazard site remediation for local property values and public health. Our findings demonstrate that there is substantial heterogeneity in site remediation quality depending on site-specific factors. Moreover, these findings add to a significant literature on environmental justice. We show that a lower willingness-to-pay or ability-to-pay for environmental remediation can elicit lighter regulation and reduced remediation quality, which in turn yields disparities in the exposure to pollution by race and socioeconomic advantage.

More broadly, our study speaks to the optimal design of mechanisms for tasking private third-party agents to serve in assessment and policy implementation capacities. Recent research highlights the importance of monitoring the actions of government agents and of

maintaining strong economic incentives for their honesty. Our findings illustrate that discretion by third-party evaluators can exacerbate incentives for misbehavior.

References

A. Banerjee, R. Hanna, B. A. Olken, and D. Sverdlin-Lisker. Social protection in the developing world. Working paper, 2022.

H. S. Banzhaf and R. P. Walsh. Do people vote with their feet? An empirical test of Tiebout. *The American Economic Review*, 98(3):843–63, 2008.

S. Banzhaf, L. Ma, and C. Timmins. Environmental justice: The economics of race, place, and pollution. *Journal of Economic Perspectives*, 33(1):185–208, 2019.

J. A. Blonz. The welfare costs of misaligned incentives: Energy inefficiency and the principal-agent problem. Resources for the Future Working Paper 18-28, 2018.

O. Borcan, M. Lindahl, and A. Mitrut. Fighting corruption in education: What works and who benefits? *American Economic Journal: Economic Policy*, 9(1):180–209, 2017.

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

E. Calvo, R. Cui, and J. Camilo Serpa. Oversight and efficiency in public projects: A regression discontinuity analysis. *Management Science*, Forthcoming, 2019.

Z. Chen, Z. Liu, J. C. S. Serrato, and D. Y. Xu. Notching R&D investment with corporate income tax cuts in China. *American Economic Review*, 111(7):2065–2100, 2021.

J. Cohn and T. Deryugina. Firm-level financial resources and environmental spills. NBER Working Paper w24516, 2018.

K. Crowder and L. Downey. Interneighborhood migration, race, and environmental hazards: Modeling microlevel processes of environmental inequality. *American Journal of Sociology*, 115(4):1110–1149, 2010.

J. Currie, M. Greenstone, and E. Moretti. Superfund cleanups and infant health. *The American Economic Review: Papers and Proceedings*, 101(3):435–441, 2011.

T. S. Dee, W. Dobbie, B. A. Jacob, and J. Rockoff. The causes and consequences of test score manipulation: Evidence from the New York regents examinations. *American Economic Journal: Applied Economics*, 11(3):382–423, 2019.

B. Depro, C. Timmins, and M. O'Neil. White Flight and coming to the nuisance: Can residential mobility explain environmental injustice? *Journal of the Association of Environmental and Resource Economists*, 2(3):439–468, 2015.

R. Diamond and P. Persson. The long-term consequences of teacher discretion in grading of high-stakes tests. NBER Working Paper w22207, 2016.

J. DiNardo and D. S. Lee. Economic impacts of new unionization on private sector employers: 1984–2001. *The Quarterly Journal of Economics*, 119(4):1383–1441, 2004.

E. Duflo, M. Greenstone, R. Pande, and N. Ryan. What does reputation buy? Differentiation in a market for third-party auditors. *The American Economic Review: Papers and Proceedings*, 103(3):314–319, 2013a.

E. Duflo, M. Greenstone, R. Pande, and N. Ryan. Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. *The Quarterly Journal of Economics*, 128(4):1499–1545, 2013b.

R. Fisman and Y. Wang. The distortionary effects of incentives in government: Evidence from China's "death ceiling" program. *American Economic Journal: Applied Economics*, 9(2):202–218, 2017.

S. Gamper-Rabindran and C. Timmins. Hazardous waste cleanup, neighborhood gentrification, and environmental justice: Evidence from restricted access Census Block data. *The American Economic Review: Papers and Proceedings*, 101(3):620–24, 2011.

S. Gamper-Rabindran and C. Timmins. Does cleanup of hazardous waste sites raise housing values? Evidence of spatially localized benefits. *Journal of Environmental Economics and Management*, 65(3):345–360, 2013.

K. Gillingham, S. Houde, and A. van Benthem. Consumer myopia in vehicle purchases: Evidence from a natural experiment. NBER Working Paper w25845, 2019.

W. B. Gray and R. J. Shadbegian. 'Optimal' pollution abatement – Whose benefits matter, and how much? *Journal of Environmental Economics and Management*, 47(3):510–534, 2004.

M. Greenstone and J. Gallagher. Does hazardous waste matter? Evidence from the housing market and the Superfund program. *The Quarterly Journal of Economics*, 123(3):951–1003, 2008.

S. Gupta, G. Van Houtven, and M. Cropper. Paying for permanence: An economic analysis of EPA's cleanup decisions at Superfund sites. *The RAND Journal of Economics*, pages 563–582, 1996.

K. Haninger, L. Ma, and C. Timmins. The value of brownfield remediation. *Journal of the Association of Environmental and Resource Economists*, 4(1):197–241, 2017.

C. Hausman and S. Stolper. Inequality, information failures, and air pollution. *Journal of Environmental Economics and Management*, 110:102552, 2021.

G. Z. Jin and J. Lee. A tale of repetition: Lessons from Florida restaurant inspections. *The Journal of Law and Economics*, 61(1):159–188, 2018.

H. J. Kleven. Bunching. *Annual Review of Economics*, 8(1):435–464, 2016.

M. Lavelle and M. Coyle. Unequal protection: The racial divide in environmental law. *National Law Journal*, 15(3):S1–S12, 1992.

T. D. Lytton and L. K. McAllister. Oversight in private food safety auditing: Addressing auditor conflict of interest. *Wisconsin Law Review*, 2014(2):289–336, 2014.

Massachusetts Department of Environmental Protection. The Massachusetts waste site cleanup program appendices: Measures of program performance 1993-2001. Technical report, Massachusetts Bureau of Waste Site Cleanup, 2007.

Massachusetts Department of Environmental Protection. Downloadable contaminated site lists [dataset], 2019. URL https://www.mass.gov/service-details/downloadable-contaminated-site-lists.

Massachusetts Department of Geographic Information. MassGIS data: MassDEP regions [dataset], 2006. URL https://www.mass.gov/info-details/massgis-data-massdep-regions.

Massachusetts Department of Geographic Information. MassGIS data: Municipalities [dataset], 2020. URL https://www.mass.gov/info-details/massgis-data-municipalities.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008.

P. Oliva. Environmental regulations and corruption: Automobile emissions in Mexico City. *Journal of Political Economy*, 123(3):686–724, 2015.

C. Persico, D. Figlio, and J. Roth. The developmental consequences of superfund sites. *Journal of Labor Economics*, 38(4):1055–1097, 2020.

M. Reynaert and J. Sallee. Who benefits when firms game corrective policies? CEPR Discussion Paper No. DP13755, 2019.

M. Seifter. Rent-a-regulator: Design and innovation in privatized governmental decision-making. *Ecology Law Quarterly*, 33:1091–1148, 2006.

R. J. Shadbegian and W. B. Gray. Spatial patterns in regulatory enforcement. In H. S. Banzhaf, editor, *The Political Economy of Environmental Justice*, chapter 9, pages 225–248. Stanford University Press, 2012.

J. P. Shimshack. The economics of environmental monitoring and enforcement. *Annual Review of Resource Economics*, 6:339–360, 2014.

United States Census Bureau. 1990 Census of population and housing, summary tape files 301, 310, 321, and 333 [dataset], 1990. URL https://www2.census.gov/census_1990/CD90_3A_26/.

United States Census Bureau. 2010 Census of population and housing, summary tape file DP1 [dataset], 2010a. URL http://factfinder2.census.gov.

United States Census Bureau. 2006-2010 American Community Survey 5-year estimates, tables B25077, DP03, and S1501 [dataset], 2010b. URL http://factfinder2.census.gov.

W. K. Viscusi and J. T. Hamilton. Are risk regulators rational? Evidence from hazardous waste cleanup decisions. *The American Economic Review*, 89(4):1010–1027, 1999.

J. West. Racial bias in police investigations. UC Santa Cruz working paper, 2018.

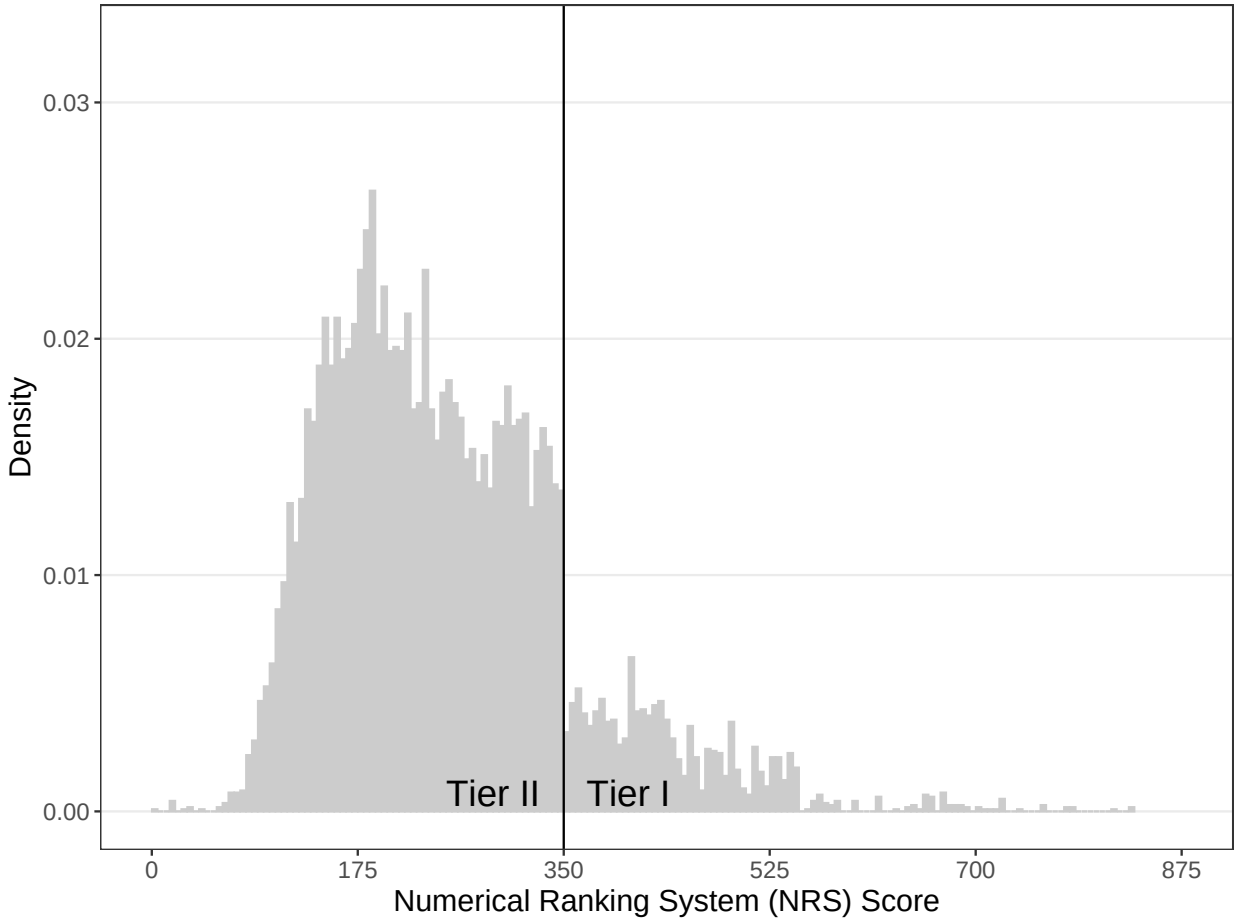
L. J. White. Markets: The credit rating agencies. *Journal of Economic Perspectives*, 24(2): 211–226, 2010.

Table 1: Summary statistics on sites in the Numerical Ranking System

	Mean	SD
Panel [A] Site scoring and cleanup quality		
NRS total score	250.092	103.620
NRS score above 350 (Tier I)	0.130	0.337
NRS component VI score	−0.462	9.965
Negative component VI score	0.054	0.226
Permanent Solution of A1 or A2	0.583	0.493
Permanent Solution includes AUL	0.213	0.410
Panel [B] Predetermined Census Tract covariates		
Household earned income ($000)	34.501	13.143
Median home value ($000)	167.862	64.944
White population share (%)	87.455	18.612
Adult pop. with any college (%)	48.709	16.903
Number of sites	11,347	

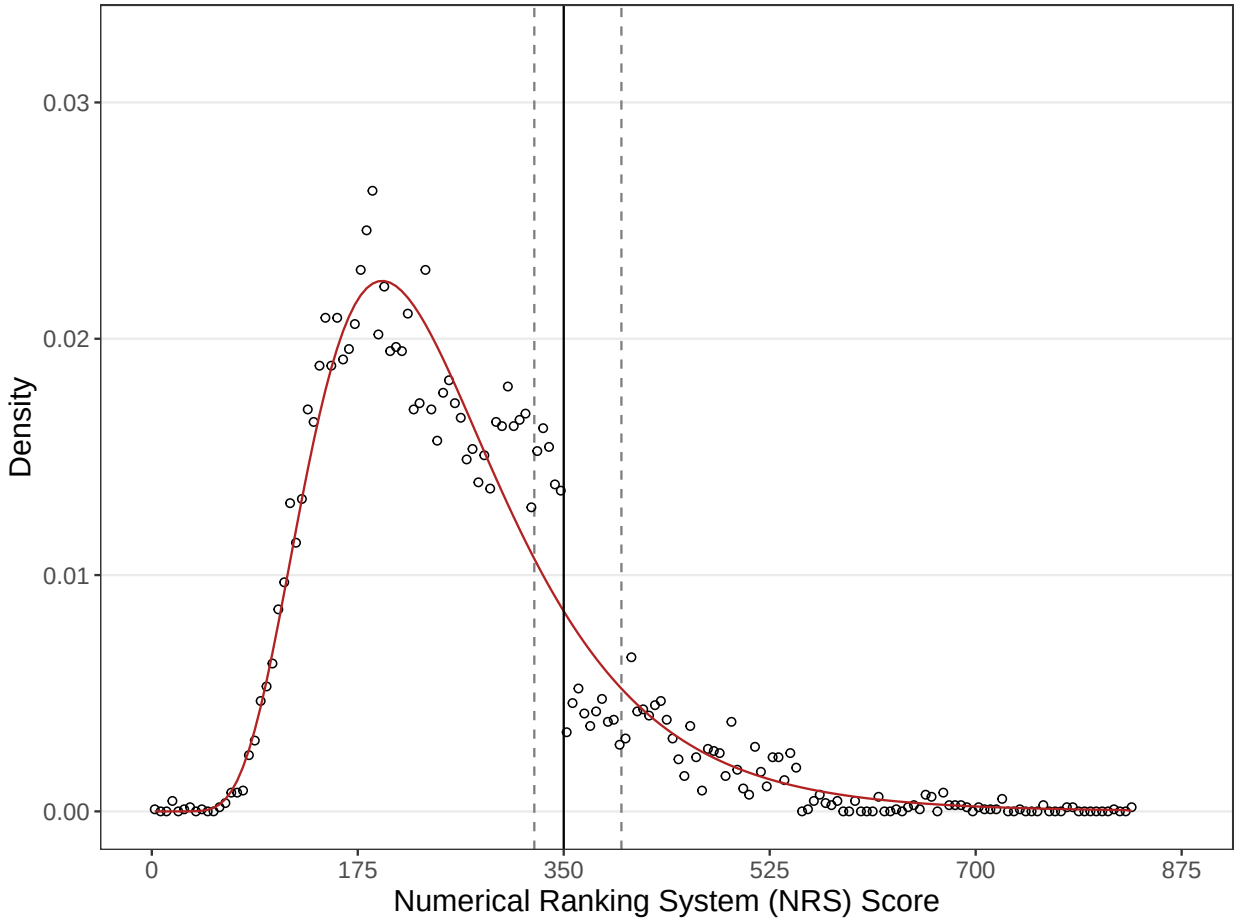
Notes: Summary statistics are for hazardous waste sites in the Numerical Ranking System (NRS). Panel [A] includes measures of site scoring and of the resulting cleanup quality for sites that have established a Permanent Solution through a Release Action Outcome. Panel [B] includes 1990 Census Tract economic and demographic covariates for the neighborhoods containing each site. The NRS component VI score is an ad hoc adjustment determined by the LSP for "mitigating disposal site-specific conditions" and has values between -50 and +50 points. A Permanent Solution of A1 or A2 is the highest possible cleanup quality and entails "No Significant Risk" to local human and ecological populations. An Activity Use Limitation (AUL) means that remediation resolution was obtained in part via land use restrictions rather than complete removal of the hazardous material. Adult pop. is persons over age 25.

Figure 1: Distribution of site scores in the Numerical Ranking System (NRS)



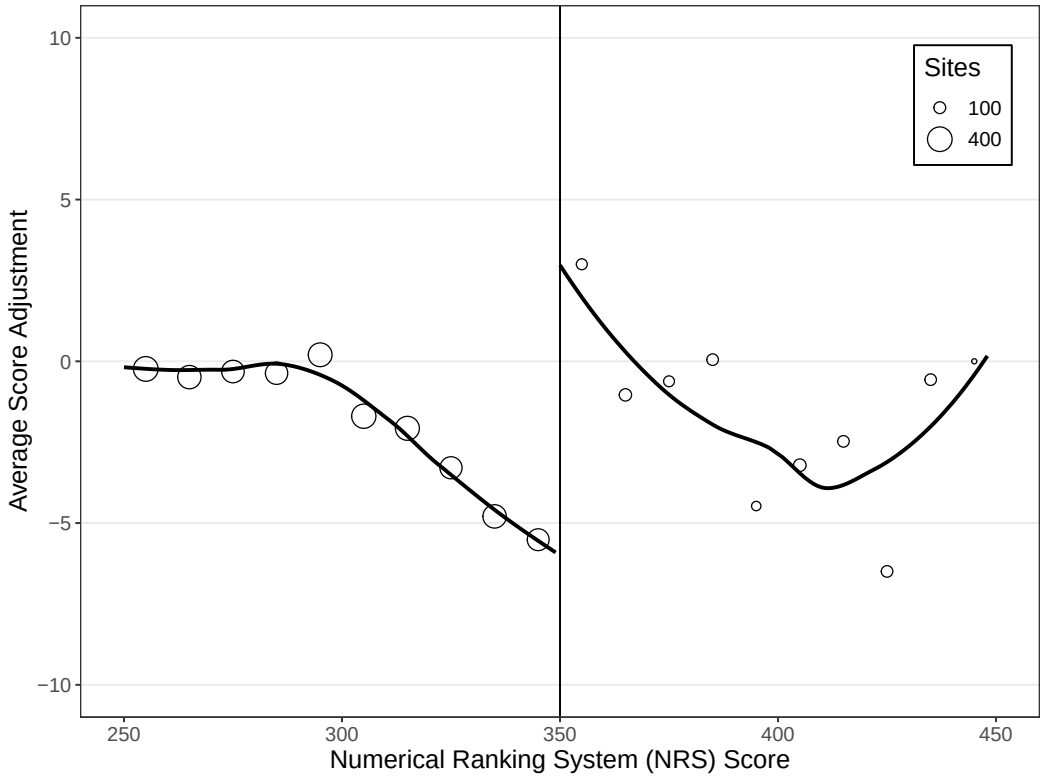
Notes: The figure plots the distribution of hazardous waste site scores in the Numerical Ranking System using a bin width of five points and showing the full set of 11,347 scores. The solid vertical line indicates the cutoff at 350 points between the Tier II and Tier I regulatory categories.

Figure 2: Estimated manipulation region and counterfactual density

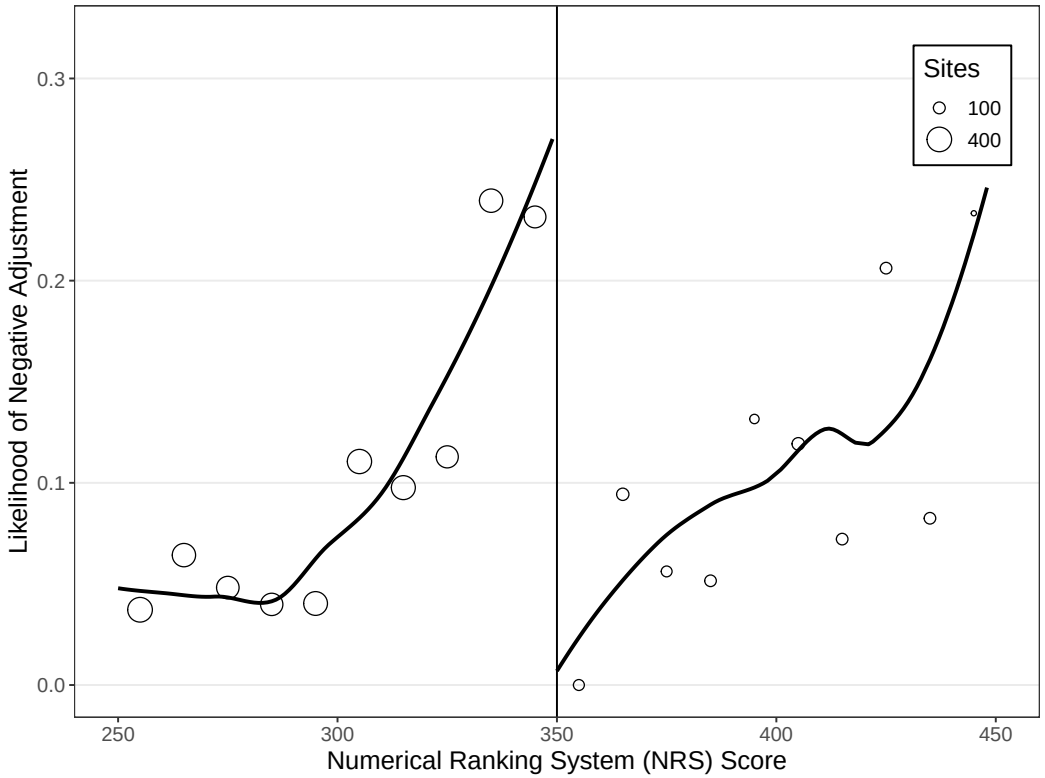


Notes: The figure plots the distribution of hazardous waste site scores in the Numerical Ranking System using a bin width of five points and showing the full set of 11,347 scores. The solid vertical line indicates the cutoff at 350 points between the Tier II and Tier I regulatory categories. The dashed vertical lines depict the estimated region over which score manipulation is present. The solid red curve shows a log-normal density function that is fit to the data excluding this manipulation region.

Figure 3: Score adjustments for "mitigating disposal site-specific conditions"



(a) Average NRS component VI score adjustment



(b) Likelihood of a negative NRS component VI score adjustment

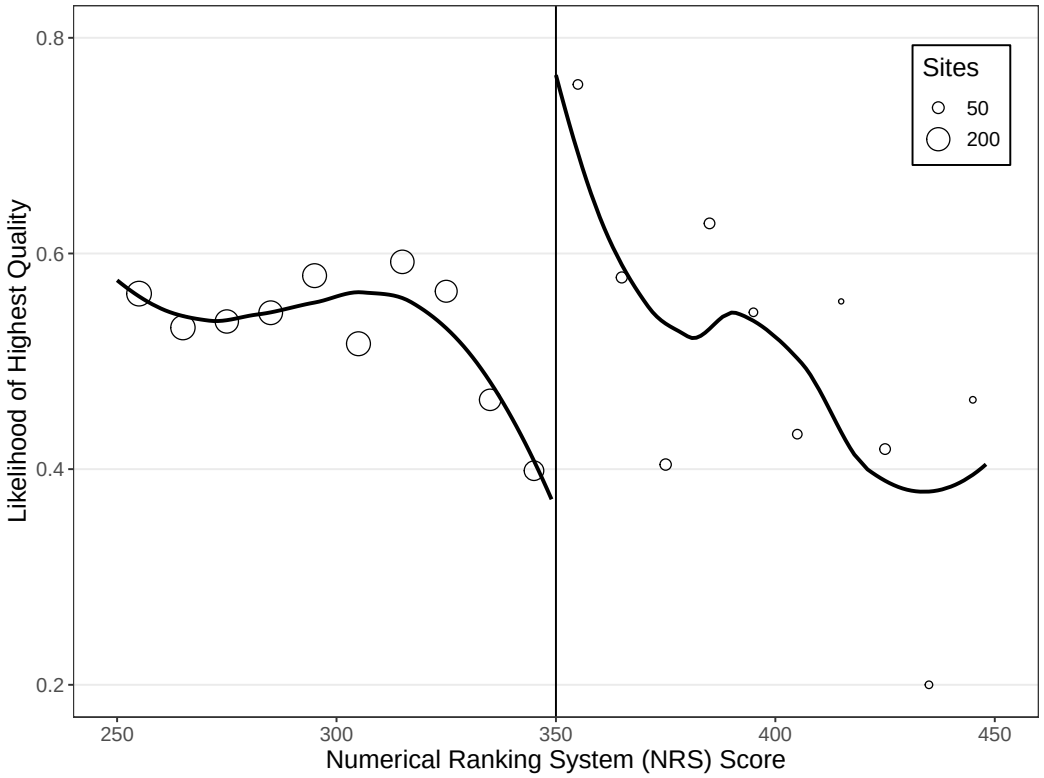
Notes: The figure plots local averages for the use of NRS component VI ad hoc score adjustments (ranging -50 to +50 points) against the total site NRS score, using a bin size of 10 points. The curves show a LOESS fit to the data separately on each side of the tier cutoff.

Table 2: NRS site scoring: Regression discontinuity estimates

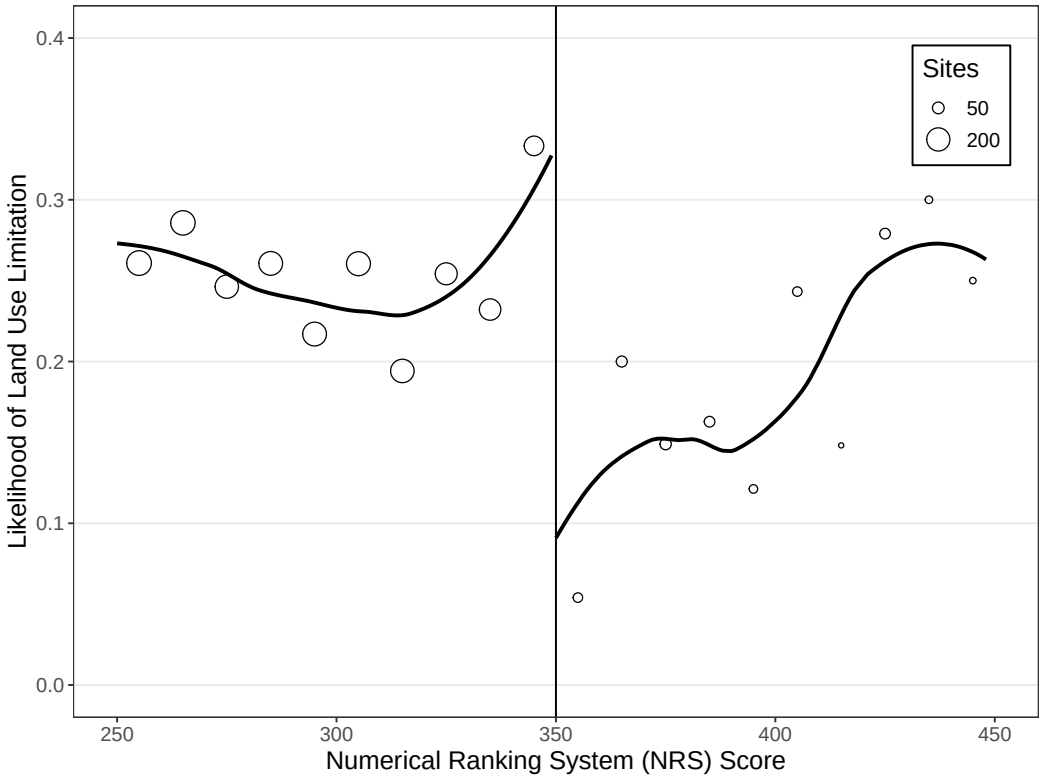
	(1)	(2)	(3)	(4)	(5)
Panel [A] NRS component VI score					
I{Tier I}	8.935	8.826	8.489	8.273	8.675
	(1.455)	(1.470)	(1.450)	(1.438)	(1.721)
Bandwidth	48.3	48.6	49.5	47.6	50
Observations	2,127	2,127	2,190	2,058	2,184
Panel [B] Has negative NRS component VI score					
I{Tier I}	-0.2661	-0.2801	-0.271	-0.2556	-0.26
	(0.030)	(0.031)	(0.030)	(0.031)	(0.036)
Bandwidth	45.2	44.3	45.5	44.8	50
Observations	1,982	1,966	1,982	1,962	2,184
BW selection	Optimal	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes	Yes
Region FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes

Notes: Each column presents results from a separate regression discontinuity estimation for how the outcome in each panel varies where crossing the Tier II to Tier I threshold at 350 total points in the Numerical Ranking System. All regressions use the "rdrobust" software package developed and provided by Calonico et al. (2014). Heteroskedasticity-robust bias-corrected standard errors are selected using the same package, as are optimal bandwidths using a triangular kernel. Where included, tier-assignment year FE are fixed effects for each year (1994-2013) of NRS site scoring, and region FE are fixed effects for each of the four MassDEP office regions.

Figure 4: Measures of cleanup quality for sites with a Permanent Solution



(a) Permanent Solution of A1 or A2: "No Significant Risk" (highest quality)



(b) Permanent Solution involves an Activity and Use Limitation for the property

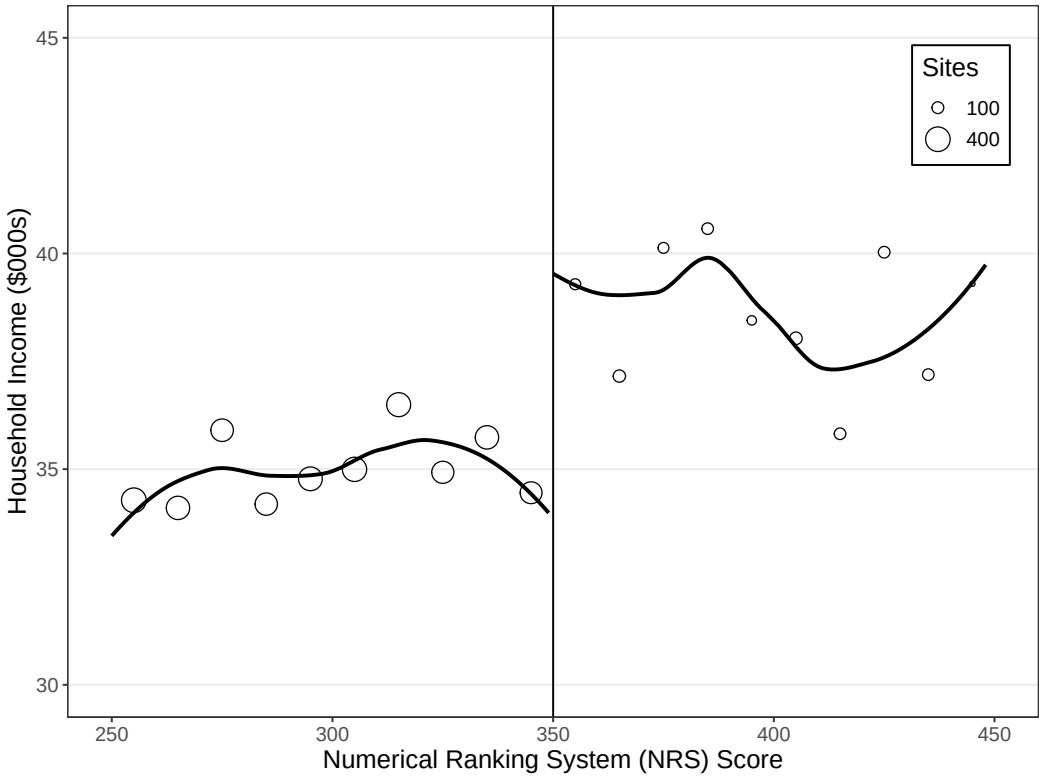
Notes: The figure plots local averages for measures of cleanup quality for sites with a Response Action Outcome Permanent Solution against the total NRS score, using a bin size of 10 points. The curves show a LOESS fit to the data separately on each side of the tier cutoff.

Table 3: Site remediation quality: Regression discontinuity estimates

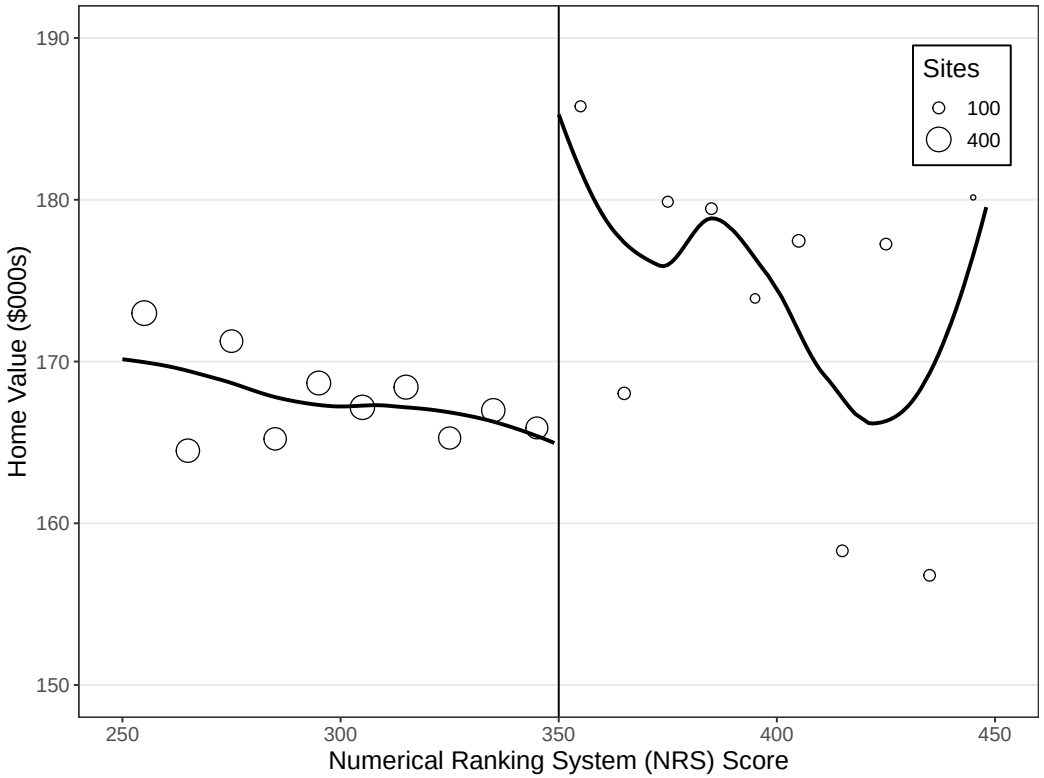
	(1)	(2)	(3)	(4)	(5)
Panel [A] Highest quality: "No Significant Risk"					
I{Tier I}	0.3112	0.2676	0.2603	0.2416	0.3646
	(0.081)	(0.078)	(0.076)	(0.076)	(0.104)
Bandwidth	59.4	60.7	61.6	62.5	50
Observations	1,354	1,362	1,366	1,390	1,093
Panel [B] Has land use limitation (AUL)					
I{Tier I}	-0.2019	-0.1872	-0.1709	-0.1559	-0.2105
	(0.061)	(0.062)	(0.059)	(0.057)	(0.081)
Bandwidth	61.6	63.1	68.1	69.8	50
Observations	1,364	1,437	1,556	1,586	1,093
BW selection	Optimal	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes	Yes
Region FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes

Notes: Each column presents results from a separate regression discontinuity estimation for how the outcome in each panel varies where crossing the Tier II to Tier I threshold at 350 total points in the Numerical Ranking System. All regressions use the "rdrobust" software package developed and provided by Calonico et al. (2014). Heteroskedasticity-robust bias-corrected standard errors are selected using the same package, as are optimal bandwidths using a triangular kernel. Where included, tier-assignment year FE are fixed effects for each year (1994-2013) of NRS site scoring, and region FE are fixed effects for each of the four MassDEP office regions.

Figure 5: Predetermined economic characteristics for neighborhood of site



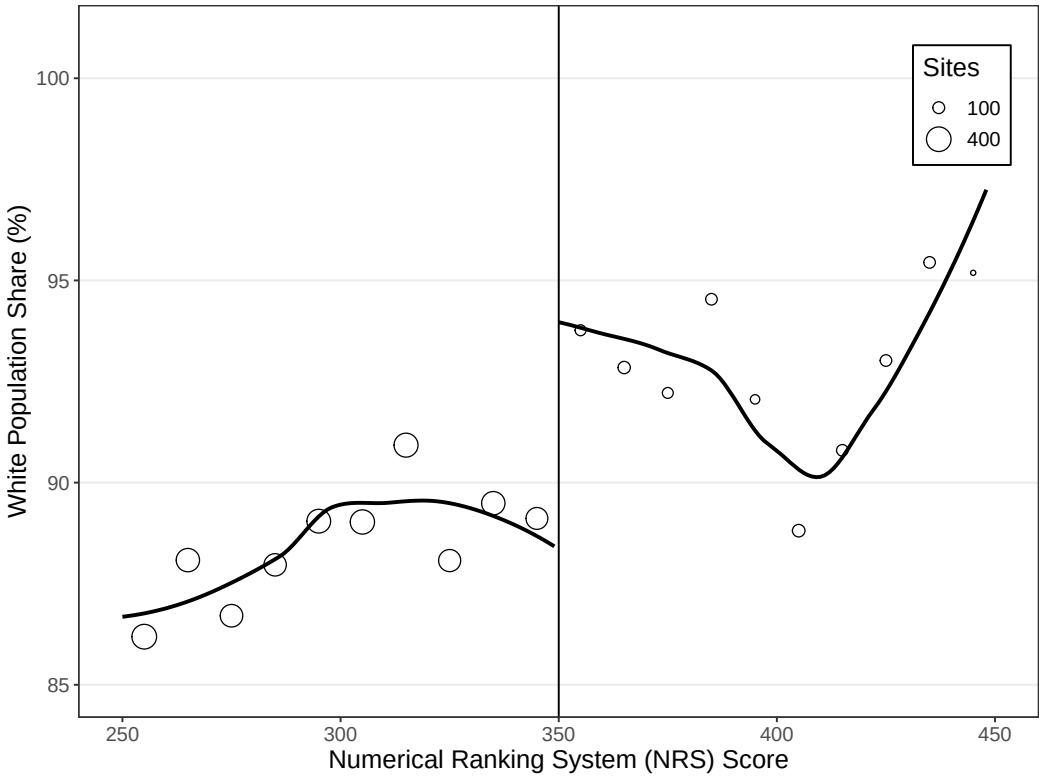
(a) 1990 Census Tract-level average household earned income ($000s)



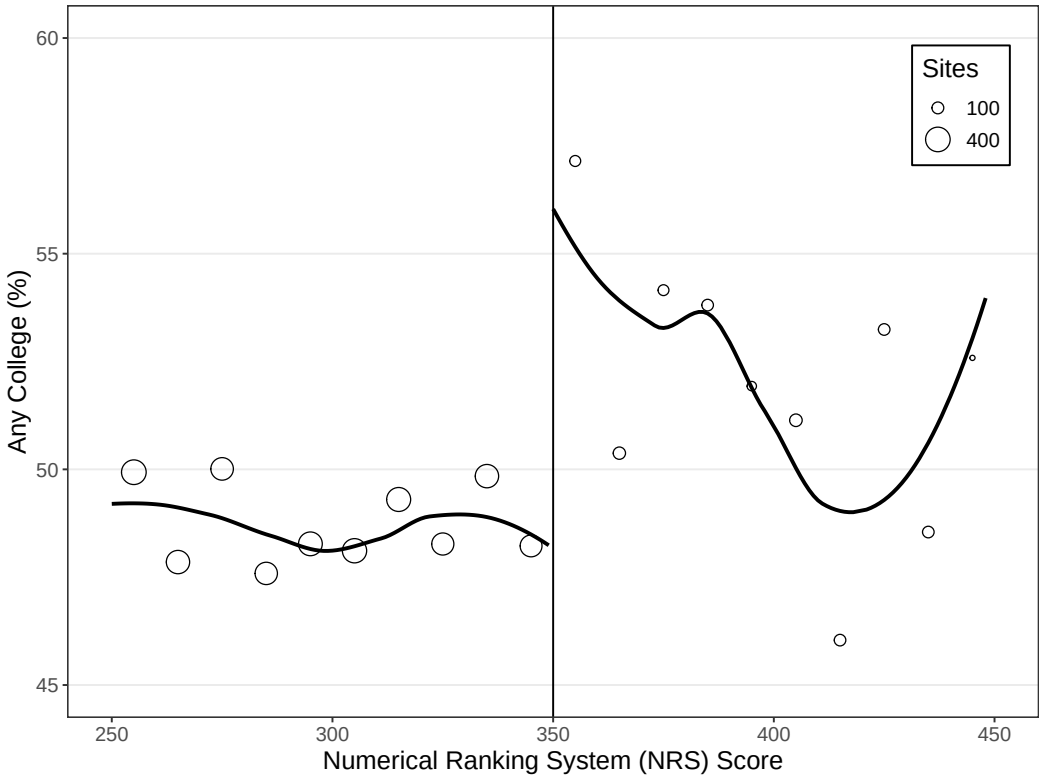
(b) 1990 Census Tract-level median home property value ($000s)

Notes: The figure plots local averages for 1990 Census Tract-level average household earned income and median home value against the total site NRS score, using a bin size of 10 points. The curves show a LOESS fit to the data separately on each side of the tier cutoff.

Figure 6: Predetermined demographic and education characteristics of neighborhood



(a) 1990 Census Tract-level white population share of residents



(b) 1990 Census Tract-level fraction of adult residents with any college education

Notes: The figure plots local averages for 1990 Census Tract-level demographic composition and adult (aged 25^{+}) college education against the total site NRS score, using a bin size of 10 points. The curves show a LOESS fit to the data separately on each side of the tier cutoff.

Table 4: Predetermined neighborhood characteristics: Regression discontinuity estimates

	(1)	(2)	(3)	(4)	(5)
Panel [A] Average household income (percentile in state)					
I{Tier I}	0.1314	0.1195	0.1186	0.07163	0.08668
	(0.033)	(0.033)	(0.032)	(0.026)	(0.037)
Bandwidth	52.2	49.4	47.4	55.1	50
Observations	2,273	2,184	2,058	2,435	2,184
Panel [B] Median home value (percentile in state)					
I{Tier I}	0.07592	0.07244	0.1129	0.07684	0.07604
	(0.024)	(0.029)	(0.023)	(0.019)	(0.030)
Bandwidth	101	67.7	66.8	66.4	50
Observations	4,378	2,905	2,867	2,867	2,153
Panel [C] White population share (percentile in state)					
I{Tier I}	0.1736	0.1508	0.0991	0.07948	0.1056
	(0.032)	(0.030)	(0.027)	(0.025)	(0.033)
Bandwidth	41.1	42.4	47.6	46	50
Observations	1,774	1,822	2,058	1,992	2,184
Panel [D] Adult pop. with any college (percentile in state)					
I{Tier I}	0.1296	0.1234	0.1308	0.08973	0.128
	(0.027)	(0.028)	(0.027)	(0.022)	(0.035)
Bandwidth	61.6	58	56.4	66.2	50
Observations	2,708	2,605	2,450	2,918	2,184
BW selection	Optimal	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes	Yes
Region FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes

Notes: Each column presents results from a separate regression discontinuity estimation for how the outcome in each panel varies where crossing the Tier II to Tier I threshold at 350 total points in the Numerical Ranking System. All regressions use the "rdrobust" software package developed and provided by Calonico et al. (2014). Heteroskedasticity-robust bias-corrected standard errors are selected using the same package, as are optimal bandwidths using a triangular kernel. Where included, tier-assignment year FE are fixed effects for each year (1994-2013) of NRS site scoring, and region FE are fixed effects for each of the four MassDEP office regions.

Table 5: Counterfactual density and manipulated density: Structural estimates

	Manip. region (1)	Counterf. density (2)	Missing density (3)	Share manip. (4)	Obs.
Full sample	[350, 399]	0.0669	0.0265 (0.0018)	0.3965 (0.0272)	11,347
Below median income	[350, 397]	0.0554	0.0299 (0.0022)	0.5406 (0.0391)	5,583
Above median income	[350, 383]	0.0554	0.0154 (0.0025)	0.2771 (0.0451)	5,738
Below median home value	[350, 398]	0.0645	0.0300 (0.0024)	0.4654 (0.0373)	5,452
Above median home value	[350, 388]	0.0573	0.0172 (0.0025)	0.3008 (0.0437)	5,612
Below median white pop.	[350, 392]	0.0497	0.0266 (0.0021)	0.5364 (0.0418)	5,471
Above median white pop.	[350, 398]	0.0759	0.0214 (0.0028)	0.2819 (0.0369)	5,850
Below median college	[350, 400]	0.0650	0.0358 (0.0023)	0.5516 (0.0358)	5,631
Above median college	[350, 387]	0.0562	0.0135 (0.0027)	0.2406 (0.0472)	5,690

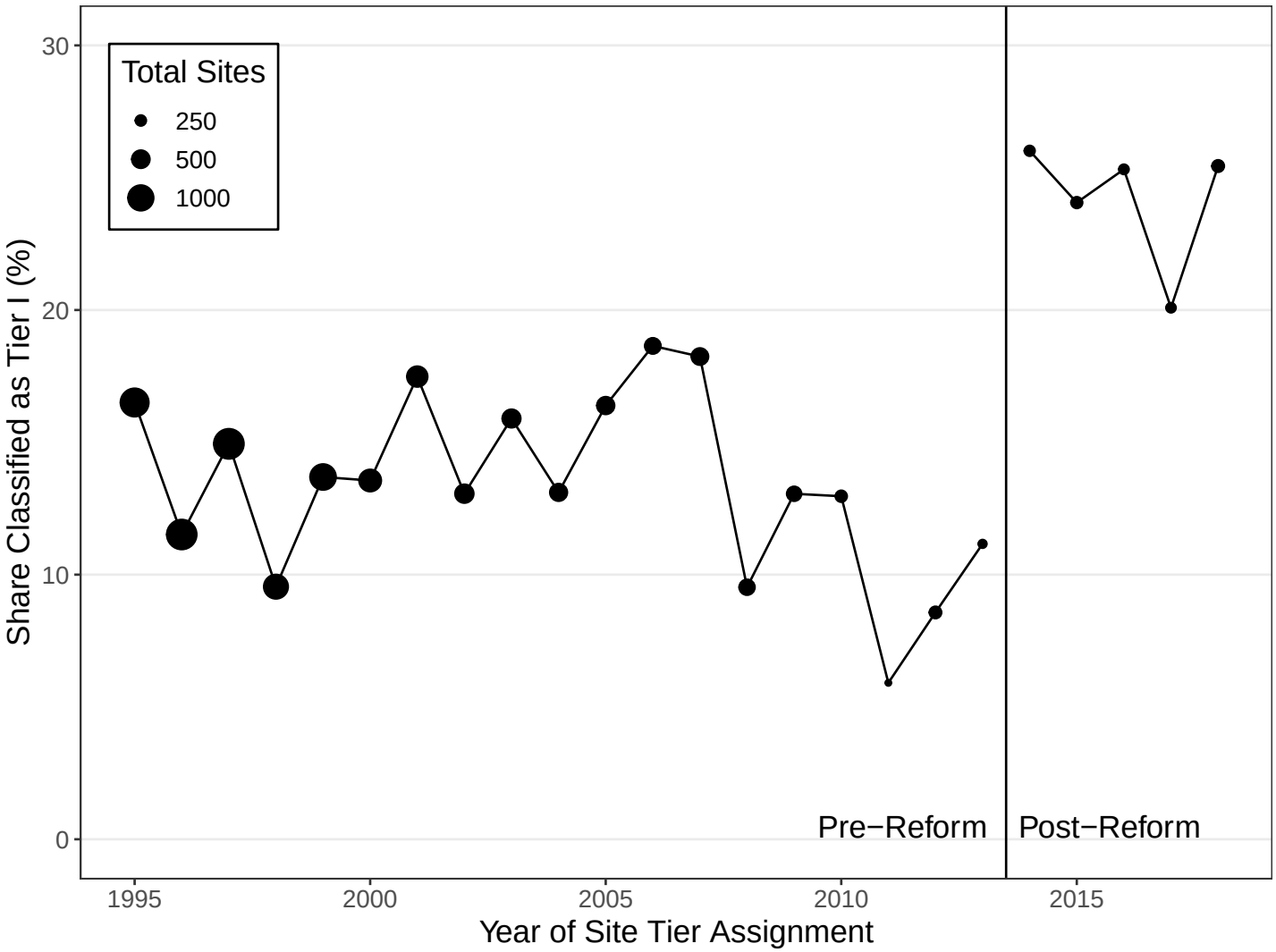
Notes: This table presents the results of the structural estimation for the counterfactual density and manipulation region. Each panel uses sites within the indicated subsample of Census Tracts. Column (1) shows the estimated score domain of the manipulation region above the Tier I threshold of 350 points. Column (2) shows the estimated counterfactual density within this region. Column (3) shows the estimated score density that is manipulated to be below 350. Column (4) shows the estimated fraction of scores in the manipulation region that are manipulated to be below 350. Bootstrapped standard errors are in parentheses.

Table 6: Relationship between NRS scores of above 350 and neighborhood characteristics: Linear regression estimates

	Dep. variable: Score between 350-400							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Household earned income	0.159				0.005	−0.009	−0.044	−0.020
	(0.033)				(0.051)	(0.050)	(0.052)	(0.065)
Median home value		0.130			−0.016	−0.005	0.084	0.082
		(0.034)			(0.049)	(0.049)	(0.059)	(0.063)
White population share			0.186		0.142	0.127	0.111	0.079
			(0.032)		(0.036)	(0.036)	(0.039)	(0.043)
Adult pop. with any college				0.195	0.166	0.167	0.128	0.106
				(0.034)	(0.052)	(0.052)	(0.054)	(0.058)
Dep. variable mean	0.208	0.208	0.208	0.208	0.208	0.208	0.208	0.208
Year fixed effects	No	No	No	No	No	Yes	Yes	Yes
Region fixed effects	No	No	No	No	No	No	Yes	Yes
County fixed effects	No	No	No	No	No	No	No	Yes
Observations	2,178	2,178	2,178	2,178	2,178	2,178	2,178	2,178

Notes: Each column presents results from a linear regression of a binary indicator for whether the NRS score is between 350-400 on the four 1990 Census Tract covariates, expressed as percentiles within the state. Only sites with an NRS score of between 300 and 400 are included in these regressions. Heteroskedasticity-robust standard errors are in parentheses. Where included, the year fixed effects are for the year of tier assignment and region fixed effects are for each of the four MassDEP office regions.

Figure 7: Tier composition of newly-classified sites by year during 1995-2018



Notes: The figure plots the annual share of hazardous waste sites that were classified each year by Licensed Site Professionals as being a Tier I site. The size of the markers indicates the total number of newly-classified waste sites each year. The solid vertical line indicates the state's overhaul of the Numerical Ranking System and revisions to the tier classification process that went into effect in 2014.

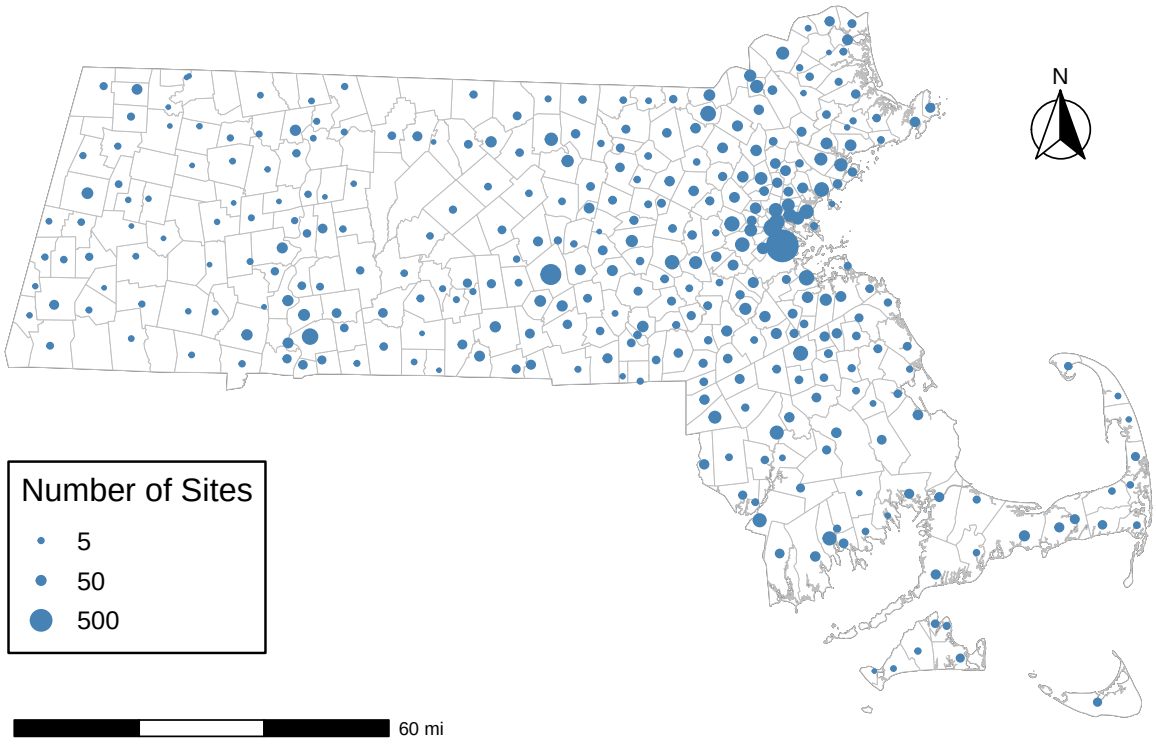
Table 7: Tier reform and neighborhood characteristics: Difference in differences estimates

	Dep. variable: Percentile in state			
	Income	Home value	White pop.	College
	(1)	(2)	(3)	(4)
I{Tier I}	0.149	0.077	0.126	0.074
	(0.024)	(0.026)	(0.024)	(0.025)
I{Tier I} X I{Post-reform}	−0.091	−0.097	0.014	−0.093
	(0.030)	(0.031)	(0.031)	(0.031)
Years included	2010-2019	2010-2019	2010-2019	2010-2019
Year fixed effects	Yes	Yes	Yes	Yes
Observations	2,236	2,236	2,236	2,236

Notes: Each column presents results from a separate difference in differences regression for how the Census 2010 Tract-level outcome indicated in the column titles changes following the 2014 reform to the tier classification process. Column (1) uses average household earned income. Column (2) uses the median home value. Column (3) uses the white population share. Column (4) uses the share of adults (25 or older) with any college attainment. Each outcome is expressed as the Census Tract's percentile within the state. Heteroskedasticity-robust standard errors are in parentheses. The year fixed effects are for the year of tier assignment.

A Appendix figures and tables

Figure A1: Map of site locations across municipalities in Massachusetts



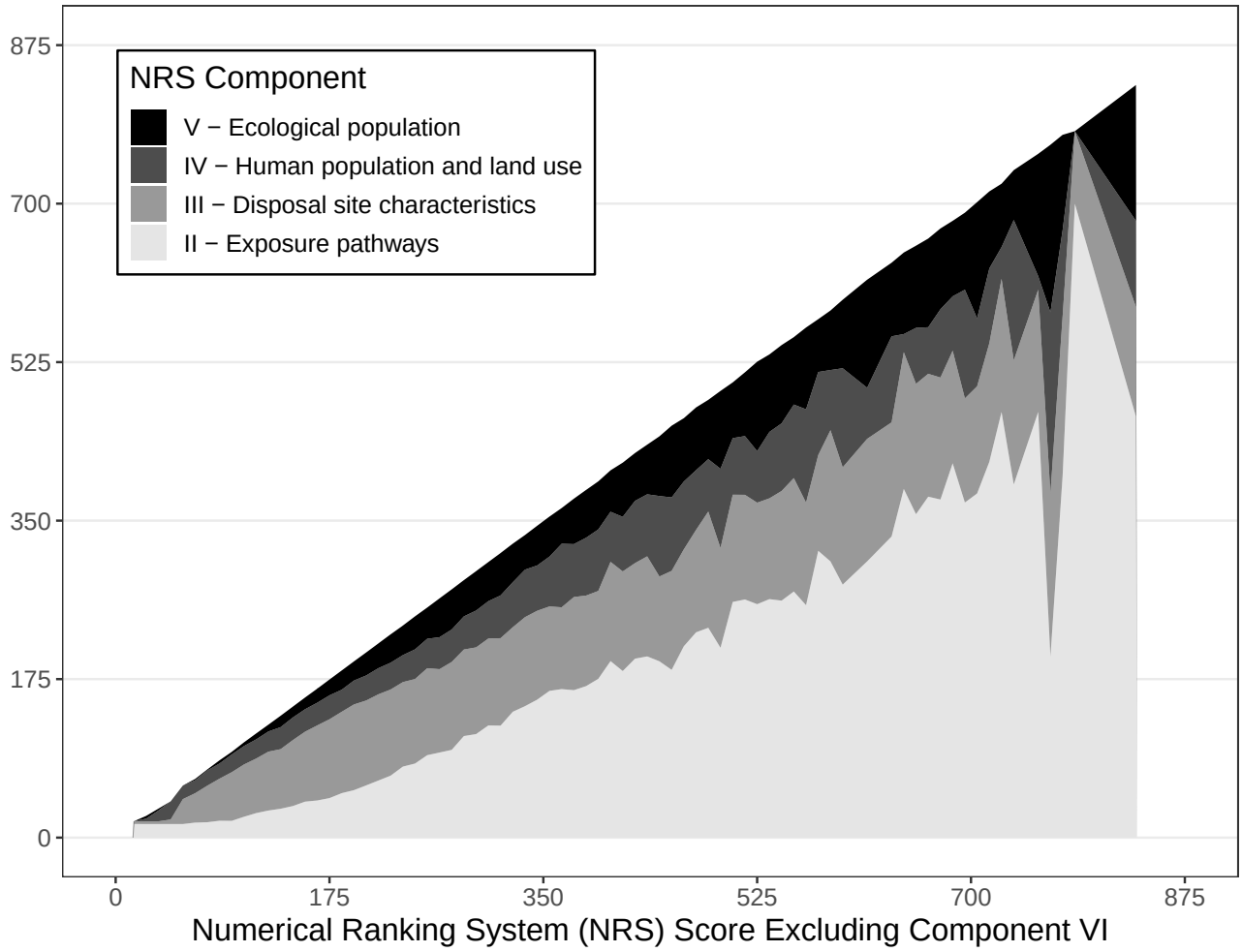
Notes: The figure shows the locations of hazardous waste sites scored using the Numerical Ranking System. The polygons show boundaries for municipalities, and the points are centered at the average coordinates of sites within each municipality. The size of the points indicates the total number of sites scored within each town.

Table A1: Numerical Ranking System components and possible score ranges

Component	Score range
I. Disposal site information	*[Not scored]*
II. Exposure pathways	*[15 – 700]*
Soil (likely presence, human exposure)	0 – 150
Groundwater (likely presence, human exposure)	0 – 150
Surface water (likely presence, human exposure)	0 – 150
Air (likely presence, affecting occupied buildings)	0 – 200
Number of sources (one, two, three or more)	0 – 50
III. Disposal site characteristics	*[3 – 180]*
Toxicity score (substance type, amount)	1 – 80
How many highly toxic substances? (none/one, more than one)	0 – 30
Substance mobility and persistence (low, medium, high)	0 – 50
Site hydrogeology (depth to groundwater, soil permeability)	2 – 20
IV. Human population and land uses	*[0 – 205]*
Population (people <0.5 mi., institutions <500ft., on-site workers)	0 – 40
Above an aquifer (no, potentially productive, or sole source)	0 – 40
Water use (proximity to public and private water supplies)	0 – 125
V. Ecological populations	*[0 – 185]*
Resource area analysis (wetlands, fish habitat, protected species)	0 – 150
Environmental toxicity analysis (substance types, concentration)	1 – 35
VI. Mitigating disposal site-specific conditions	*[± 0 – 50]*
Statutory total score range	18 – 1320
Empirical total score range	3 – 831

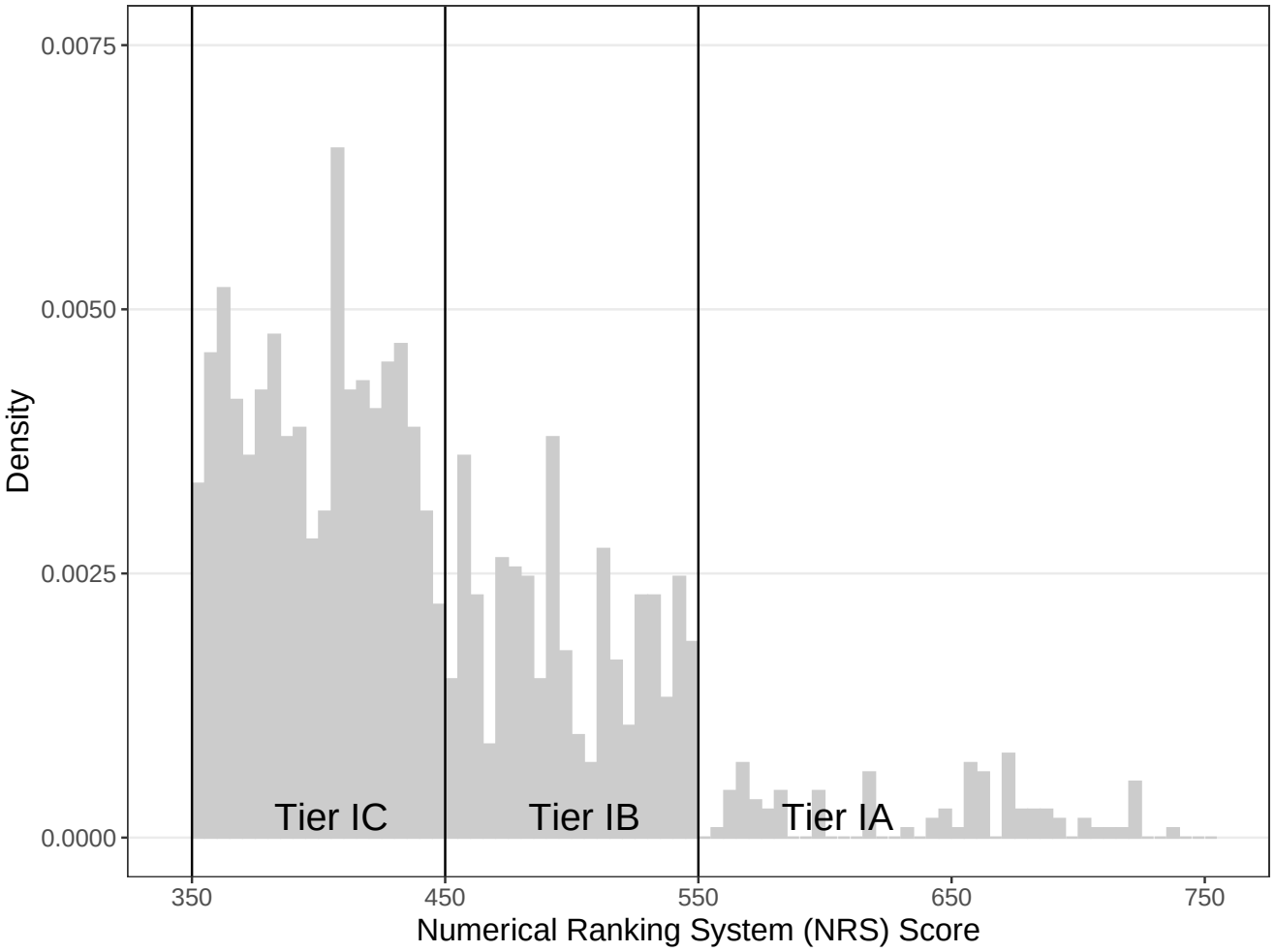
Notes: Values are sourced from the Numerical Ranking System Guidance Manual (310 CMR 40.1500). This manual of more than 80 pages is "written to assist users of the Numerical Ranking System developed by the Massachusetts Department of Environmental Protection to classify disposal sites as defined by the Massachusetts Contingency Plan and Massachusetts General Law."

Figure A2: Component contributions to total scores of sites in the Numerical Ranking System



Notes: The figure plots stacked area regions for the four component sub-scores in the Numerical Ranking System, excluding the discretionary component VI (which can take values between +/- 50 points). Note that there are very few sites with scores at the far right tail of the distribution.

Figure A3: Distribution of site scores in the NRS zoomed-in to 350-750



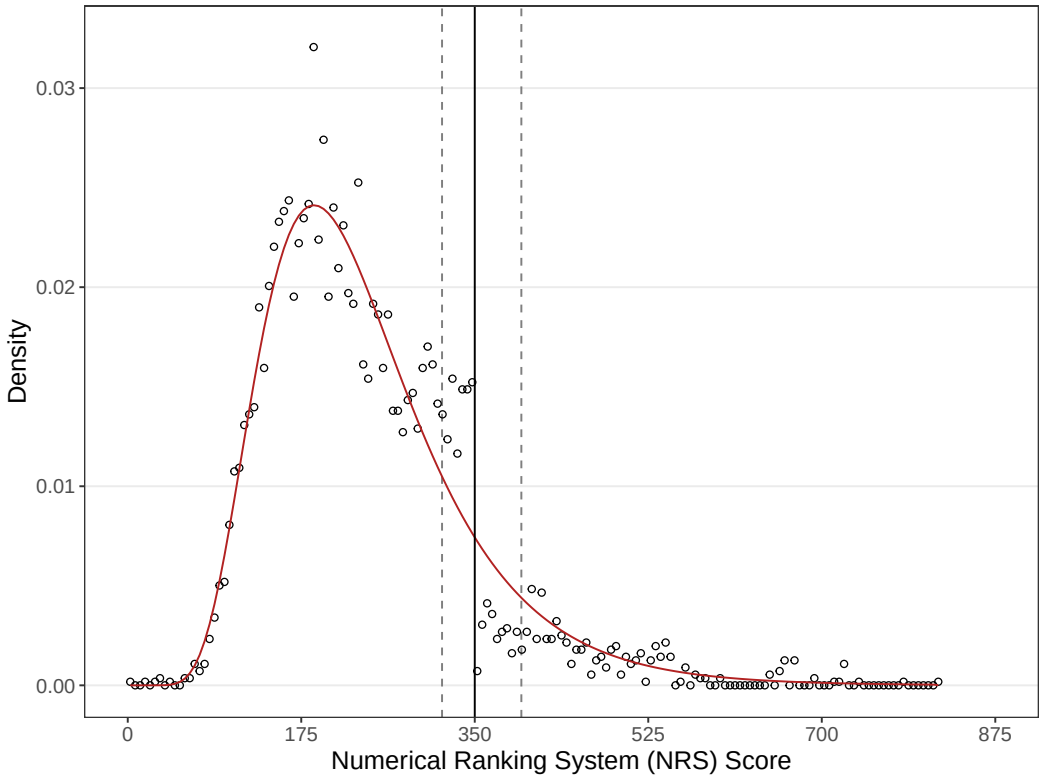
Notes: The figure plots a portion of the distribution of hazardous waste site scores in the Numerical Ranking System using a bin width of 10 points and showing the set of Tier I scores with values between 350-750. The solid vertical lines indicate the cutoffs at 350 points, 450 points, and 550 points, respectively between the Tier II/IC, Tier IC/IB, and Tier IB/IA regulatory categories.

Table A2: Predetermined neighborhood characteristics: Regression discontinuity estimates

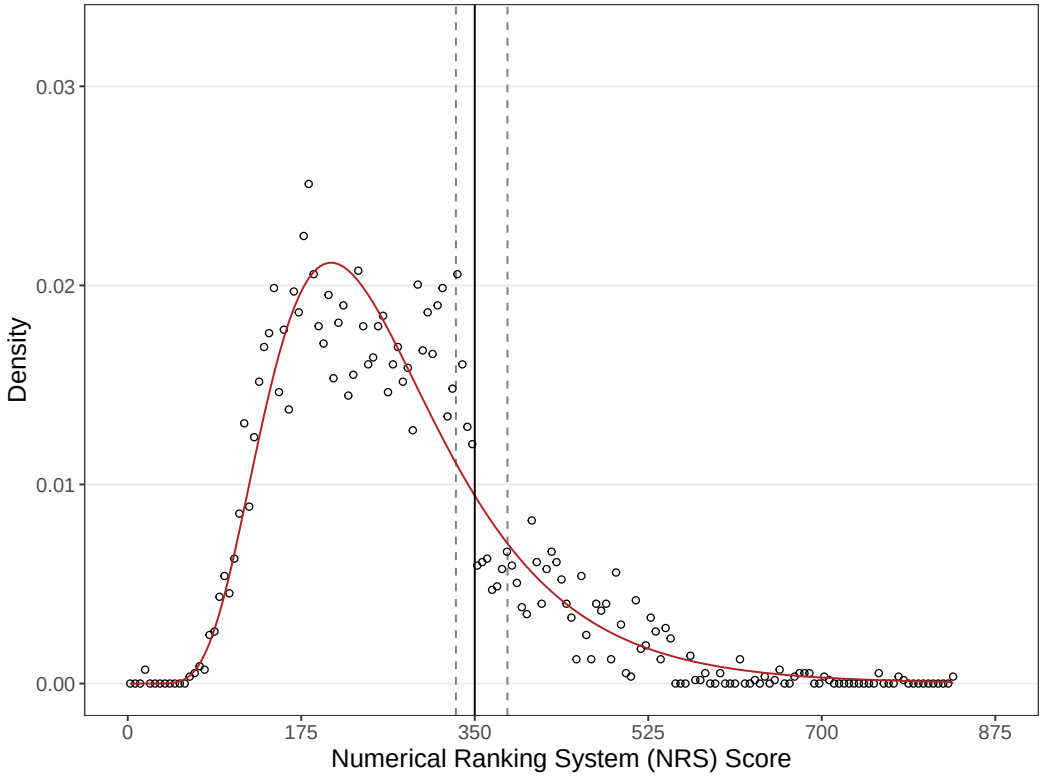
	(1)	(2)	(3)	(4)	(5)
Panel [A] Average household income ($000)					
I{Tier I}	4.841	4.58	4.656	3.048	2.754
	(1.471)	(1.361)	(1.379)	(1.170)	(1.902)
Bandwidth	66.4	65.1	58.4	69.1	50
Observations	2,918	2,898	2,605	3,101	2,184
Panel [B] Median home value ($000)					
I{Tier I}	17.77	17.22	26.78	20.89	23.91
	(7.105)	(7.520)	(6.893)	(5.408)	(11.172)
Bandwidth	73.4	65.3	58.5	72.9	50
Observations	3,184	2,847	2,561	3,129	2,153
Panel [C] White population share (%)					
I{Tier I}	5.968	5.291	2.951	0.7617	-0.3713
	(1.431)	(1.385)	(1.063)	(1.004)	(1.928)
Bandwidth	51	54.4	98.1	101.2	50
Observations	2,225	2,415	4,358	4,471	2,184
Panel [D] Adult pop. with any college (%)					
I{Tier I}	6.858	6.649	7.181	4.798	6.876
	(1.751)	(1.763)	(1.734)	(1.480)	(2.330)
Bandwidth	62.8	60.5	58.5	67.4	50
Observations	2,759	2,692	2,605	2,957	2,184
BW selection	Optimal	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes	Yes
Region FE	No	No	Yes	Yes	Yes
County FE	No	No	No	Yes	Yes

Notes: Each column presents results from a separate regression discontinuity estimation for how the outcome in each panel varies where crossing the Tier II to Tier I threshold at 350 total points in the Numerical Ranking System. All regressions use the "rdrobust" software package developed and provided by Calonico et al. (2014). Heteroskedasticity-robust bias-corrected standard errors are selected using the same package, as are optimal bandwidths using a triangular kernel. Where included, tier-assignment year FE are fixed effects for each year (1994-2013) of NRS site scoring, and region FE are fixed effects for each of the four MassDEP office regions.

Figure A4: Estimated manipulation regions and counterfactual densities for subsamples of sites in below and above median Census Tracts based on average household earned income



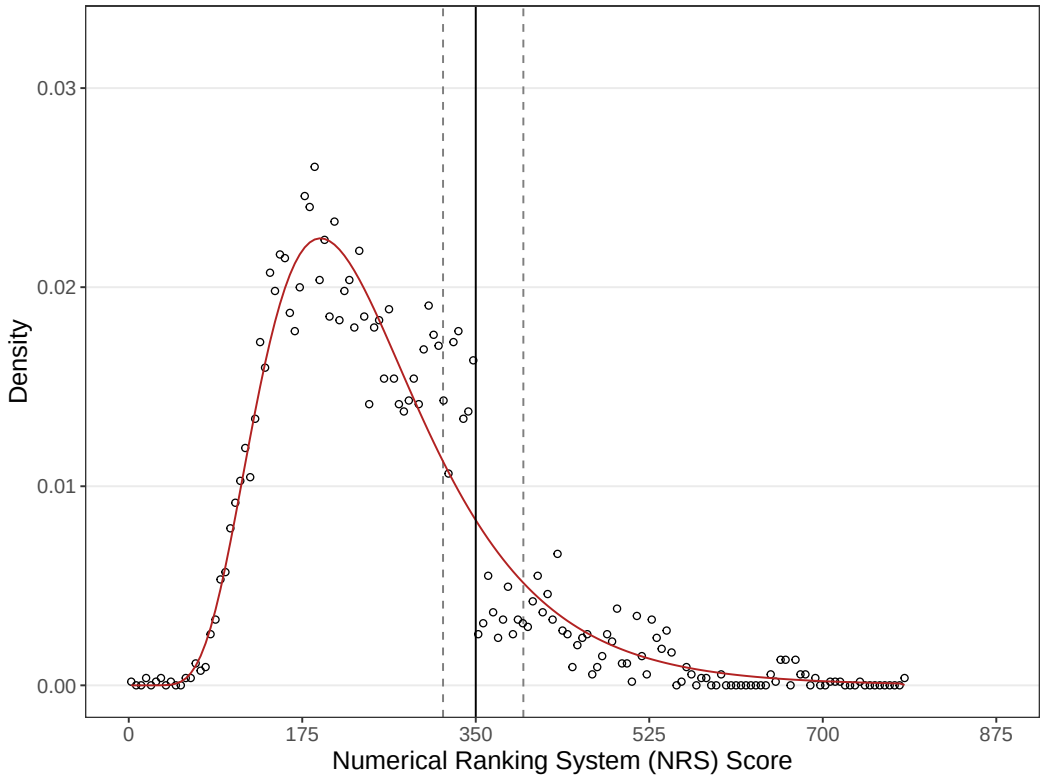
(a) Below median Census Tracts



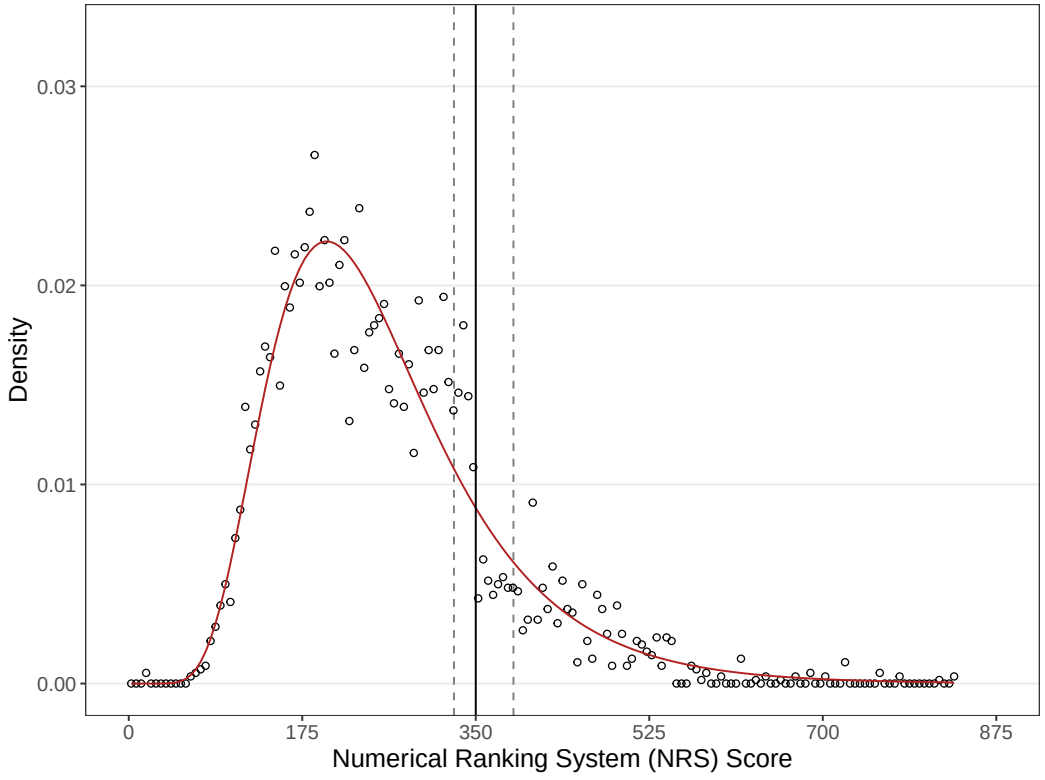
(b) Above median Census Tracts

Notes: The figures plot the score distributions for the indicated subsamples, the estimated regions over which score manipulation is present, and the counterfactual density functions.

Figure A5: Estimated manipulation regions and counterfactual densities for subsamples of sites in below and above median Census Tracts based on median home value



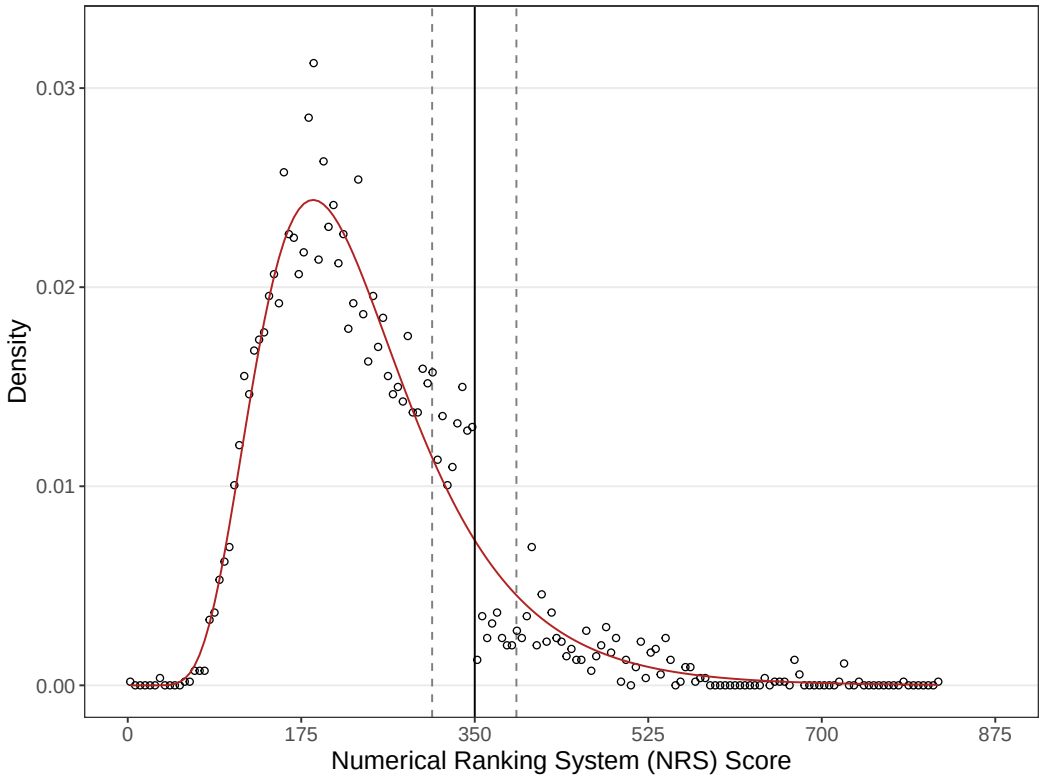
(a) Below median Census Tracts



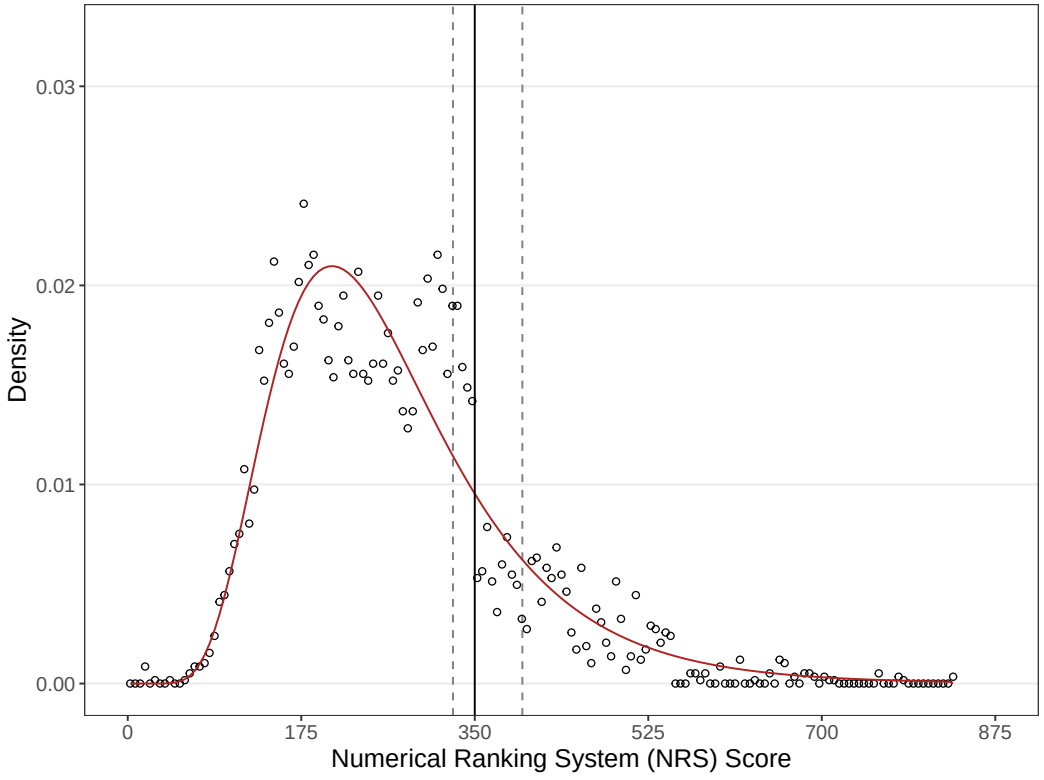
(b) Above median Census Tracts

Notes: The figures plot the score distributions for the indicated subsamples, the estimated regions over which score manipulation is present, and the counterfactual density functions.

Figure A6: Estimated manipulation regions and counterfactual densities for subsamples of sites in below and above median Census Tracts based on white population share



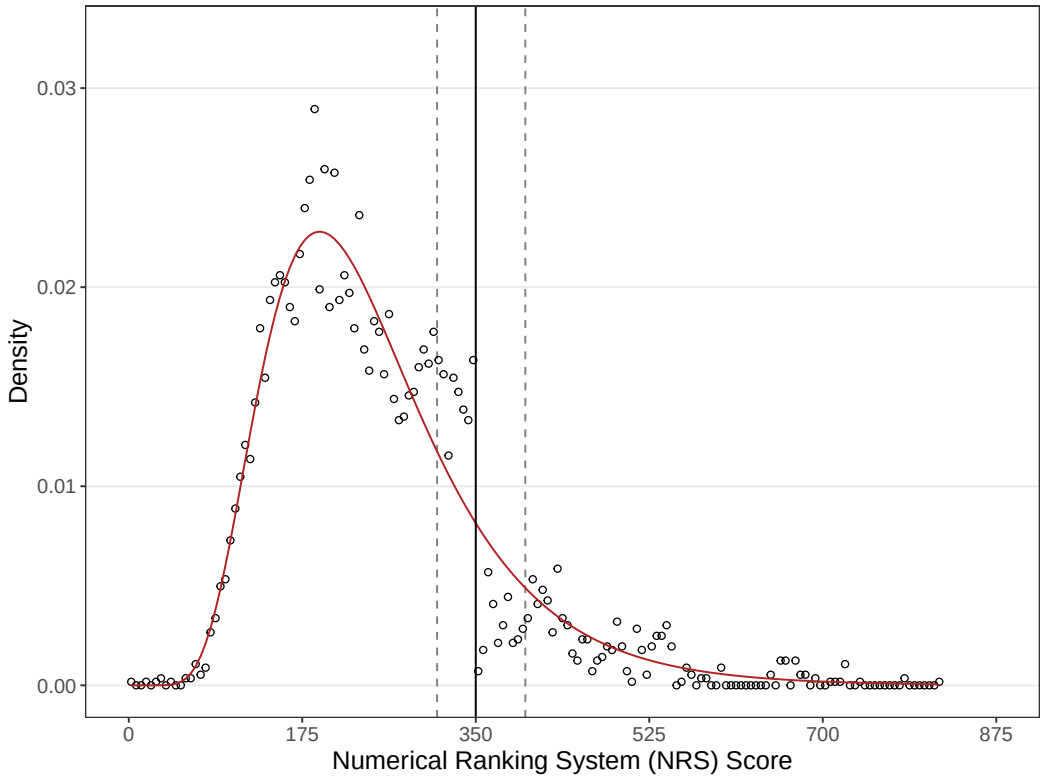
(a) Below median Census Tracts



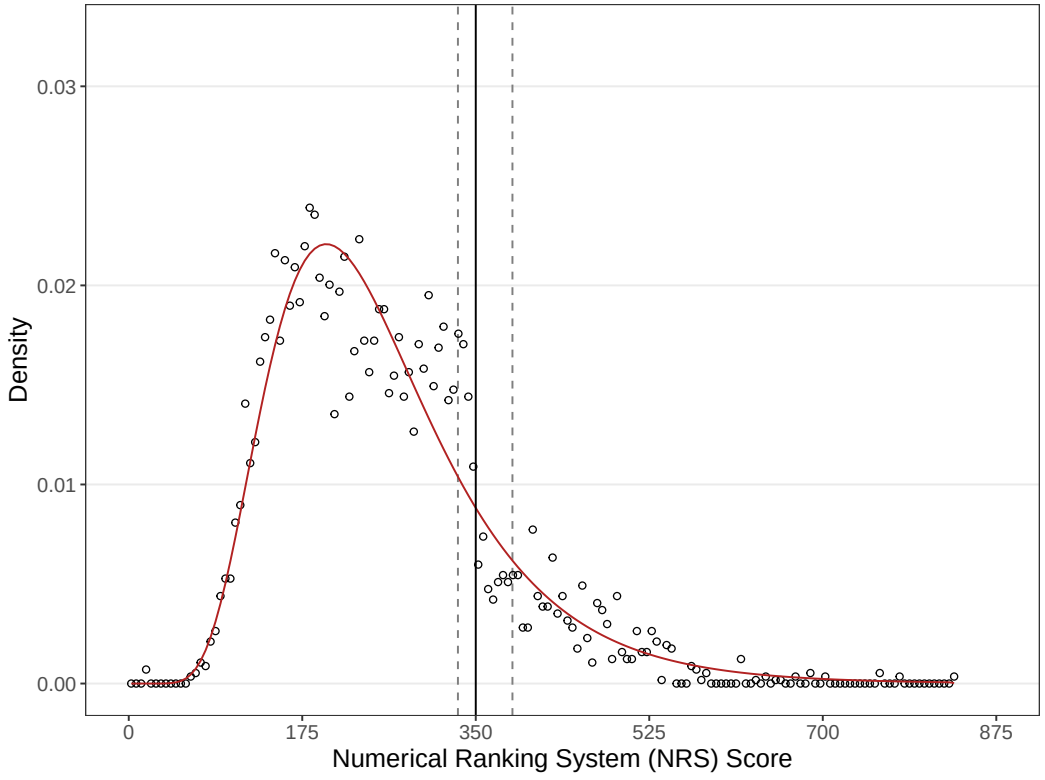
(b) Above median Census Tracts

Notes: The figures plot the score distributions for the indicated subsamples, the estimated regions over which score manipulation is present, and the counterfactual density functions.

Figure A7: Estimated manipulation regions and counterfactual densities for subsamples of sites in below and above median Census Tracts based on adult population with any college



(a) Below median Census Tracts



(b) Above median Census Tracts

Notes: The figures plot the score distributions for the indicated subsamples, the estimated regions over which score manipulation is present, and the counterfactual density functions.

Table A3: Relationship between the estimated excess density at each site's NRS score and neighborhood characteristics: Linear regression estimates

	Dep. variable: Excess density X 1000							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Household earned income	−0.408				−0.074	−0.076	−0.108	−0.433
	(0.080)				(0.134)	(0.135)	(0.138)	(0.167)
Median home value		−0.233			0.038	0.032	−0.655	−0.547
		(0.081)			(0.133)	(0.133)	(0.160)	(0.169)
White population share			−0.673		−0.611	−0.605	−0.324	−0.186
			(0.079)		(0.088)	(0.089)	(0.094)	(0.104)
Adult pop. with any college				−0.333	−0.136	−0.128	0.211	0.409
				(0.082)	(0.144)	(0.144)	(0.152)	(0.165)
Dep. variable mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Year fixed effects	No	No	No	No	No	Yes	Yes	Yes
Region fixed effects	No	No	No	No	No	No	Yes	Yes
County fixed effects	No	No	No	No	No	No	No	Yes
Observations	11,064	11,064	11,064	11,064	11,064	11,064	11,064	11,064

Notes: Each column presents results from a linear regression of the estimated excess density at a site's NRS score on the four 1990 Census Tract covariates, expressed as percentiles within the state. Heteroskedasticity-robust standard errors are in parentheses. Where included, the year fixed effects are for the year of tier assignment and region fixed effects are for each of the four MassDEP office regions.

B Methodological appendix

This appendix section provides additional details and discussion about the bunching estimator that we use to estimate the width of the manipulation region and the counterfactual density function. As described in Section 4.1 of the paper, our estimator is adapted from Diamond and Persson (2016), Kleven (2016), and Chen et al. (2021). The methodology uses k-fold cross-validation with a grid search over possible widths of the manipulation region. After the data-driven approach selects a manipulation region, we then use the full sample of data outside of the chosen manipulation region to estimate the counterfactual log-normal distribution. We compare the observed data to this estimated counterfactual to quantify manipulation, using a bootstrap procedure for inference.

To recover the unmanipulated distribution and the width of the manipulation region, we first collapse the data of Numerical Ranking System (NRS) scores into the density at each score value. The domain of empirical score values spans from three to 831 points. Let the density at score value s be defined as d_s, which we model as:

$$\underbrace{d_s}_{\substack{\text{Observed}\\ \text{density}}} = \underbrace{\Phi(\theta, s)}_{\substack{\text{Unmanipulated}\\ \text{distribution}}} + \underbrace{\sum_{j=\underline{s}}^{349} \gamma_j \cdot 1[s=j]}_{\substack{\text{Excess}\\ \text{density}}} - \underbrace{\sum_{j=350}^{\overline{s}} \gamma_j \cdot 1[s=j]}_{\substack{\text{Missing}\\ \text{density}}} + \underbrace{\epsilon_s}_{\substack{\text{Sampling}\\ \text{error}}} \quad (3)$$

The counterfactual density at each score value is obtained as the predicted value of Equation (3) omitting the contribution of the dummies in the excluded region $[\underline{s}, \overline{s}]$, i.e. $\hat{d}_s = \Phi(\hat{\theta}, s)$. Excess (missing) mass is then estimated as the difference between the observed and counterfactual density for each score value in $[\underline{s}, \overline{s}]$.[28] We specify that the unmanipulated distribution Φ is log-normal, parameterized by mean μ and standard deviation σ.[29]

We determine the manipulation region using k-fold (k=5) cross-validation with a grid search over all possible combinations of $\underline{s} \in [300, 345]$ and $\overline{s} \in [355, 400]$.[30] For each guess of the manipulation region, $[\underline{s}, \overline{s}]$, we use a constrained optimization by linear approximations

[28] As discussed in the main text, the NRS also has tier thresholds at 450 and 550 points, however, there is very little density in the right tail of the score distribution. Less than five percent of sites are scored above 450 points. Our estimator focuses only on the manipulation region around the 350 point Tier I/II threshold.

[29] Bunching estimators in the literature often specify the unmanipulated density function as a linear combination of polynomial basis functions. In our setting, this leads to much worse quality-of-fit and implausible estimated manipulation regions, compared to imposing log-normal structure for the density function.

[30] The grid search is computationally intensive and run-time scales exponentially with the size of the grid, so our primary procedure constrains the possible manipulation region to be within 50 points of the threshold. This also reduces the bias from any manipulated mass around the Tier IC/IB 450-point threshold. We confirmed that the estimated manipulated density is unchanged if we further expand the grid size.

(COBYLA) direct search algorithm to estimate the counterfactual density. Using only score values outside of $[\underline{s}, \overline{s}]$, the nonlinear optimization solves for the log-normal parameters μ and σ that minimize the sum of squared errors between the counterfactual density and the observed density:

$$\underset{\{\mu,\sigma\}}{\text{Min}} \sum_{s \notin [\underline{s},\overline{s}]} (d_s - \Phi(\mu, \sigma, s))^2 \tag{4}$$

This estimation is done using the 80 percent training sample.[31] We then calculate the out-of-sample mean squared error (MSE) for the 20 percent hold-out sample using the estimated $\hat{\mu}$ and $\hat{\sigma}$. This is done separately for each of the five folds, and then we average the MSE over the five folds for each combination of $[\underline{s}, \overline{s}]$. We then select the manipulation region, $[\underline{s}, \overline{s}]$, that yields the smallest out-of-sample MSE, conditional on passing a statistical test that total excess mass equals total missing mass. To operationalize this statistical test, we pool the five folds and calculate the residuals, $d_s - \Phi(\hat{\mu}, \hat{\sigma}, s)$, using the estimated $\hat{\mu}$ and $\hat{\sigma}$. Our test criteria is that the absolute value of the total prediction error in the manipulation region, $\sum_{\underline{s}}^{\overline{s}}(d_s - \Phi(\hat{\mu}, \hat{\sigma}, s))$, is smaller than the 10th percentile of the absolute value of these residuals. In practice, the selected manipulation region has a difference in total excess mass and total missing mass that is very close to zero (≈ 0.001).

Once we have determined the out-of-sample MSE-minimizing manipulation region, $[\underline{s}, \overline{s}]$, and counterfactual distribution, $\Phi(\hat{\mu}, \hat{\sigma})$, we compute the total missing density in $[350, \overline{s}]$ using Equation (3). Likewise, we compute the share of scores in $[350, \overline{s}]$ that are manipulated by dividing the total estimated missing density by the total counterfactual density in $[350, \overline{s}]$.

We bootstrap standard errors for the estimated missing density and share of scores manipulated by drawing score values with replacement from the original score distribution and collapsing to score density, and then re-estimating Equation (4) and calculating missing density using Equation (3) for each bootstrap sample, with 1000 repetitions.

[31] Following the literature, when randomly binning the data into five groups for cross-validation, we sample from the score density values, d_s, instead of sampling from the uncollapsed data on waste sites.