Dirty Business: Principal-Agent Problems in Hazardous Waste Remediation

Justin Marion[a]* Jeremy West[a,b]

[a]University of California at Santa Cruz
[b]The E2e Project

January 2022

Abstract

Governments often privatize the administration of regulations to third-party specialists paid by the regulated parties. We study the resulting conflict of interest for hazardous waste sites in Massachusetts, where the responsible parties must hire private firms to quantify environmental contamination. We find significant bunching of site severity scores just below thresholds that determine the intensity of government oversight throughout the remediation. We show this client favoritism in evaluations is enabled by discretion afforded to evaluators. Favoritism is associated with inferior remediation quality and is most pronounced in lower socioeconomic status neighborhoods, 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 Spencer Banzhaf, James Bushnell, Peter Christensen, Kenneth Gillingham, Joshua Graff Zivin, Wayne Gray, Alex Hollingsworth, Koichiro Ito, David Keiser, Ashley Langer, Matthew Neidell, 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 Spring Conference, the 2021 European Urban Economics Conference, the 2021 International Industrial Organization Conference, UC-Santa Cruz, and the UC-Environmental Economics Workshop. 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. The authors have no relevant or material financial interests related to the research described in this paper.

1 Introduction

Government regulations are often officially enforced 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 to government agencies. It leverages the expertise of firms and their cost-containment motive, and it shifts 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 result from this conflict of interest have been established in several empirical studies, for instance in Jin and Leslie (2003). Less well studied are the distributional consequences, 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 efficiency and distributional 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.

We begin by presenting a model of the incentives for misreporting site severity assessments that demonstrates a novel channel through which inequities in pollution exposure can arise from regulatory enforcement. In our model, remediation raises the value of the property where the contamination occurred, providing a private benefit to the responsible party for site cleanup in addition to the positive externality for the broader neighborhood. When the property value gain is small, the polluter wants to put inefficiently low effort into site cleanup and will demand an inaccurate assessment of site severity to obtain less government oversight. Misreporting by the assessor will therefore be more common for sites in poor neighborhoods with a low willingness-to-pay for environmental amenities, and cleanup outcomes will be worse. This 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 racial minorities 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 racial minority population share.

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 this discontinuous regulatory process in several ways.

By examining the the distribution of NRS scores, we find substantial bunching just below the tier thresholds. We view this bunching as compelling evidence 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 Tier II (less severe) and Tier I (more severe), due to its location within the NRS distribution.[2] 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 40 percent.

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

[1] Responsible parties can share this surplus with LSPs either explicitly or implicitly via repeated business.

[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.

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, or one-third 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 estimating how predetermined socioeconomic characteristics of site Census Tracts vary across tier thresholds. We find that income, property values, education, and white population share all increase discontinuously at the Tier I/II threshold. This implies that sites located in neighborhoods with lower socioeconomic status are more likely to be manipulated to fall under the Tier I threshold, and therefore the adverse impacts of NRS manipulation are concentrated among disadvantaged populations. This result is consistent with the prediction of our theoretical model that score manipulation is less likely in neighborhoods with greater willingness-to-pay (or ability-to-pay) for environmental amenities such as high-quality site remediation.

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.,

[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.

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 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. In addition to better-aligning economic incentives for honesty, recent work supports that increased oversight also improves the behavior of government agents (Borcan et al., 2017; West, 2018; Calvo et al., 2019). Our findings speak to a third mechanism: we show that discretion by third-party evaluators can exacerbate incentives for misbehavior.

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., 2019). Prior work on hazard sites demonstrates 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. As described in a detailed review of this literature by Banzhaf et al. (2019), differential exposure can arise due to the initial siting of pollution, from household sorting by willingness-to-pay for environmental amenities, or by disparities in the enforcement of regulation. Our study provides new evidence for the last of these channels, 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.

Finally, our results relate to the literature on willingness-to-pay for environmental ameni-

[4]See Shimshack (2014) for a broader discussion of the literature on environmental compliance monitoring.

ties. 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. We show that polluters seek lighter regulation of waste remediation and reduce remediation quality in lower socioeconomic status neighborhoods, which is expected 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 responsible party client, 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 and reports to the government. The severity score that the agent reports determines the site's regulatory treatment. The agent uncovers the true severity score, z^*. However, the government does not directly observe z^*, so the agent may choose to report a score $z < z^*$ to obtain more favorable regulatory treatment. Misreporting is costly and the agent's loss function associated with score manipulation is given by $\phi(z, z^*)$, with $\phi(z^*, z^*) = 0$ under truth-telling, and ϕ increasing as $z^* - z$ grows. This cost can represent loss of credibility, legal penalties, or a disutility of dishonesty. Depending on whether the reported score is above or below a threshold z_0, there are two regulation categories: Tier I sites are more severe and face r^I if $z \geq z_0$ whereas Tier II sites are less severe and face r^{II} if $z < z_0$.

Upon learning the site severity, the responsible party decides how much effort, $e \in \{0, 1\}$, to devote to site cleanup. The cost of remediation, $c(e, x, z^*)$, is influenced by effort, the true z^*, and other site characteristics, x. Effort also reduces pollution levels and is capitalized into

the property value of the site: $v(e, x, z^*)$. By comparing these private costs and benefits, the responsible party selects a preferred effort level, $\tilde{e}$. If the returns to effort outweigh the cost, i.e. $\Delta v \geq \Delta c$, then high effort is chosen and $\tilde{e}$=1, where $\Delta v \equiv v(e = 1, x, z^*) - v(e = 0, x, z^*)$ and $\Delta c \equiv c(e = 1, x, z^*) - c(e = 0, x, z^*)$.

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 $z^* \geq z_0$, the responsible party's desired effort is $\tilde{e} = 0$ 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 below z_0 increases the client's surplus by $w(x, z^*) = \Delta c - \Delta v$. Suppose the agent receives a share of this surplus, $\lambda \in (0, 1]$.[5] The agent will misreport the site severity if the agent's surplus exceeds the misreporting cost:

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

Let $\overline{z}(x)$ be the largest z^* for which the relationship in Equation (1) holds. This misreporting threshold depends on site characteristics x, which determine both Δv and Δc.[6]

The model yields three testable predictions.

Prediction 1. For sites with $z_0 \leq z^* \leq \overline{z}(x)$, the agent will report z just below z_0 and the score distribution will therefore have excess mass just below z_0.

This first prediction pertains to the extent of misreporting of severity scores. For sites with a true severity score not too far above the threshold, the cost of misreporting may be low enough to be outweighed by the gains to the agent, depending on site-specific factors (x). The empirically-testable hypothesis is that the distribution of observed scores should exhibit excess mass just under the tier threshold and missing mass above the threshold.

Prediction 2. Let e^- and e^+ denote the value of e approaching z_0 from below and above, respectively. As z approaches z_0, e^- is discontinuously lower than e^+.

[5] We do not model how surplus is shared between the responsible party and the agent, which could be done either explicitly or via implicit contract enforced by repeated interactions. In practice, the agent is often hired to also conduct the remediation, but under reasonable assumptions – a competitive remediation market and low switching costs – this feature will not influence the agent's scoring decision.

[6] Although Equation (1) may not hold for any z^* for certain values of x, the premise of our study is that there are at least some sites with a x for which $\lambda w(x, \overline{z}(x)) > \phi(z_0, \overline{z}(x))$. For these nontrivial cases, under the mild assumption that the cost of manipulation increases with z^* more steeply than the benefit, $\partial w/\partial z^* < \partial \phi/\partial z^*$, then there is a single crossing of $\lambda w(x, z^*)$ and $\phi(z_0, z^*)$ and thus a unique $\overline{z}(x)$. This assumption holds as long as it is more difficult for the agent to credibly manipulate scores that are farther from the tier threshold, which is likely in practice given that the government reviews submitted scoresheets.

This second prediction is a test of the incentives for score manipulation. The regulatory intensity changes discretely at the Tier I threshold, and therefore so will the chosen effort. Specifically, there should be lower remediation effort for sites barely below the tier threshold compared to those barely above the threshold. By extension, we should observe comparatively lower quality cleanups of sites just below the Tier I threshold, which is an empirically-testable hypothesis. Note that, because regulatory intensity is weaker for Tier II sites, the cleanup effort and quality might be discontinuously worse for these sites even absent score manipulation. Prediction 2 and the associated empirical hypothesis are a test that score manipulation facilitates reduced effort.

Prediction 3. Let x^- and x^+ denote the value of x approaching z_0 from below and above, respectively. The signs of $(x^+ - x^-)$, $\partial \Delta v / \partial x$, and $-\partial \Delta c / \partial x$ are equal.

This third and final prediction of the model relates to the characteristics of the sites that are manipulated. Site characteristics affect score manipulation through second-order effects on the net return to cleanup effort. The misreporting threshold $\overline{z}(x)$ decreases with attributes that are positively related to Δv or negatively related to Δc. Sites with higher levels of these attributes should be less likely to be misreported and, as a result, these attributes should be discontinuously larger on average just above the Tier I regulatory threshold.

This prediction relates closely to the literature that evaluates manipulation of the running variable for regression discontinuity designs. As described by Lee (2008), in the absence of manipulation, predetermined characteristics should be smooth across categorical thresholds. The socioeconomic status of the neighborhood, x^{SES}, is determined prior to the decision to misreport the score. Without manipulation, the expected value of x^{SES} is the same at (barely) each side of the threshold. If scores are manipulated, then selection generates discontinuities in the expected value of x^{SES} at the threshold, with the sign of this discontinuity depending on the relationship between x^{SES} and the terms Δv and Δc. Likewise, Δc and Δv varies discontinuously at the threshold, though these objects are unobserved. In practice, any variable that influences these objects, including unobserved site severity for instance, will vary discontinuously at the threshold due to selective manipulation.

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.[7] 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 evaluation components (along with a site information component) which are summed to form an overall score ranging from 18 to 1320 points. Four 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. Appendix Figure A1 shows the empirical contribution of each of these components to the total NRS score. Additionally, there is a component allowing for ad hoc adjustments of ± 0-50 points

[7]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.

for "mitigating site-specific conditions," determined at the discretion of the LSP.[8]

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.[9] Most scored sites (84.5 percent) fall into this tier. Sites scored 350 or above are more serious and obtain a classification of Tier I. This tier is further subdivided into Tier IC (350-449, 10.64 percent of sites), Tier IB (450-549, 3.15 percent of sites), and Tier IA ($\geq$ 550, 1.7 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).[10]

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.[11]

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

[8]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.

[9]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.

[10]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.

[11]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.

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.

3.2 Data

This study relies on data from several sources. Our starting point is a database provided by MassDEP that contains the universe of hazardous contamination sites in Massachusetts for spills that occurred during 1984 to present.[12] The database includes details on each site location, the chemical(s) that were spilled, and the history of official actions taken throughout the remediation process (e.g. 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 and LSP identifiers 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. Where noted, we also use data from the 2010 Census and American Community Survey.

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, so consequently only 15.5 percent of sites are Tier I (A, B, or C). 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. 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).

[12]The hazard site database is available in flat file format from the Massachusetts Department of Environmental Protection at https://www.mass.gov/service-details/downloadable-contaminated-site-lists.

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 12.55 percent nonwhite; 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 the 57th, 61st, 72nd, and 52nd percentiles across all Tracts statewide (unweighted).

4 Results

The model and predictions derived in Section 2 guide our empirical work. We examine discontinuities at the NRS Tier I threshold in the distribution of scores, 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 higher racial minority population shares, lower adult educational attainment, lower household incomes, and lower home values.

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.[13] Because of this scoring discreteness, manipulation can lead to excess mass in the score distribution even inframarginal to the tier cutoff.[14]

[13]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.

[14]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.

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 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 A2, 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 on the share of sites categorized Tier I, we conduct a back-of-the envelope calculation comparing the empirical distribution in Figure 1 to a parametric log-normal distribution fit to the mean (250) and standard deviation (104) of the data. As the figure shows, this log-normal distribution closely fits the data for scores that are far from tier cutoffs. However, there is noticeably more mass in the empirical distribution for scores between 290 and 349. We calculate that there are 615 "excess" sites in this range than would be the case if the empirical distribution followed the fitted log-normal, which is sizable – more than forty percent – when compared against the 1478 Tier I sites actually observed.

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.[15] Figure 2(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.

[15]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.

Unsurprisingly, Figure 2(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. For scores away from tier thresholds, this relationship holds in Figure 2(b). 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 2, 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.[16] These three columns use optimal bandwidths calculated using the methods of Calonico et al. (2014), while Column (4) 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 higher just above the threshold compared to just below (se = 1.46). 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 -0.268 (se = 0.031) at the tier threshold, and again the estimate and significance change little across specifications.

The evidence shown in Figure 2 and Table 2 provide a clear indication that the excess bunching of the site score distribution is intentional and that Component VI is a substantial factor. Setting this component to zero and holding the other components constant would increase the share of Tier I sites by 12.86 percent, and therefore this component alone explains almost one-third of the total excess bunching.

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 Prediction 2 of 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

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

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 3 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 3. As described above, all RD estimates use kernel-based local linear regression. Most specifications use the optimal bandwidth for that specification, while Column (4) uses a constant bandwidth of 50 points across all outcomes. Panel [A] shows estimates for the discontinuity in the likelihood of sites' permanent solutions entailing "no significant risk." Consistent with the figure, we find that barely-Tier I sites are 32.8 percent more likely to achieve this highest quality of permanent solution (se = 8.3 percent). This finding is robust to the inclusion of year and region fixed effects. When a fixed bandwidth of 50 is used in Column (4), the RD estimate increases somewhat, to 38.6 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 and region fixed effects, and to specifying a 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 by examining 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 racial/ethnic minority (nonwhite) population share, and the share of the adult (25 or older) population with any college.

Results for these four neighborhood characteristics are presented visually in Figures 4 and 5, 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, lower minority 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. Panel [A] shows that the discontinuity in average annual household earned income is \$4,852 (se = \$1,479). The estimate changes little with the inclusion of year and region fixed effects and is actually somewhat higher using the common bandwidth of 50. This gap in local income is both large and economically significant, about fourteen percent of the sample mean. A similar pattern is shown for home values in Panel [B]. We find a discontinuity of \$18,104 (se = \$6,948), which changes little with the inclusion of year fixed effects and increases somewhat when including region fixed effects or switching to the

common bandwidth of 50 points. Again, the difference is economically significant, more than ten percent of the sample mean.

The latter two panels of Table 4 also show large and significant discontinuities in Census characteristics at the Tier I/II threshold. The nonwhite population share in Panel [C] declines by 5.97 percentage points at the tier threshold (se = 1.48), a magnitude that is nearly 50 percent of the sample mean. This point estimate is also unaffected by the inclusion of year effects, but is somewhat attenuated to -2.65 (se = 1.02) when conditioning on region effects. This attenuation is perhaps not that surprising, given that two of the four MassDEP regions have fairly little racial variation. Panel [D] shows that the college share rises by an estimated 6.89 percentage points (se = 1.74) at the threshold. This estimate barely changes with the inclusion of year effects or the further inclusion of region effects, and it grows to 11.48 percentage points when using the common bandwidth of 50 points in Column (4). As with the other three Census outcomes, these estimated discontinuities are large and economically significant, at least 14 percent of the sample mean.

Ultimately, these Census attributes capture spatial variation, and the four SES measures we consider are correlated with one another. A discontinuity in one measure might simply be due to a scoring choice based on another neighborhood characteristic. As an attempt to evaluate each SES attribute's marginal contribution to score manipulation, we consider sites within 50 points of the Tier I/II threshold. This "manipulation region" is both the range in which we predominantly find excess 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 SES terms predict the likelihood that the site was scored above the Tier I threshold using the following regression:

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

where i, j, and t index site, MassDEP region, and year of tier assignment. So that the coefficients of interest in the vector B will be comparable in magnitude, we convert each SES measure into the Tract's percentile across all tracts in the state, unweighted and scaled to range 0 to 1. These percentile SES measures are captured in the vector $\tilde{X}_i^{SES}$. The specification also includes region fixed effects, ρ_j, and year fixed effects, γ_t.

Table 5 presents results of this estimation. In Column (1), we include only the four SES measures as regressors. Columns (2) and (3) respectively add year and region fixed effects. Across each specification, we find that the racial/ethnic minority share has a negative and statistically significant coefficient while the education measure has a positive and significant

relationship. For each ten percentile increase in the Tract's minority population share, the likelihood that a site is scored above 350 is reduced by 1.4 percentage points (se = 0.37). Similarly, each ten percentile increase in the college population share raises the likelihood of a site score being above 350 by 1.59 percentage points (se = 0.54). These estimates are statistically significant, robust across the specifications, and quite sizable, especially given that only 21 percent of the sites in the "manipulation region" are Tier I.

At face value, these results show that NRS score manipulation is less likely in neighborhoods that have higher educational attainment and a smaller minority population share, even conditional on local income and property values.[17] 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.[18] 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.

Altogether, the relationships between Census attributes and site score manipulation indicate that the principal-agent problem we document above has a much more pronounced impact on socioeconomically disadvantaged neighborhoods. Each measure varies discontinuously at the threshold in a common direction. For two sites with an equal true severity above the tier threshold, it is the site in the lower SES neighborhood that is more likely to be manipulated into Tier II status. Coupled with the results shown in Section 4.2 of inferior cleanup quality for barely-Tier II sites, the implication is that low SES 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

[17] 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.

[18] 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.

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.[19]

We utilize this reform to provide additional evidence supporting the disparate impact of score manipulation on low SES neighborhoods. In Section 4.3, we showed that education, race, income and housing values 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 6, 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 grows 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 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. To provide more unequivocal evidence, 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. Our evaluation uses difference-in-differences specifications of the form:

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

[19]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."

The dependent variables are the SES measures examined earlier – average household income, median housing values, nonwhite 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 6 presents these estimates for 2010 Census Tract-level attributes, using a sample period spanning 2010-2019. Prior to the reform, the α_1 coefficients for the four measures all indicate generally similar SES differences as those shown above for the local averages near the tier threshold. Turning to the difference-in-differences coefficients of interest, all four coefficients indicate some reversal of the pre-reform disparities. Moreover, three of the four characteristics show that gaps are statistically eliminated after the reform. The only exception is the share of nonwhite residents, for which the Tier I-II gap is large pre-reform (10.44 percentage points lower in Tier I sites) and does not substantially decline (moves 1.8 percentage points closer). 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

As the complexity of the economy and the scope of government responsibilities continue to grow, public policymakers increasingly turn to the private sector to assist with the administration of regulations. 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 racial minority share.

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 socioeconomic status.

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

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.

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.

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.

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.

G. Z. Jin and P. Leslie. The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, 118(2):409–451, 2003.

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

D. S. Lee. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2):675–697, 2008.

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.

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 (forthcoming)*, 2019.

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.

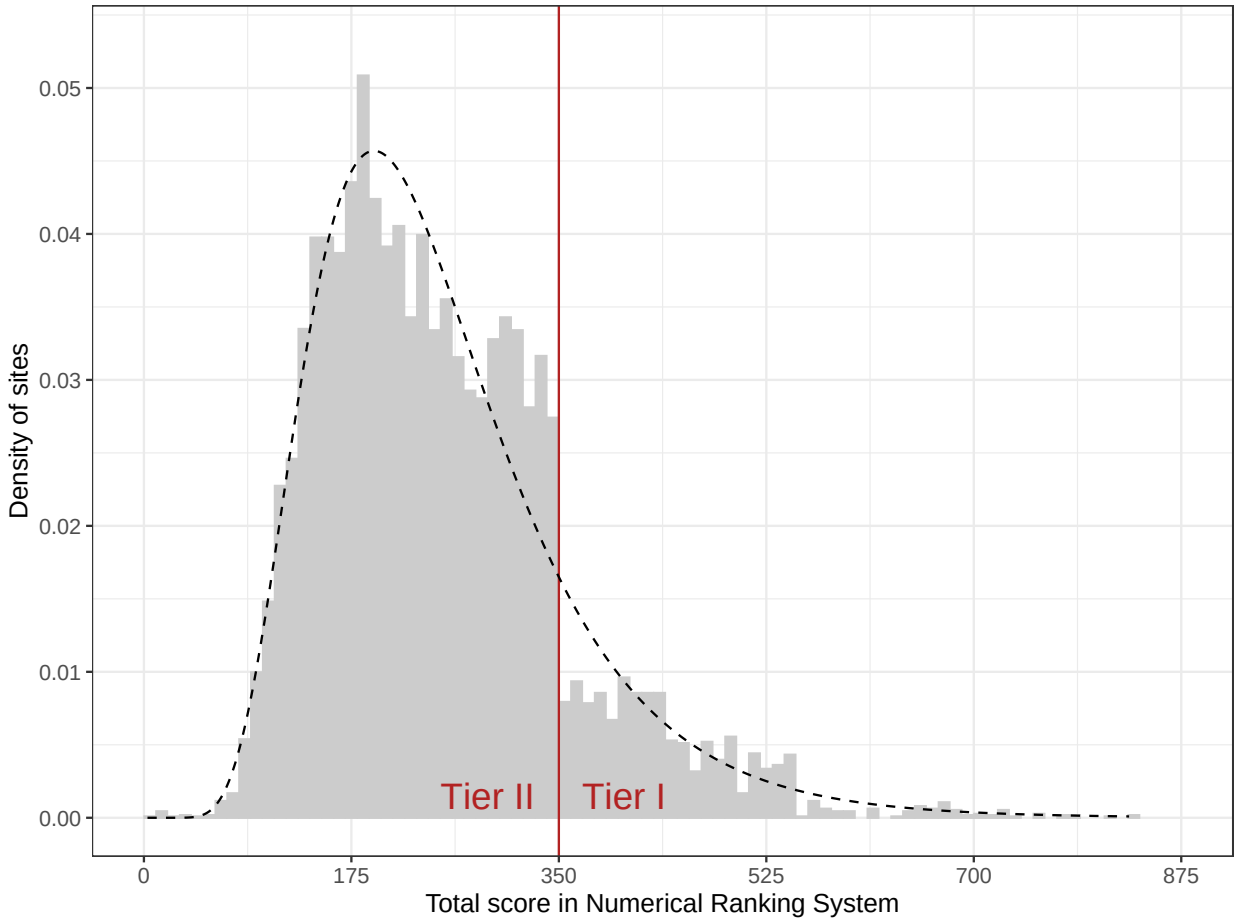
W. K. Viscusi and J. T. Hamilton. Are risk regulators rational? Evidence from hazardous waste cleanup decisions. *The American Economic Review*, 89(4):1010–1027, 1999.

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.

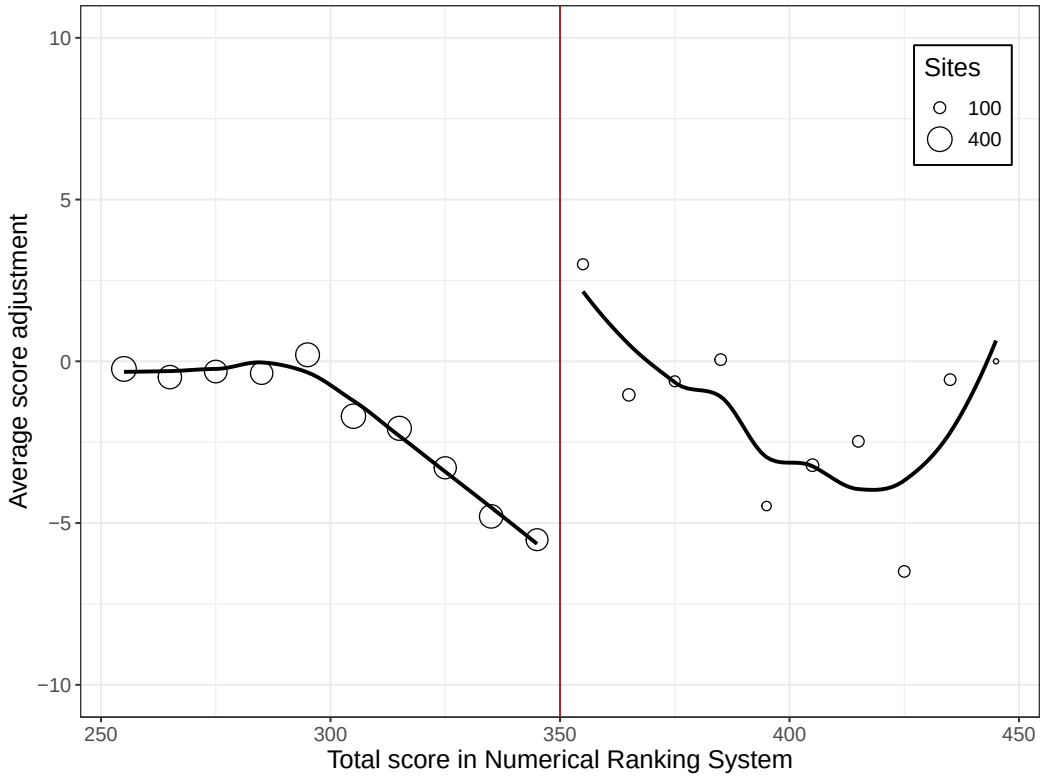
Figures and tables

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

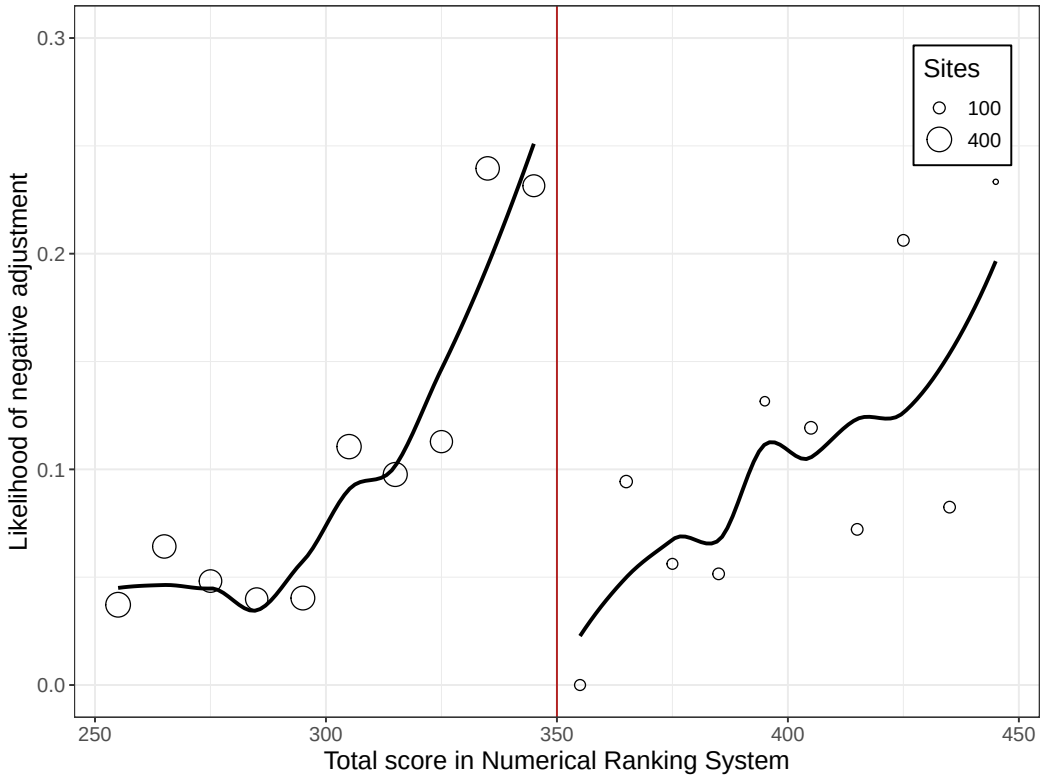


Notes: Figure 1 plots the distribution of hazardous waste site scores in the Numerical Ranking System using a bin width of 10 points and showing the full set of 11,347 scores. The dashed black line shows a log-normal distribution fit to the mean (250) and standard deviation (104) of the set of scores. The solid vertical red line indicates the cutoff at 350 points between the Tier II and Tier I regulatory categories.

Figure 2: Score adjustments for "mitigating disposal site-specific conditions"



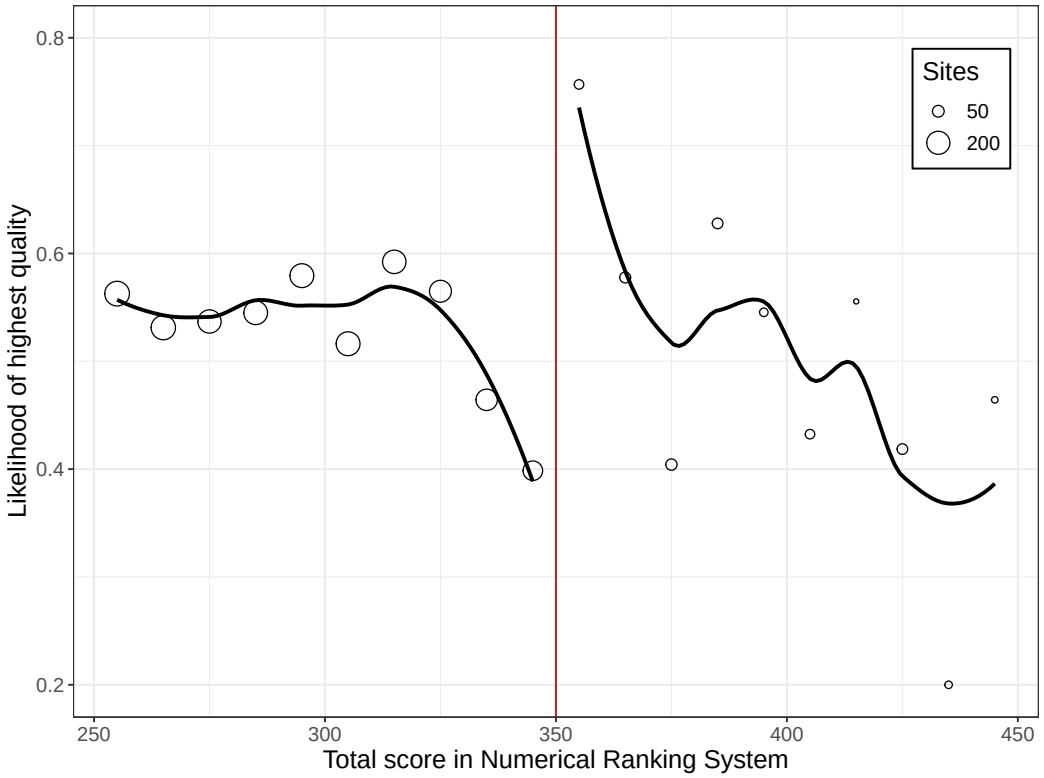
(a) Average NRS component VI score adjustment



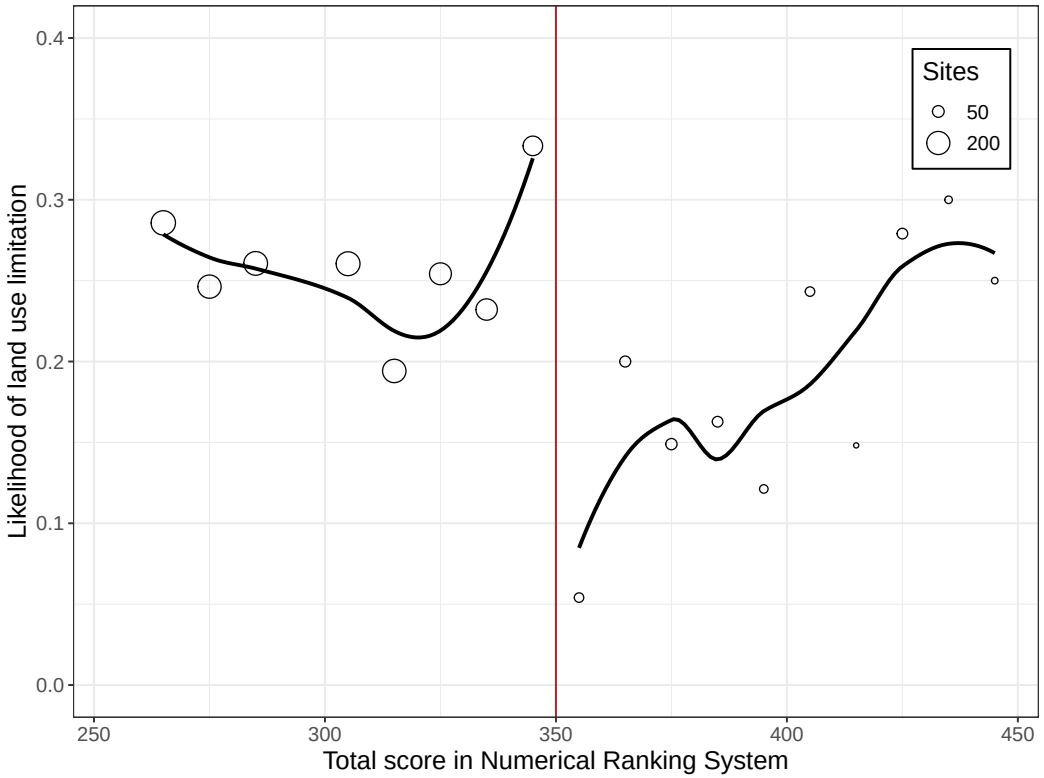
(b) Likelihood of a negative NRS component VI score adjustment

Notes: Figure 2 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.

Figure 3: Measures of cleanup quality for sites with a Permanent Solution



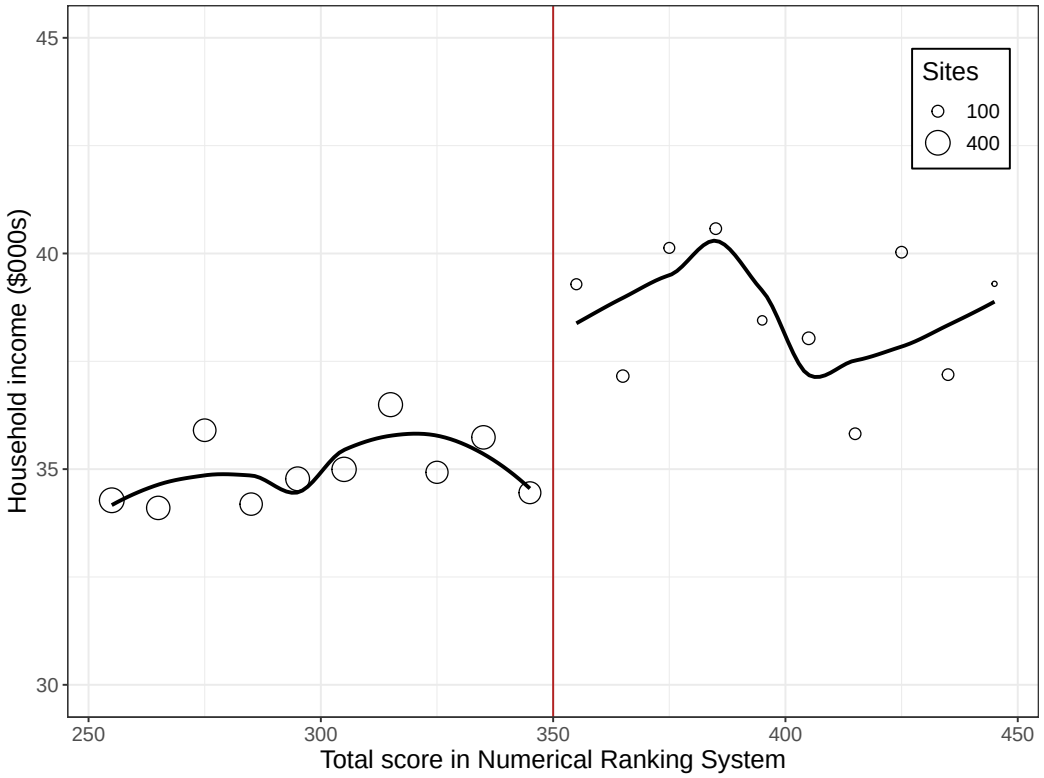
(a) Permanent Solution of A1 or A2: "No Significant Risk" (highest quality)



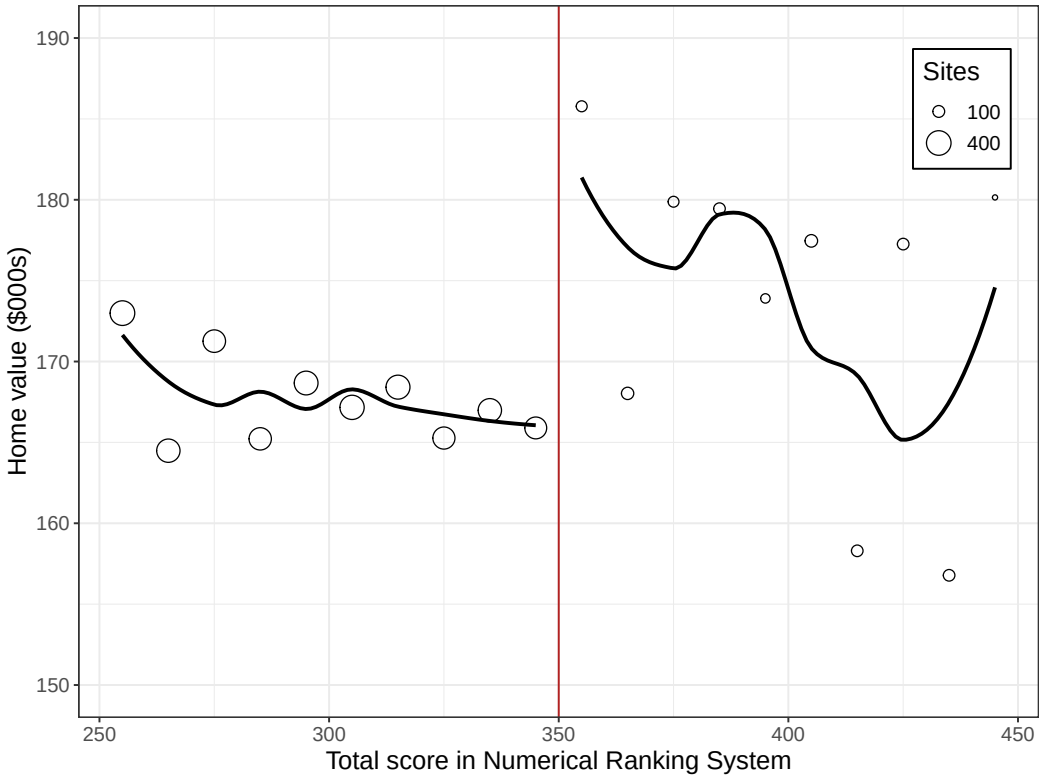
(b) Permanent Solution involves an Activity and Use Limitation for the property

Notes: Figure 3 plots local averages for measures of cleanup quality for sites with a Response Action Outcome Permanent Solution 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 4: Predetermined economic characteristics for neighborhood of site



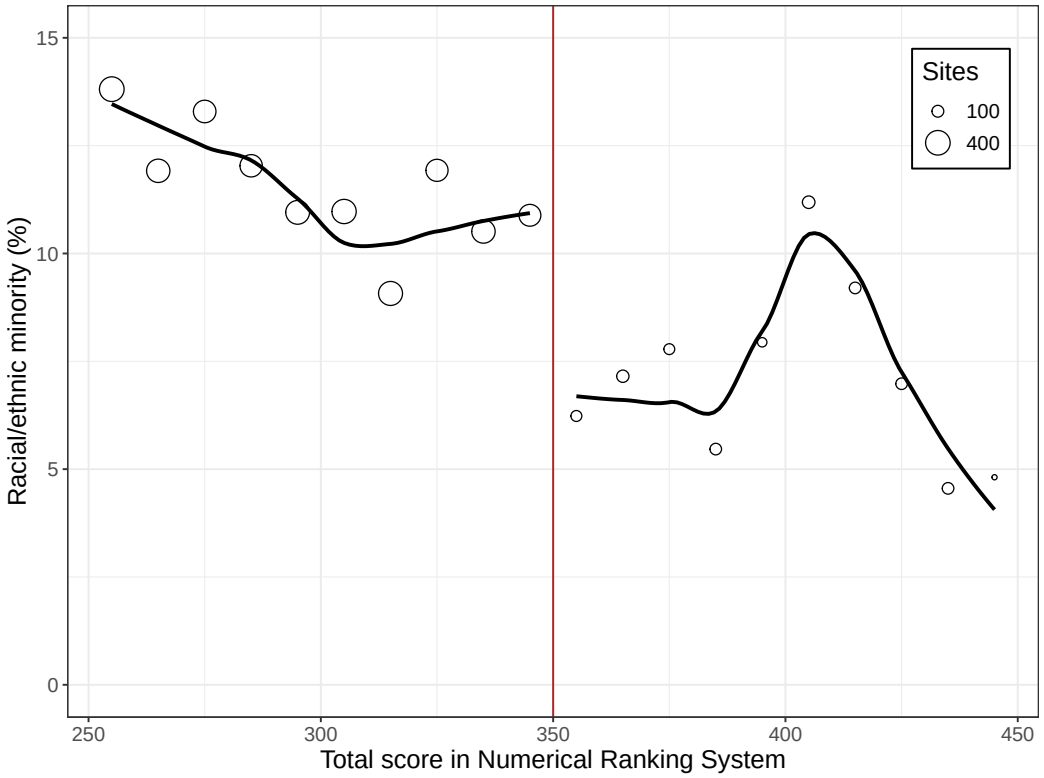
(a) 1990 Census Tract-level average household earned income ($000s)



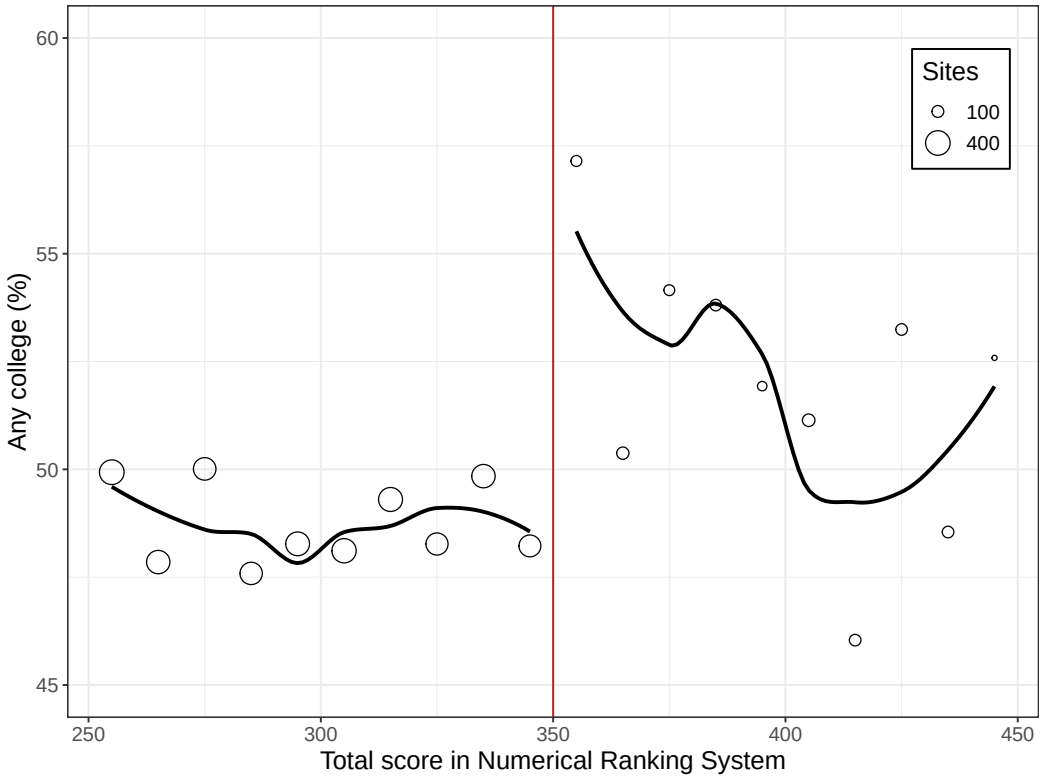
(b) 1990 Census Tract-level median home property value ($000s)

Notes: Figure 4 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 5: Predetermined demographic and education characteristics of neighborhood



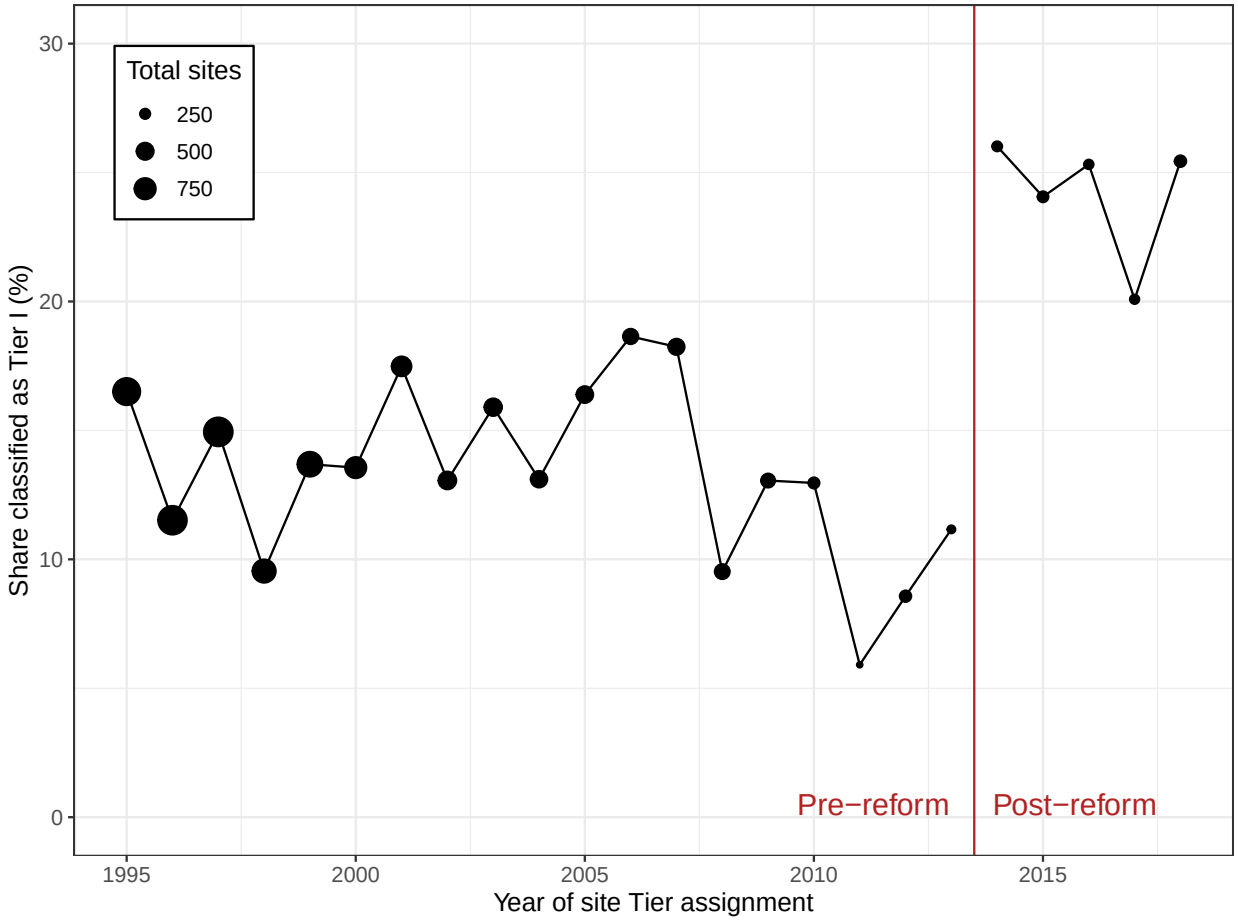
(a) 1990 Census Tract-level racial/ethnic minority share of residents (nonwhite)



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

Notes: Figure 5 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.

Figure 6: Tier composition of newly-classified sites by year during 1995-2018



Notes: Figure 6 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 red 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 1: Summary statistics on sites in the Numerical Ranking System

	Mean	St. Dev.
Panel [A] Site scoring and cleanup quality		
Tier I	0.155	0.362
NRS total score	250.092	103.620
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	0.218	0.413
Panel [B] Predetermined Census Tract covariates		
Household earned income ($000)	34.501	13.143
Median home value ($000)	167.862	64.944
Non-white population (%)	12.545	18.612
Adult pop. with any college (%)	48.709	16.903
Number of sites	11,347	

Notes: Table 1 presents summary statistics 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 defined as persons over age 25.

Table 2: NRS site scoring: Regression discontinuity estimates

	(1)	(2)	(3)	(4)
Panel [A] NRS component VI score				
I{Tier I}	8.915***	8.725***	8.373***	8.489***
	(1.458)	(1.471)	(1.447)	(1.223)
Bandwidth	46.2	45.8	47.7	50
Observations	1,996	1,982	2,063	2,190
Panel [B] Has negative NRS component VI score				
I{Tier I}	-0.268***	-0.283***	-0.271***	-0.283***
	(0.031)	(0.031)	(0.030)	(0.026)
Bandwidth	42.4	40.3	43.1	50
Observations	1,826	1,753	1,907	2,190
BW selection	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes
Region FE	No	No	Yes	Yes

*p<0.1; **p<0.05; ***p<0.01 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 3: Site remediation quality: Regression discontinuity estimates

	(1)	(2)	(3)	(4)
Panel [A] Highest quality: "No Significant Risk"				
I{Tier I}	0.328***	0.286***	0.281***	0.386***
	(0.083)	(0.080)	(0.080)	(0.074)
Bandwidth	54.4	53.8	53.4	50
Observations	1,216	1,187	1,187	1,095
Panel [B] Has land use limitation (AUL)				
I{Tier I}	-0.202***	-0.186***	-0.175***	-0.23***
	(0.061)	(0.063)	(0.061)	(0.059)
Bandwidth	58.8	58.5	62.7	50
Observations	1,306	1,306	1,390	1,095
BW selection	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes
Region FE	No	No	Yes	Yes

*p<0.1; **p<0.05; ***p<0.01 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 4: Predetermined neighborhood characteristics: Regression discontinuity estimates

	(1)	(2)	(3)	(4)
Panel [A] Average household earned income ($000)				
I{Tier I}	4.852***	4.568***	4.625***	7.78***
	(1.479)	(1.357)	(1.435)	(1.352)
Bandwidth	65.6	65.1	55.6	50
Observations	2,898	2,898	2,435	2,184
Panel [B] Median home value ($000)				
I{Tier I}	18.1***	17.24**	26.83***	40.46***
	(6.948)	(7.475)	(6.888)	(7.050)
Bandwidth	76.6	66.5	58.9	50
Observations	3,275	2,867	2,561	2,153
Panel [C] Racial/ethnic minority share (%)				
I{Tier I}	-5.974***	-5.178***	-2.654***	-3.032**
	(1.477)	(1.471)	(1.024)	(1.259)
Bandwidth	48.1	48	107.8	50
Observations	2,122	2,058	4,748	2,184
Panel [D] Adult population with any college (%)				
I{Tier I}	6.885***	6.653***	7.187***	11.48***
	(1.743)	(1.755)	(1.738)	(1.603)
Bandwidth	62.7	60.7	57.5	50
Observations	2,759	2,692	2,514	2,184
BW selection	Optimal	Optimal	Optimal	Fixed
Year FE	No	Yes	Yes	Yes
Region FE	No	No	Yes	Yes

*p<0.1; **p<0.05; ***p<0.01 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: Estimating whether NRS score is above 350 by neighborhood characteristics

	Dep. variable: Score between 350-400		
	(1)	(2)	(3)
Household earned income	0.003	−0.010	−0.039
	(0.047)	(0.047)	(0.049)
Median home value	−0.007	0.004	0.081
	(0.048)	(0.048)	(0.056)
Racial/ethnic minority share	−0.140***	−0.124***	−0.109***
	(0.037)	(0.036)	(0.039)
Adult pop. with any college	0.159***	0.161***	0.130**
	(0.054)	(0.054)	(0.056)
Year fixed effects	No	Yes	Yes
Region fixed effects	No	No	Yes
Observations	2,209	2,209	2,209

*p<0.1; **p<0.05; ***p<0.01 Notes: Each column presents results from a 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, 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 6: Tier reform and neighborhood characteristics: Difference in differences estimates

	Income	Home val.	Nonwhite	College
	(1)	(2)	(3)	(4)
I{Tier I}	15.947***	26.952*	−10.441***	4.856***
	(3.372)	(13.949)	(2.420)	(1.790)
I{Tier I} X I{Post-reform}	−9.767**	−35.210**	1.804	−5.511***
	(3.980)	(16.464)	(2.857)	(2.113)
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

*p<0.1; **p<0.05; ***p<0.01 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. The outcome in Column (1) is average household earned income in thousands of dollars. In Column (2) it is the median home value in thousands of dollars. The Column (3) outcome is the percentage of racial/ethnic minority persons as a share of Tract population. In Column (4) the outcome is the percentage share of adult (25 or older) population with any college attainment. Heteroskedasticity-robust standard errors are in parentheses. The tier-assignment year FE are fixed effects for each year (2010-2019).

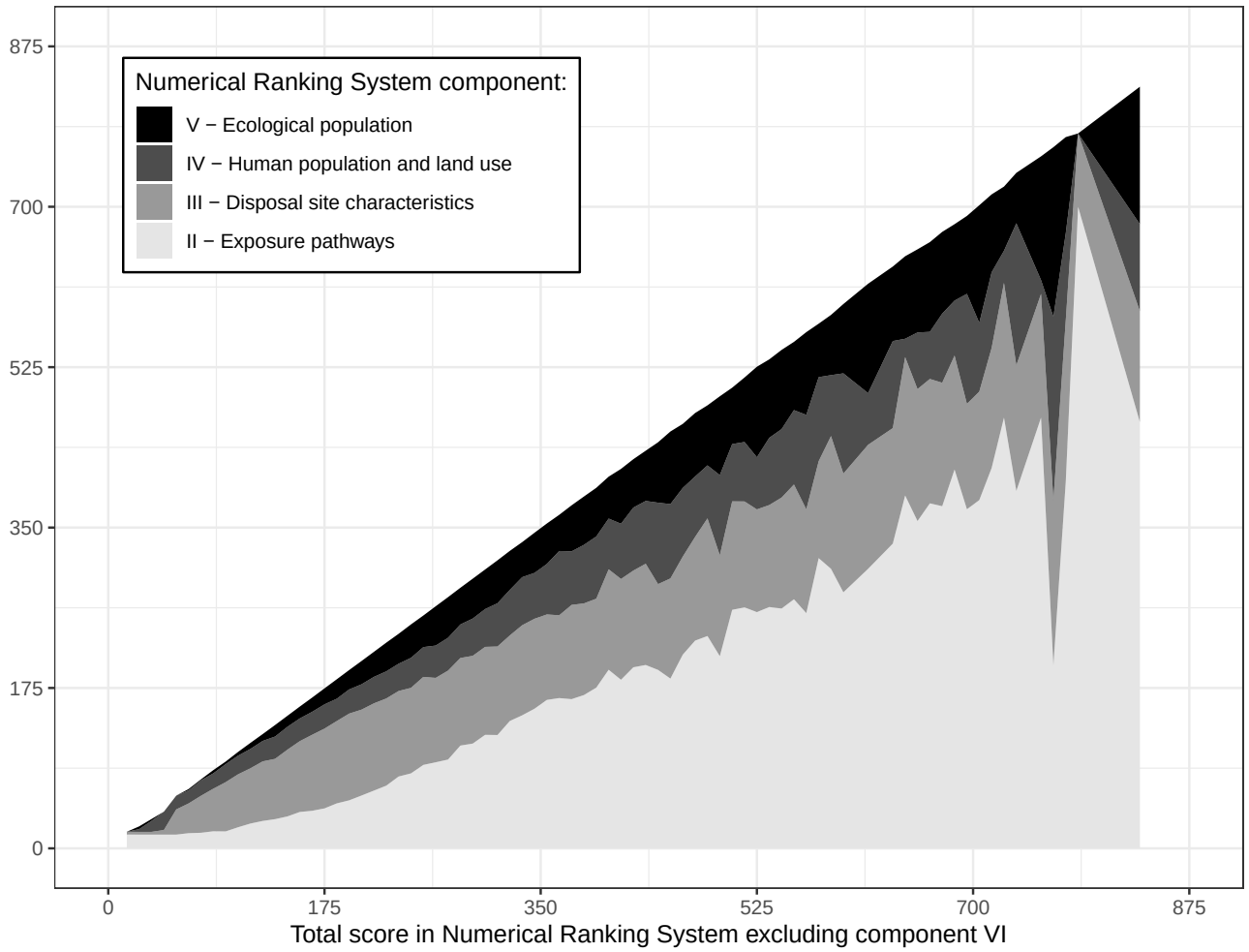
A Appendix tables and figures

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

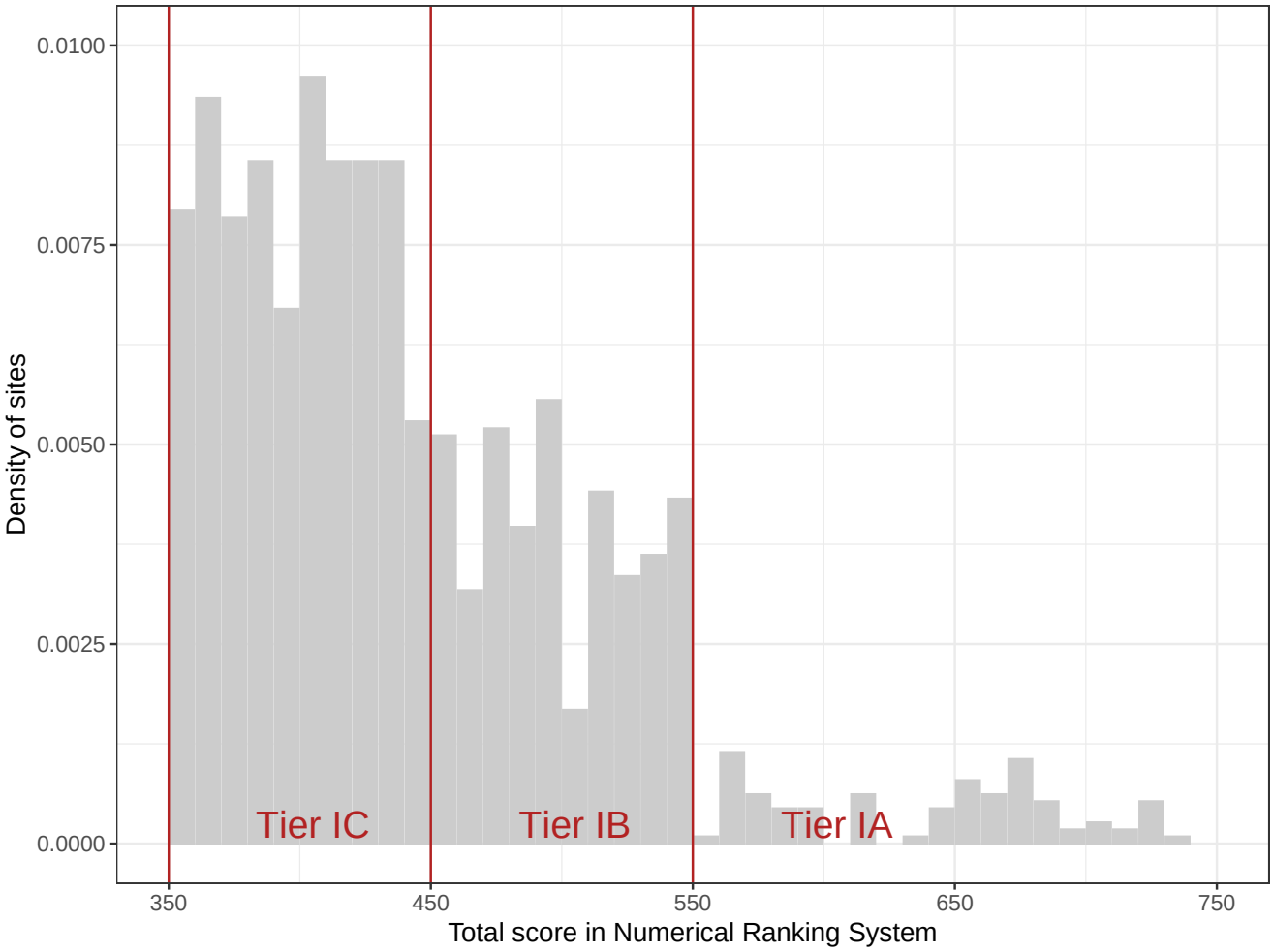
Notes: Values in this table are sourced from the Numerical Ranking System Guidance Manual (310 CMR 40.1500). This manual of more than 80 pages is "written to assist users of the Numerical Ranking System developed by the Massachusetts Department of Environmental Protection to classify disposal sites as defined by the Massachusetts Contingency Plan and Massachusetts General Law." At least one of the exposure pathways must be assigned a value of at least 15 points, as to do otherwise would indicate that no spill occurred.

Figure A1: Component contributions to total scores of sites in the Numerical Ranking System



Notes: Appendix Figure A1 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, as depicted in Figure 1, there are very few sites with scores at the far right tail of the distribution.

Figure A2: Distribution of site scores in the NRS zoomed-in to 350-750



Notes: Figure A2 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 red 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.