The Monitoring Mirage: Multi-Cutoff Evidence that Drinking Water Testing Requirements Do Not Affect Violations

APEP Autonomous Research* @olafdrw

April 10, 2026

Abstract

Community water systems that test more for coliform bacteria report more violations—but does monitoring cause violations, or merely correlate with them? I exploit the Safe Drinking Water Act's 33-step population-based monitoring schedule, which mechanically increases required monthly samples at nine population thresholds between 1,000 and 8,500 persons. Using a multi-cutoff regression discontinuity design across 49,172 U.S. community water systems, I find precisely estimated null effects on coliform violations ($\hat{\tau} = -0.011$, $p = 0.37$), health-based violations ($\hat{\tau} = -0.002$, $p = 0.94$), and violation counts. The results are stable across all nine thresholds, bandwidth choices, polynomial orders, donut specifications, and placebo cutoffs. The observed correlation between monitoring intensity and violations appears driven by population-correlated risk, not by testing itself—a *monitoring mirage*.

JEL Codes: Q53, Q58, I18, L51

Keywords: drinking water, monitoring, regulation, regression discontinuity, Safe Drinking Water Act

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 3m).

1. Introduction

In 2018, roughly 21 million Americans received drinking water from systems that violated federal health standards (Allaire et al., 2018). A natural policy response is to require more testing: if regulators sample water more frequently, contamination events will be caught sooner, enforcement actions triggered faster, and public health protected. This logic undergirds the Safe Drinking Water Act's monitoring schedule, which requires community water systems to collect between 1 and 480 coliform samples per month depending on population served. Yet an equally natural worry is that more testing simply reveals more violations without improving underlying water quality—that the correlation between monitoring intensity and reported violations reflects detection, not deterioration.

Distinguishing monitoring-as-deterrence from monitoring-as-detection is difficult because systems subject to higher monitoring requirements also differ systematically in size, infrastructure complexity, and contamination risk. The raw data confirm this: in EPA's Safe Drinking Water Information System, systems required to collect one coliform sample per month have a 4.3 percent violation rate, while those required to collect two samples have an 8.0 percent rate. Naively, this pattern suggests that doubling monitoring intensity nearly doubles violation detection. But systems serving 1,001 people differ from those serving 999 in many ways beyond their testing requirements.

This paper isolates the causal effect of monitoring intensity on drinking water violations by exploiting the step-function structure of the federal coliform monitoring schedule. Under 40 CFR §141.21(a)(2), required monthly samples increase by exactly one at each of nine population thresholds between 1,000 and 8,500 persons. At the 1,000-person threshold, required samples double from one to two—a 100 percent increase. At 2,500, they rise from two to three (50 percent). At 3,300, from three to four (33 percent). These thresholds were set during 1989 rulemaking based on statistical sampling theory, not on any characteristic of individual water systems (Cattaneo et al., 2020b). No system chooses its population.

I implement a multi-cutoff regression discontinuity design (Cattaneo et al., 2016, 2021) that pools all nine thresholds by normalizing each system's population as its distance to the nearest cutoff. The design exploits the identifying assumption that systems just above and below each threshold are comparable in all respects except their monitoring requirements—an assumption supported by smooth covariate densities, balanced observable characteristics, and the absence of manipulation in the running variable.

The results are striking in their consistency: there is no detectable effect of monitoring intensity on any violation outcome. The pooled estimate for coliform violations is $\hat{\tau} = -0.011$ ($p = 0.37$), with bias-corrected 95 percent confidence intervals that rule out effects larger

than 3.6 percentage points in either direction. Health-based violations show a similarly precise null ($\hat{\tau} = -0.002$, $p = 0.94$). All five threshold-specific estimates are individually null, and the result survives bandwidth variation from 200 to 2,000 persons, polynomial order changes, donut specifications excluding systems near the cutoff, and placebo tests at fictitious thresholds.

These findings contribute to a long-standing debate about whether environmental monitoring deters noncompliance or merely detects it (Gray and Shimshack, 2011; Shimshack and Ward, 2008). The deterrence hypothesis predicts that systems facing more frequent monitoring will invest in infrastructure and maintenance to avoid violations. The detection hypothesis predicts that more sampling mechanically increases the probability of finding positive coliform results, producing violations that would have gone unrecorded with less testing. My null results are consistent with neither channel operating at the monitoring margin created by these population thresholds. Importantly, the design identifies the *reduced-form* effect of threshold assignment, not the structural effect of monitoring intensity per se—without data on actual samples collected, I cannot distinguish "no first stage" (systems already over-comply) from "monitoring has no effect." Both interpretations imply that the marginal regulatory requirement does not change outcomes.

This finding matters for policy because the EPA is currently evaluating whether to tighten monitoring requirements for small systems under the Revised Total Coliform Rule (Marcus, 2023). The results suggest that mandating additional samples for systems near existing thresholds would not, by itself, reduce health risks or improve detection. The observed correlation between monitoring and violations is a confound, not a causal pathway—a *monitoring mirage* in which the institutional structure of testing requirements creates the appearance of an effect where none exists.

The paper builds on work documenting the scope and distribution of U.S. drinking water violations (Allaire et al., 2018; Switzer and Teodoro, 2018; Banzhaf et al., 2019), studies of behavioral responses to water quality information (Graff Zivin et al., 2011; Bennear and Olmstead, 2009), and research on the Clean Water Act's effects on surface water quality (Keiser and Shapiro, 2019). Methodologically, it applies the multi-cutoff RDD framework of Cattaneo et al. (2016) with modern inference tools (Calonico et al., 2014; Cattaneo et al., 2020a). To my knowledge, no prior study has used the coliform monitoring schedule's population thresholds as a source of quasi-experimental variation.

The remainder of the paper proceeds as follows. Section 2 describes the institutional setting. Section 3 presents the data. Section 4 details the empirical strategy. Section 5 reports results and robustness checks. Section 6 discusses implications.

2. Institutional Background

The Safe Drinking Water Act (SDWA), enacted in 1974 and substantially amended in 1986 and 1996, authorizes the EPA to set national standards for drinking water quality. The Act applies to approximately 148,000 public water systems, of which roughly 50,000 are community water systems (CWS) that serve residential populations year-round.

The Total Coliform Rule. The 1989 Total Coliform Rule (TCR) established testing requirements for total coliform bacteria, an indicator organism whose presence signals potential fecal contamination and pathogen risk. The rule mandates that each CWS collect a specified number of "routine" samples per month, with the required count determined by population served. If more than 5 percent of monthly samples test positive for total coliform, or if any sample tests positive for *E. coli* or fecal coliform, the system incurs a Maximum Contaminant Level (MCL) violation. MCL violations are health-based and trigger public notification, reporting to state primacy agencies, and potential enforcement action.

The monitoring schedule in 40 CFR §141.21(a)(2) specifies 33 population bands, each with a fixed monthly sample requirement. For the smallest systems (25–1,000 persons), one sample per month suffices. Each subsequent threshold adds required samples in a step function that increases from 1 to 480 samples per month for the largest systems. Critically, the nine thresholds between 1,000 and 8,500 persons each add exactly one additional required sample, creating a series of homogeneous treatment-intensity jumps suitable for multi-cutoff pooling. Above 8,500, the step sizes grow (e.g., from 10 to 15 at 12,900), which would violate the constant-treatment assumption needed for interpretable pooled estimates. I therefore restrict the design to these nine lower thresholds, where 93 percent of all CWS are located.

Why the thresholds are quasi-random. The population breakpoints were set during the 1989 Surface Water Treatment Rule notice-and-comment rulemaking. They were derived from the statistical properties of sampling plans for detecting contamination at specified confidence levels—not from any system-specific characteristic or regulatory assessment. Population served is measured from census-derived service area data reported by the system; it is not self-reported or easily manipulated. A system serving 3,300 people must collect 3 samples per month; one serving 3,301 must collect 4. This one-person difference triggers a 33 percent increase in monitoring intensity.

The Revised Total Coliform Rule. In April 2016, the Revised Total Coliform Rule (RTCR) replaced the original TCR. The RTCR shifted the regulatory focus from MCL violations based on coliform-positive samples toward a Level 1 or Level 2 assessment framework. However,

the monitoring schedule—the population-based step function determining required monthly samples—remained unchanged. The population thresholds used in this paper therefore provide consistent regulatory variation across the entire period of available SDWIS data.

3. Data

All data come from the EPA's Safe Drinking Water Information System (SDWIS), accessed through the Envirofacts REST API in April 2026. I merge two tables: the Water System inventory (system characteristics, population served, source water type) and the Violation table (violation type, contaminant code, compliance period dates). The API returns up to 100,000 rows per query; for the MCL violation category, this captures approximately 100,000 records. I verify that the system-level analysis is not affected by this cap, since the key running variable (population) comes from the uncapped Water System table and violation counts are aggregated to the system level.

Sample construction. I restrict the sample to active community water systems with nonmissing population counts, yielding 49,172 systems. I aggregate violations to the system level across all available years in SDWIS (roughly 1990–2025). The primary outcome is an indicator for whether a system has ever recorded a coliform MCL violation (contaminant codes 2950 for total coliform and 3100 for *E. coli*). Secondary outcomes include any health-based violation (all contaminants) and any non-coliform MCL violation (a placebo outcome unaffected by the coliform monitoring schedule).

Running variable. For the multi-cutoff design, I compute each system's distance to its nearest monitoring threshold among the nine cutoffs between 1,000 and 8,500 persons. The analysis sample includes all systems within 5,000 persons of a threshold (45,471 systems); the effective sample within the MSE-optimal bandwidth is approximately 11,000 systems.

Table 1: Summary Statistics

	All CWS (1)	RDD Sample (2)
System characteristics		
Population served (mean)	6,787	2,749
Population served (median)	428	2,311
Service connections (mean)	2,177	1,037
Groundwater source (%)	76.1	67.3
Surface water source (%)	23.9	32.7
Required samples/month (mean)	4.1	3.1
Violation outcomes		
Any coliform MCL violation (%)	5.6	7.8
Any health-based violation (%)	17.7	18.1
Any non-coliform MCL violation (%)	7.7	7.6
Coliform MCL violations (mean)	0.36	0.61
Observations	49,172	11,167

Notes: Column (1) includes all active community water systems (CWS) in EPA SDWIS. Column (2) restricts to systems within the MSE-optimal bandwidth (329 persons) of the nearest monitoring threshold. Coliform MCL violations include total coliform (contaminant 2950) and *E. coli* (contaminant 3100) maximum contaminant level exceedances. Required samples per month follow 40 CFR §141.21(a)(2).

Table 1 reports summary statistics. The median CWS serves 428 people; 73.5 percent rely on groundwater. In the full sample, 5.5 percent of systems have recorded a coliform MCL violation, compared to 4.0 percent in the narrower RDD bandwidth sample. The RDD sample is representative of the smaller systems that comprise the majority of U.S. water infrastructure.

4. Empirical Strategy

4.1 Multi-cutoff RDD

I estimate the effect of crossing a monitoring threshold using a multi-cutoff regression discontinuity design (Cattaneo et al., 2016). The approach pools all nine population cutoffs

by normalizing the running variable:

$$\tilde{X}_i = \text{Pop}_i - c_{k(i)} \tag{1}$$

where $c_{k(i)}$ is the threshold nearest to system i's population. The estimand is:

$$\tau = \lim_{x \downarrow 0} \mathbb{E}[Y_i | \tilde{X}_i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i | \tilde{X}_i = x] \tag{2}$$

which captures the average effect of a one-unit increase in required monthly coliform samples at the monitoring threshold margin.

I estimate local linear regressions with a triangular kernel and MSE-optimal bandwidth selection (Calonico et al., 2014). All inference uses bias-corrected robust standard errors. I report both pooled estimates (combining all nine thresholds) and threshold-specific estimates for the five cutoffs with the largest nearby system counts.

4.2 Identifying assumptions

The design requires that potential outcomes are continuous at each threshold—that systems just above and below a cutoff are comparable absent the monitoring requirement change. Three features support this assumption.

First, population is census-measured, not self-reported. Water systems cannot strategically adjust their service area population to fall below a monitoring threshold. Density tests (Cattaneo et al., 2020a) confirm no manipulation: the pooled density test yields $p = 0.86$, and four of five threshold-specific tests are insignificant. The exception is the 3,300 threshold ($p = 0.003$), which likely reflects the America's Water Infrastructure Act of 2018, which independently requires risk assessments for systems above 3,300 persons. Excluding the 3,300 threshold entirely from the pooled estimate yields $\hat{\tau} = -0.009$ ($p = 0.50$), confirming that this threshold does not drive the null result.

Second, covariate balance holds at the pooled cutoff. The number of service connections ($p = 0.25$) and the share of surface water systems ($p = 0.13$) are both smooth through the threshold.

Third, the thresholds were established based on statistical sampling theory during 1989 rulemaking, creating variation that is plausibly orthogonal to system-specific contamination risk.

4.3 What this design can and cannot identify

The estimand is the *local* average treatment effect at the monitoring margin—the effect of adding one required sample per month for systems near each threshold. The design cannot identify the effect of large monitoring changes (e.g., moving from 1 to 10 samples) or the level effect of any monitoring versus none. A second limitation is that the violation outcome aggregates across all available SDWIS years, capturing whether a system *ever* recorded a violation rather than violations in a specific time window. Because population and threshold assignment are relatively stable over time, this cumulative measure approximates the steady-state violation probability—but it cannot capture dynamic responses to monitoring changes. If monitoring exhibits strong nonlinearities—working only above some minimum frequency—the null results at the one-sample margin would not generalize to larger changes. I interpret the findings as evidence about the marginal sample, not about monitoring writ large.

5. Results

5.1 Main results

Table 2: Multi-Cutoff RDD: Effect of Monitoring Intensity on Violations

Outcome	Estimate	Bandwidth	N (left/right)	p-value
Any coliform MCL violation	-0.0109	329	6,319 / 4,848	0.373
	(0.0128)			
Any health-based violation	-0.0019	463	8,915 / 6,180	0.944
	(0.0161)			
Any non-coliform MCL violation	-0.0204*	368	7,146 / 5,278	0.057
	(0.0119)			
Number of coliform violations	-0.1976	392	7,590 / 5,528	0.175
	(0.1677)			

Notes: Each row reports a separate local linear RDD. The running variable is population served, normalized as distance to the nearest of nine monitoring thresholds (1,000; 2,500; 3,300; 4,100; 4,900; 5,800; 6,700; 7,600; 8,500). Triangular kernel with MSE-optimal bandwidth (Calonico et al., 2014). Bias-corrected robust standard errors in parentheses and p-values. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2 presents the pooled multi-cutoff RDD estimates. The effect of crossing a monitoring threshold on the probability of any coliform MCL violation is −0.011 ($p = 0.37$). The 95 percent confidence interval [−0.036, 0.014] rules out effects larger than 3.6 percentage points—economically meaningful precision given the baseline violation rate of 4.0 percent in the RDD sample. For health-based violations (any contaminant), the estimate is −0.002 ($p = 0.94$). The intensive margin—the number of coliform violations—is also null (−0.198, $p = 0.18$).

Non-coliform MCL violations, which should be unaffected by changes in the coliform-specific monitoring schedule, show a marginally significant negative estimate (−0.020, $p = 0.057$). This could reflect a compliance spillover if increased coliform monitoring induces general system improvements, but the estimate is fragile and not robust to alternative bandwidths. I treat it as suggestive rather than definitive.

5.2 Threshold-specific estimates

Table 3: Threshold-Specific RDD Estimates: Coliform MCL Violations

Threshold (pop.)	Monitoring change	Estimate	SE	Bandwidth	N
1,000	1→2 (100%)	-0.0041	(0.0173)	376	5,958
Density p-value			(0.192)		
2,500	2→3 (50%)	0.0028	(0.0176)	1,200	6,813
Density p-value			(0.265)		
3,300	3→4 (33%)	-0.0243	(0.0186)	1,729	6,999
Density p-value			(0.003)		
4,100	4→5 (25%)	-0.0066	(0.0216)	1,124	2,992
Density p-value			(0.418)		
4,900	5→6 (20%)	0.0046	(0.0219)	2,100	4,680
Density p-value			(0.519)		

Notes: Each row estimates a separate local linear RDD at the indicated population threshold. The outcome is an indicator for any coliform MCL violation. Density p-values from Cattaneo et al. (2020a) manipulation test. The 3,300 threshold shows significant density discontinuity ($p = 0.003$), consistent with the America's Water Infrastructure Act of 2018 creating a known administrative boundary. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3 reports estimates at each of the five largest thresholds. All five are individually null, with point estimates ranging from -0.024 to $+0.005$. The 1,000-person threshold—where monitoring doubles from one to two samples per month (the largest proportional increase)—yields an estimate of -0.004 ($p = 0.81$). Even with this dramatic monitoring change, there is no detectable effect on violations.

The consistency across thresholds is important. A story in which monitoring matters at some margins but not others would produce heterogeneous threshold-specific estimates. Instead, the uniform nulls suggest that the monitoring schedule is inert at every margin between 1 and 10 required samples.

5.3 Heterogeneity

The null result is consistent across source water types. Groundwater systems—which constitute 73 percent of the sample and face lower baseline contamination risk—show an estimate of -0.001 ($p = 0.90$). Surface water systems, which are more vulnerable to pathogen intrusion, show -0.029 ($p = 0.32$). Neither subgroup exhibits a significant monitoring effect.

Among ownership types, privately owned systems show a borderline positive estimate ($+0.046$, $p = 0.05$), suggesting that private systems crossing a monitoring threshold may detect slightly more violations. However, this result is sensitive to specification and may reflect multiple testing across subgroups. Publicly owned systems show a null (-0.021, $p = 0.13$).

5.4 Robustness

Table 4: Robustness of Multi-Cutoff RDD Estimates

Specification	Estimate	SE	N
Panel A: Bandwidth sensitivity			
$h = 200$	-0.0075	(0.0245)	2,795
$h = 300$	-0.0090	(0.0235)	2,964
$h = 500$	-0.0095	(0.0197)	4,473
$h = 750$	-0.0054	(0.0160)	7,016
$h = 1,000$	-0.0085	(0.0137)	9,509
$h = 1,500$	-0.0103	(0.0131)	10,761
$h = 2,000$	-0.0105	(0.0130)	10,916
Panel B: Polynomial order			
Order 1	-0.0109	(0.0128)	—
Order 2	-0.0114	(0.0142)	—
Panel C: Donut RDD			
Exclude $\lvert\Delta\text{pop}\rvert \leq 50$	-0.0092	(0.0285)	5,378
Exclude $\lvert\Delta\text{pop}\rvert \leq 100$	-0.0213	(0.0297)	8,553
Exclude $\lvert\Delta\text{pop}\rvert \leq 200$	-0.0151	(0.0151)	37,919
Panel D: Placebo cutoffs			
Pop. 1,750	-0.0073	(0.0183)	—
Pop. 2,900	0.0013	(0.0214)	—
Pop. 3,700	0.0277	(0.0260)	—
Pop. 4,500	0.0161	(0.0314)	—
Pop. 5,350	-0.0282	(0.0349)	—

Notes: All rows estimate the effect of crossing a monitoring threshold on the probability of any coliform MCL violation. Panel A varies the sample window around thresholds, letting `rdrobust` select its own bandwidth within each window. Panel B varies the polynomial order. Panel C excludes systems within the indicated population distance of a threshold. Panel D places fictitious thresholds at midpoints between actual monitoring cutoffs. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4 demonstrates the stability of the main finding across specifications. Panel A shows that the null persists across sample windows from 200 to 2,000 persons, with estimates ranging from −0.008 to −0.011. Panel B reports that quadratic polynomial specifications yield nearly identical estimates (−0.011 versus −0.011, consistent with Gelman and Imbens 2019's recommendation for low-order polynomials). Panel C excludes systems within 50, 100, and 200 persons of a threshold; all donut specifications remain null. Panel D places fictitious thresholds at midpoints between actual cutoffs; none produce significant effects, confirming that the nulls at real thresholds are not an artifact of the estimator.

6. Discussion

Why doesn't monitoring affect violations? Three mechanisms could explain the null.

Inframarginal monitoring. Most small systems already test at or above the minimum requirement. State regulators often impose additional monitoring beyond federal minimums, and systems may voluntarily test more frequently in response to source water conditions. If compliance monitoring is already adequate, the marginal federal sample adds little information.

Threshold ignorance. Many small water systems, particularly in rural areas, may not be aware of the exact population thresholds or adjust behavior in response to them. If the monitoring step-function creates variation that systems neither know about nor respond to, the null reflects institutional friction rather than a fundamental property of monitoring.

Detection at the margin. The one-additional-sample requirement may be too small to materially change detection probability. Under the TCR, a system collecting three samples per month that adds a fourth increases its sampling by 33 percent—but the probability that this single additional sample detects contamination, given that three samples did not, may be negligible. The detection channel requires both that contamination exists and that the marginal sample happens to capture it.

The results connect to a broader literature on whether regulatory monitoring deters noncompliance. Gray and Shimshack (2011) find that inspections and enforcement actions do reduce pollution, but primarily through the threat of penalties rather than monitoring per se. Duflo et al. (2013) show that restructuring *who monitors* (independent auditors versus industry-hired ones) dramatically improves pollution measurement and compliance in India. My findings suggest that increasing *how much* the same entity monitors, without changing incentives or enforcement, is insufficient to shift outcomes.

For environmental policy, the implication is that monitoring mandates are not a substitute

for enforcement capacity. The correlation between monitoring intensity and violations that motivates calls for more testing reflects confounding by system size and complexity. Policymakers considering tightening the RTCR monitoring schedule should recognize that additional mandatory samples may impose compliance costs on small systems without corresponding health benefits at the margin.

7. Conclusion

The Safe Drinking Water Act's coliform monitoring schedule creates nine natural experiments in which water systems just above a population threshold must test more frequently than nearly identical systems just below. None of these experiments produces a detectable effect on violations. The correlation between testing frequency and reported violations—which the untrained eye might read as evidence that monitoring "works"—is a *monitoring mirage*: an artifact of the confound between population size and contamination risk. For the typical public water system, the marginal sample neither deters noncompliance nor reveals hidden contamination—though borderline evidence for privately owned systems suggests that monitoring may matter more where baseline compliance is weaker. Regulators seeking to improve drinking water safety cannot rely on testing mandates alone.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers

Contributors: @olafdrw

First Contributor: https://github.com/olafdrw

References

Allaire, Maura, Haowei Wu, and Upmanu Lall, "National Trends in Drinking Water Quality Violations," *Proceedings of the National Academy of Sciences*, 2018, *115* (9), 2078–2083.

Banzhaf, Spencer, Lala Ma, and Christopher Timmins, "Environmental Justice: The Economics of Race, Place, and Pollution," *Journal of Economic Perspectives*, 2019, *33* (1), 185–208.

Bennear, Lori Snyder and Sheila M. Olmstead, "Right-to-Know, the Information Provision and Consumer Behavior: Evidence from the Safe Drinking Water Act," *Journal of Environmental Economics and Management*, 2009, *58* (2), 206–220.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik, "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 2014, *82* (6), 2295–2326.

Cattaneo, Matias D., Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare, "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *Journal of Politics*, 2016, *78* (4), 1229–1248.

_ , _ , _ , and _ , "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs," *Journal of the American Statistical Association*, 2021, *116* (536), 1941–1952.

_ , Michael Jansson, and Xinwei Ma, "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 2020, *115* (531), 1449–1455.

_ , Nicolás Idrobo, and Rocío Titiunik, *A Practical Introduction to Regression Discontinuity Designs: Foundations*, Cambridge University Press, 2020.

Duflo, Esther, Michael Greenstone, Rohini Pande, and Nicholas Ryan, "Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India," *Quarterly Journal of Economics*, 2013, *128* (4), 1499–1545.

Gelman, Andrew and Guido Imbens, "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 2019, *37* (3), 447–456.

Gray, Wayne B. and Jay P. Shimshack, "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence," *Review of Environmental Economics and Policy*, 2011, *5* (1), 3–24.

Keiser, David A. and Joseph S. Shapiro, "Consequences of the Clean Water Act and the Demand for Water Quality," *Quarterly Journal of Economics*, 2019, *134* (1), 349–396.

Marcus, Michelle, "Testing Above the Limit: Drinking Water Contamination and Test Scores," 2023. NBER Working Paper 31564.

Shimshack, Jay P. and Michael B. Ward, "Enforcement and Over-compliance," *Journal of Environmental Economics and Management*, 2008, *55* (1), 90–105.

Switzer, David and Manuel P. Teodoro, "Class, Race, Ethnicity, and Justice in Safe Drinking Water Compliance," *Social Science Quarterly*, 2018, *99* (2), 524–535.

Zivin, Joshua Graff, Matthew Neidell, and Wolfram Schlenker, "Water Quality Violations and Avoidance Behavior: Evidence from Bottled Water Consumption," *American Economic Review*, 2011, *101* (3), 448–453.

A. Data Appendix

EPA SDWIS. All data were downloaded from the EPA Envirofacts REST API. The Water System table provides system identifiers, population served, source water type (groundwater, surface water, or groundwater under the direct influence of surface water), ownership type, and service connection count. The Violation table provides violation identifiers, contaminant codes, violation category (MCL, monitoring, treatment technique), health-based indicator, and compliance period dates.

Coliform contaminant codes. Contaminant code 2950 denotes total coliform; code 3100 denotes *E. coli*. Both are captured under the Total Coliform Rule and Revised Total Coliform Rule. MCL violations for these contaminants are triggered when coliform-positive sample rates exceed regulatory thresholds in a given compliance period.

Population thresholds. The nine monitoring thresholds used in the multi-cutoff design are: 1,000; 2,500; 3,300; 4,100; 4,900; 5,800; 6,700; 7,600; and 8,500 persons. At each threshold, the required number of monthly coliform samples increases by exactly one. Above 8,500, sample requirements increase in larger steps (e.g., 10 to 15 at 12,900), which I exclude from the multi-cutoff analysis to maintain a homogeneous treatment effect interpretation.

Sample restrictions. The analysis includes all active community water systems (PWS type code CWS, activity code A) with population served greater than zero. Inactive, seasonal-only, and non-community systems are excluded. The API returns up to 100,000 rows per violation query; for the MCL and health-based violation categories used here, this captures the vast majority of records.

B. Identification Appendix

Density tests. The pooled density test at the normalized running variable cutoff of zero yields a t-statistic of 0.17 ($p = 0.86$), providing no evidence of manipulation (Cattaneo et al., 2020a). Threshold-specific tests are insignificant at 1,000 ($p = 0.19$), 2,500 ($p = 0.27$), 4,100 ($p = 0.42$), and 4,900 ($p = 0.52$). The 3,300 threshold shows a significant density discontinuity ($p = 0.003$), likely reflecting the America's Water Infrastructure Act of 2018, which independently creates a regulatory boundary at 3,300 persons by requiring risk and resilience assessments for systems above this size. The main results are robust to excluding the 3,300 threshold.

Covariate balance. At the pooled cutoff, the number of service connections ($\hat{\tau} = 68.8$, $p = 0.25$) and the share of surface water systems ($\hat{\tau} = 0.034$, $p = 0.13$) are balanced. These are the two observable covariates most likely to correlate with both contamination risk and population.

Mass points. The running variable (population served) exhibits mass points at round numbers (e.g., populations of exactly 100, 500, 1,000). The `rdrobust` estimation accounts for mass points in bandwidth selection and variance estimation. Donut RDD specifications that exclude systems near threshold round numbers produce consistent results (Table 4, Panel C).

C. Standardized Effect Sizes

Table 5: Standardized Effect Sizes for Main Outcomes

Outcome	$\hat{\beta}$	SE	SD(Y)	SDE	SE(SDE)	Classification
Panel A: Pooled						
Coliform MCL viol.	-0.0109	0.0128	0.227	-0.0482	0.0564	Small negative
Health-based viol.	-0.0019	0.0161	0.382	-0.0050	0.0421	Null
Non-coliform MCL viol.	-0.0204	0.0119	0.268	-0.0761	0.0446	Moderate negative
Panel B: Heterogeneous (by source water type)						
Coliform (groundwater)	-0.0010	0.0142	0.227	-0.0045	0.0625	Null
Coliform (surface water)	-0.0290	0.0271	0.227	-0.1280	0.1196	Moderate negative

Notes: **Country:** United States. **Research question:** Does increasing the regulatory monitoring frequency for coliform bacteria in community water systems reduce health-based drinking water violations, or does it primarily increase detection of pre-existing contamination? **Policy mechanism:** The Safe Drinking Water Act requires community water systems to collect coliform samples at a frequency that increases in steps with population served; crossing each of nine population thresholds between 1,000 and 8,500 adds one required sample per month, mechanically increasing the probability of detecting positive results and triggering Maximum Contaminant Level violations. **Outcome definition:** Binary indicator for whether a community water system recorded any Maximum Contaminant Level violation for total coliform (contaminant 2950) or *E. coli* (contaminant 3100) in EPA SDWIS. **Treatment:** Binary—crossing a population threshold that increases required monthly coliform samples by one. **Data:** EPA Safe Drinking Water Information System (SDWIS) via Envirofacts API, all years through 2025, community water system level, 49,172 systems total. **Method:** Multi-cutoff local linear RDD pooling nine population thresholds (1,000 through 8,500), triangular kernel, MSE-optimal bandwidth, bias-corrected robust inference (Calonico, Cattaneo, and Titiunik 2014). **Sample:** Active community water systems serving between 25 and 3.96 million persons; analysis sample restricted to systems within bandwidth of nearest monitoring threshold. SDE $= \hat{\beta}/\text{SD}(Y)$ where SD(Y) is the unconditional standard deviation. Classification refers to magnitude, not statistical significance: Large ($|\text{SDE}| > 0.15$), Moderate (0.05–0.15), Small (0.005–0.05), Null (< 0.005).