# Moral hazard under discrete information disclosure: Evidence from food-safety inspections

**John Bovay, Virginia Tech, bovay@vt.edu**

*Invited Paper prepared for presentation at the* **2021 AEA/ASSA Annual Meeting VIRTUAL, January 3-5, 2021**

# Moral hazard under discrete information disclosure: Evidence from food-safety inspections

John Bovay*

December 28, 2020

**Abstract**

This paper provides evidence on the responses of producers to mandatory disclosure of discrete quality ratings, a type of policy sometimes referred to as "naming and shaming". The context is a series of four regulatory changes undertaken by the U.S. Department of Agriculture (USDA) over 2006–2015, regarding disclosure of information about the results of tests for *Salmonella* in chicken carcasses at slaughter establishments. If establishments exceed certain rates of positive samples, they are designated "Category 2" or "Category 3". Under some policy regimes, Category 2 and 3 establishments have been listed on a public USDA website. I employ carcass-level data on *Salmonella* test results over 1999–2018 for all federally inspected establishments to explore the effects of public disclosure and other policy changes on *Salmonella* test outcomes. First, using a regression discontinuity approach, I demonstrate that: (1) When establishments fail to meet categorization thresholds but these failures do not subject them to public disclosure, *Salmonella* test performance worsens. (2) When establishments fail to meet thresholds and are therefore subjected to public disclosure, there is no statistically significant change in *Salmonella* test performance. (3) Under one policy regime, establishment operators relaxed efforts after sustained good *Salmonella* test performance ensured they would avoid public disclosure. Second, I document that when establishments have more leeway with respect to the thresholds, their performance on *Salmonella* tests worsens. Third, I use a regression discontinuity in time approach to demonstrate the effects of the series of policy changes on average *Salmonella* test results. I show that the introduction of public disclosure in 2008 reduced the average rate of positive *Salmonella* samples by about 55 percent. On the other hand, a tightening of standards in 2011 had a bifurcating effect wherein establishments that performed poorly (prior to the policy change) tended to perform even worse and middling establishments tended to improve. There was no statistically significant effect on the best-performing establishments. The net effect of the tightening of standards in 2011 was to more than double average *Salmonella* rates. My results provide evidence that the public disclosure of discrete information about *Salmonella* in chicken carcasses results in producers exerting extra effort to avoid disclosure, and less effort when disclosure status is already certain.



*Department of Agricultural and Applied Economics, Virginia Tech. bovay@vt.edu.

This paper has benefitted from helpful discussions with Aaron Adalja, Elinor Benami, Josh Beverly, Anne Byrne, Katie Harris-Lagoudakis, Pierre Mérel, Mike Ollinger, Brad Rickard, Ben Scharadin, Jay Shimshack, Will Townes, Wei Zhang, and participants at the 2020 SEA meeting.

**Moral hazard under discrete information disclosure: Evidence from food-safety inspections**

In various contexts, governments have regulated product quality to alleviate problems of moral hazard caused by information asymmetry. In some cases, policymakers and regulators have required reporting or labeling of information about product quality. Discrete information disclosure (as opposed to disclosure of information about quality expressed on some continuous scale) can be advantageous in that the information might be more easily understood by buyers. However, discrete reporting or labeling also may limit provision of quality by discouraging producers from attaining quality scores that greatly surpass the thresholds associated with each labeled category (Shewmake and Viscusi, 2015; Ito and Sallee, 2018). In other words, discrete information disclosure perhaps limits the incentives for moral hazard but does not eliminate the problem.

In this paper, I provide evidence on the responses of producers to mandatory disclosure of discrete quality ratings, a type of policy sometimes referred to as "naming and shaming". The context is a series of four regulatory changes undertaken by the U.S. Department of Agriculture (USDA) over 2006–2015,[1] regarding disclosure of information about *Salmonella* in chicken carcasses at slaughter establishments.[2] *Salmonella* in poultry is a major cause of food-borne illness in the United States, with an economic cost of up to $3.1 billion per year.[3] This paper documents the effects of categorization, publication of information about categories, and subsequent modifications to the policy on outcomes of tests for *Salmonella*.

My results provide evidence that the public disclosure of discrete information about *Salmonella* in chicken carcasses resulted in producers exerting extra effort to avoid disclosure, and less effort when disclosure status was already certain. Overall, the share of carcasses sampling positive for *Salmonella* decreased by about half when mandatory disclosure was implemented but then more than doubled after

[1] As shorthand, most references to dates refer to the policy periods beginning and ending in the referenced years, although the policy periods never coincided with calendar years. The dates that various policy changes were implemented are outlined in section 1 and in table 1.

[2] *Salmonella* is a genus of bacteria typically present in the intestines of birds and other animals. Meat and poultry can become contaminated with *Salmonella* if they come into contact with feces or the digestive tract.

[3] Hoffmann et al. (2015) report that *Salmonella* is the pathogen with the greatest economic cost of associated food-borne illnesses, $3.67 billion per year (in 2013 dollars), based on estimates of the incidence of illnesses, hospitalizations, and deaths from Scallan et al. (2011). However, there is tremendous uncertainty embedded in these estimates; Hoffmann et al. (2015) report that the 90% credible interval of the cost of illness from *Salmonella* is $193 million to $9.49 billion, in 2013 dollars. Furthermore, Painter et al. (2013) estimate that 10.1 to 29.2% of the cases of illness caused by *Salmonella enterica* are attributed to poultry. $3.1 billion = $.292 \times$ $9.49 billion, adjusted for inflation.

a subsequent tightening of standards. The details of the program—especially the discrete categorization system and the discrete sampling windows used throughout all regulatory periods—led to some shirking or moral hazard.

By testing hypotheses about the connections between public disclosure of discrete quality levels and quality outcomes, this paper contributes to a rich empirical literature. My analysis expands upon Ollinger and Bovay (2020), who study changes in annual average *Salmonella* test results at chicken-slaughter establishments, in the context of the introduction of public disclosure in 2008. By using carcass-level test result data and a longer time series, the analysis in this paper is richer and more robust. This paper also demonstrates that chicken processors responded to the incentives created by the inspection program by reducing effort related to food safety when the stakes were low. The results bear resemblance to studies reviewed by Dranove and Jin (2010), which found that hospitals and schools responded to the introduction of quality ratings by focusing on healthier patients and ignoring the sickest; with gaming behavior such as finding ways to avoid reporting scores of poorly performing students; and by teachers cheating on standardized tests (see also Dee et al., 2019). Similarly, Houde (2018) finds evidence that the energy efficiency of refrigerators is bunched just below the threshold necessary to obtain Energy Star certification, and Shewmake and Viscusi (2015) find that home builders strategically incorporate "green" features to achieve green certifications. Other related papers have studied the effects of disclosure on outcomes in the context of restaurant health-inspection scores (Jin and Leslie, 2003; Dai and Luca, 2020), drinking water (Bennear and Olmstead, 2008), and workplace safety violations (Johnson, 2020).

Under the *Salmonella* Verification Testing Program, in effect since 1999, USDA inspectors at chicken-slaughter establishments have randomly sampled chicken carcasses for *Salmonella*. No more than one carcass per establishment per day is sampled under the program. If establishments exceed certain numbers of positive samples within a "sample set" or "window" (i.e., a pre-designated number of sampled carcasses), they are designated "Category 2" or "Category 3". Under some policy regimes, Category 2 and 3 establishments have been listed on a public USDA website. This creates clear incentives for moral hazard, specifically a reduction in effort around controlling *Salmonella*. We would expect to see establishment operators reduce effort around *Salmonella* control in at least three cases. The first case is when the establishment exceeds the public-disclosure threshold before the end of a sample set. The second case is when the establishment has had very few positive samples, and it would therefore be impossible to exceed

the threshold no matter how many positive samples there were among the remaining samples. Third, when categorization is not yet determined, we would also expect to see a correlation between leeway with respect to the thresholds and *Salmonella* test performance (i.e., more leeway, worse test performance).

In this paper, I employ carcass-level data on *Salmonella* test results over 1999–2018 for all federally inspected chicken-slaughter establishments. First, using a regression discontinuity (RD) approach, I demonstrate that: (1) When establishments fail to meet categorization thresholds but these failures do not subject them to public disclosure, *Salmonella* test performance worsens. (2) When establishments fail to meet thresholds and are therefore subjected to public disclosure, there is no statistically significant change in *Salmonella* test performance. (3) Under one policy regime, establishment operators relaxed efforts after sustained good *Salmonella* test performance ensured they would avoid public disclosure.

Second, I document that when establishments have more leeway with respect to the thresholds, their performance on *Salmonella* tests worsens, as expected. The relationship between proximity to the thresholds and test outcomes is strong whether or not there is a threat of public disclosure, but tends to be stronger when the thresholds are associated with disclosure.

Third, I use a regression discontinuity in time approach to demonstrate the effects of the series of policy changes on overall *Salmonella* test results. I show that the introduction of public disclosure in 2008 reduced the overall rate of positive *Salmonella* samples by about 55 percent. On the other hand, a tightening of standards in 2011 had a bifurcating effect wherein establishments that performed poorly (prior to the policy change) tended to perform even worse and middling establishments tended to improve. Evidently, establishments were highly responsive to the incentives created by the thresholds. The effect of the 2011 tightening of standards on establishments with the lowest *Salmonella* rates was not statistically significant. The net effect of the tightening of standards in 2011 was to increase overall *Salmonella* rates by about 140 percent.

The safety of poultry processing continues to be relevant in legislation and policymaking today. This is partly because outbreaks of COVID-19 at meat and poultry processing plants made numerous headlines in the early stages of the pandemic, and were highlighted particularly for their impact on immigrants and ethnic minorities. In July 2020, bills were introduced into both chambers of the U.S. Congress to limit line speeds in chicken-slaughter establishments. According to a press release on Senator Cory Booker's

website, increased line speeds have negative implications for both worker health and food safety.[4] If increased line speeds are indeed associated with worse *Salmonella* outcomes, perhaps improved monitoring and disclosure of *Salmonella* test results could offset those welfare losses.

Section 1 provides additional background information on the chicken-slaughter industry and federal food-safety inspections. Section 2 describes the data and provides descriptive statistics. Section 3 demonstrates the effects of known categorization on *Salmonella* test outcomes using an RD design. Section 4 explores the effects of proximity to thresholds when categorization is unknown. Section 5 uses an RD in time approach to evaluate the effects of each policy change on average *Salmonella* test outcomes. Section 6 concludes. Appendices provide a description of the data-cleaning procedure and additional validation and robustness tests.

## 1. Background on the chicken-slaughter industry and food-safety inspections

Approximately nine billion meat chickens ("broilers") are produced each year in the United States, typically grown on farms under contract with slaughter and processing companies (MacDonald, 2015; USDA, 2019). In 2017, there were more than 32,000 farms growing meat chickens in the United States (USDA, 2019), and fewer than 300 federally inspected slaughter facilities. Under the Poultry Products Inspection Act, the USDA's Food Safety and Inspection Service (FSIS) is responsible for inspecting poultry and poultry products that enter interstate commerce. To facilitate traceability, poultry packages must include a USDA Mark of Inspection with the USDA-assigned establishment number.[5] In some states, state agencies also inspect poultry and poultry products for sale within the state, but the state standards must be at least as stringent as the federal inspection standards. Small processors (with fewer than 20,000 poultry processed annually) and processors with a retail business may qualify for exemption from inspection, if their products do not enter interstate commerce. During the period covered in this paper (1999 to 2018), there were up to 302 federally inspected chicken-slaughter establishments, but 75 of these exited the industry or opted for state inspection during the period. About 60% of establishments are located in the South, and 20% in the Midwest. Two large companies own about a quarter of all establishments;

[4]See https://www.booker.senate.gov/news/press/booker-introduces-bill-to-boost-safety-and-protect-meatpacking-workers-from-covid-19.

[5]See https://www.fsis.usda.gov/wps/portal/fsis/topics/food-safety-education/get-answers/food-safety-fact-sheets/food-labeling/how-to-find-the-usda-establishment-number/how-to-find-the-usda-establishment-number.

about half of all establishments are owned by 11 companies with four or more establishments.[6] Over the past two decades, establishments have closed at approximately the same rates in all regions; the rate of establishment closure by the largest companies mirrors the overall rate of closure in the industry.[7]

FSIS and its antecedent USDA agencies have regulated the safety of meat since 1890, when federal legislation enabled inspection of salted pork and bacon for export to certify that the products were trichinella-free (Olmstead and Rhode, 2015). Federal regulations on meat and poultry inspection have evolved over time, as have technologies to improve food safety. In recent decades, state public-health authorities adopted several technologies to facilitate tracing outbreaks of food-borne illness to their source, and have shared data as part of a nationwide network since 2001.[8] Around the same time, producers began to take advantage of new technologies to rapidly detect *Salmonella* and other pathogens in food, and gained greater ability to take action to prevent or limit the shipment of food products likely to be contaminated (Park et al., 2014; Page, 2018). Substantial evidence (e.g., Ollinger et al., 2004; Fulponi, 2006) suggests that private food-safety standards imposed by both buyers and producers have long surpassed government food-safety regulations; changes in private standards may have driven improvements in *Salmonella* test results just as much as changes in government standards and technology have.

As part of its regulatory activities, FSIS randomly samples chicken carcasses for pathogens including *Salmonella* and *Campylobacter*. Under the *Salmonella* Verification Testing Program, from 1999 to 2015, FSIS inspectors assigned ratings or categories to chicken-slaughter establishments based on the number of positive samples during recent "sample sets" (in FSIS terminology) of 51 carcasses sampled on 51 consecutive operating days. At first, this rating was essentially binary (establishments with 12 or fewer positive samples out of 51 met the standard) and ratings were not published. Minor sanctions were imposed in the event of three consecutive sample sets with more than 12 positive samples. Between 2006 and 2015, FSIS undertook several policy changes related to testing of chicken carcasses for *Salmonella* and public disclosure of results. The series of policy changes is summarized below and in table 1.

Starting on May 30, 2006, establishments that failed to meet the regulatory standard of 12 or fewer positive samples in a 51-sample set were designated Category 3. Establishments with 7 to 12 positive samples were designated Category 2; and establishments with 6 or fewer positive samples were designated

[6]Based on establishment names; the true concentration may be higher.
[7]Technically, what I observe regarding closure is that establishments are no longer federally inspected.
[8]See https://www.cdc.gov/pulsenet/anniversary/timeline.html.

Category 1. The new category designations were conveyed to firms privately until March 28, 2008, when the names and locations of Category 2 and 3 establishments were posted publicly on the FSIS website.[9] An establishment's information remained on the website until the establishment attained Category 1 status.

On July 1, 2011, the standard was tightened so that establishments with 6 or more positive samples out of 51 were designated Category 3 and establishments with 3 to 5 positive samples were designated Category 2. Starting on the same date, only the names and locations of Category 3 establishments were published. Put differently, the threshold for disclosure was reduced from 7 positive samples to 6, out of 51. Establishments would remain on the public list until they attained Category 1 or 2 status.

Effective May 6, 2015, the 51-sample-set framework was replaced with a system of categorization based on aggregated results over rolling 52-week windows. Public disclosure was temporarily suspended until categories could be assigned according to the new system. Under the new system, categories were defined using the same shares: an establishment with more than 9.8% of samples positive (i.e., 5/51) during any window of the windows ending the previous month would be placed on the Category 3 list and would remain on that list for a three-month period. This system remained in effect until November 23, 2018.

## 2. Data and descriptive statistics

Through a Freedom of Information Act (FOIA) request, I obtained data from FSIS on all test results from the *Salmonella* Verification Testing Program for broilers from January 4, 1999 to January 25, 2018. The data set also includes the address and name of establishments and snapshot information on the FSIS district and circuit to which establishments belonged, FSIS size classifications (very small, small, and large), and indicators for whether they processed other types of meat and active operation. All of the data on establishment characteristics reflects characteristics at the time of the data pull. The data set I obtained from FSIS does not include any indication of the groups of 51 samples ("sample

[9]The names of Category 2T establishments were also posted publicly starting March 28, 2008. Category 2T establishments were those that had been designated Category 2 or 3 based on the second-most-recent sample set but had improved to Category 1 performance in the most recent sample set. Effectively, the introduction of the Category 2T designation meant that a Category 2 or 3 establishment's name would be listed until it had completed two consecutive sample sets with 6 or fewer positive samples. The introduction of the Category 2T designation would not have changed the nature of incentives related to thresholds, but would have raised the stakes associated with a single "Category 2" outcome.

sets") used to determine regulatory compliance and category designations over 1999–2015. I am able to assign observations into sample sets by identifying lengthy temporal gaps between observations. I drop observations that are not likely to have been assigned correctly into sample sets based on this procedure, as including these observations would generate noise.[10]

I now provide some evidence that establishment operators were attentive to the thresholds and may have adjusted their operations to avoid exceeding the thresholds. In figure 1, I plot histograms of the number of positive samples per sample set for each of the four policy periods over 1999–2015.[11] Establishment operators were unable to precisely manipulate the number of positive samples per set because the presence of *Salmonella* bacteria in chicken carcasses cannot be precisely controlled and because carcasses were pulled out of processing lines at random to be sampled. Nevertheless, these histograms provide some evidence that establishment operators adjusted their operations in response to the thresholds and their positions relative to the thresholds. In particular, for most thresholds, there are many more sample sets one or two positive samples below the threshold than one or two positive samples above the threshold. Indeed, the thresholds tend to be associated with discontinuous drops in the number of sample sets at each level, when binning observations this way. For example, during the 2006–08 period, about 24.0% of sample sets had 3 or 4 positive samples, and 21.0% had 5 or 6, while only 8.4% had 7 or 8 and 7.0% had 9 or 10. The sharp drop in number of sample sets at the 6-positive-sample threshold, and relatively flat distribution further from the threshold, suggests that establishment operators exerted effort to stay at or below the threshold but relaxed efforts once above the threshold. Similar results are evident at the 12-positive-sample regulatory threshold in the 1999–2006 period and the Category 3 threshold in the 2006–08, 2008–11, and 2011–15 periods. Note, however, that during the periods in which disclosure of *Salmonella* categorization was in effect, there is no evidence of bunching at the maximum number of positive samples allowed for non-disclosure (i.e., 6 positive samples in 2008–11; 5 positive samples in 2011–15); establishment operators could not control *Salmonella* precisely enough to yield such results.

To provide further evidence that establishment operators were attentive to the regulatory thresholds, figure 2 plots the share of samples testing positive for *Salmonella* by test number within each sample set. For each policy period, I provide a pair or a quartet of graphs that break down observations by

[10] In essence, if the assignment into sample sets generates sets of many fewer or many more than 51 observations, I drop the sets. Details on the sample-set assignment procedure are given in Appendix A.

[11] Because sample sets were not used after May 2015, an equivalent histogram cannot be generated for the last policy period.

establishments' current status relative to the relevant thresholds. That is, the lefthand graph in each pair shows test observations from establishments with a share of positive tests at or below the threshold, in the current sample set at the time of the test; each dot represents the average share of samples positive for a given test number within sample sets. The righthand graph shows observations from establishments that had a higher percentage of tests positive than the threshold at the time of the test. The quadratic-fitted curves show clear patterns. When an establishment's rate of positive samples is at or below (above) the threshold(s), the likelihood of a positive sample falls (rises) from the beginning to the middle of a sample set and then rises (falls) again. These patterns imply that when an establishment's test performance is good (poor), operators exert less (more) effort when approaching the end of a sample set. In sections 3 and 4, I provide more careful analysis of the effects of an establishment's status relative to thresholds on test performance.

In addition to examining *Salmonella* test performance in the context of incentives related to the categorization and public disclosure thresholds, I examine the effects of policy changes on overall industry performance on *Salmonella* tests, in section 5. As seen in figure 3, the aggregate share of samples positive declined sharply over the period during which policy changes were being implemented, from 16.2% of samples positive in 2005 to 2.4% of samples positive in 2015, or a decline of nearly 1.4 percentage points per year. Since so many changes in technology and buyer requirements for food safety were taking place concurrently with FSIS policy changes, a careful empirical approach is needed to identify the effects of disclosure policies on producer behavior with respect to *Salmonella* control.

## 3. Effects of known categorization on *Salmonella* test outcomes

Under each policy regime, establishment operators faced somewhat different incentives related to controlling *Salmonella*. In particular, the penalties associated with exceeding the 5-, 6-, and 12-positive-sample thresholds were different under the various policy regimes. In this section, I use a regression discontinuity (RD) model to demonstrate how *Salmonella* test results changed when establishments crossed thresholds within a sample set, thus ensuring a particular categorization. My hypothesis is that to the extent that categorization and public disclosure matter, establishment operators relax efforts around *Salmonella* control when either (1) additional positive samples result in crossing a threshold into a worse category (Category 2 or 3) and (2) additional negative samples ensure a better categorization outcome (Category 1

or 2).

### *3.1. Empirical approach*

A natural and intuitive approach to studying the effects of crossing the discrete 5-, 6-, and 12-positive-sample thresholds on *Salmonella* test performance would be to use the number of positive samples within the sample set as a running variable in an RD design. However, such an approach only works when the cutoffs are crossed from below (i.e., when an establishment has an additional positive sample). Consider the following example. If 5 positive samples is the relevant threshold (as it was in 2011–15), and an establishment has had zero positive samples through 45 tests within a sample set, another negative sample would guarantee that the establishment will have no more than 5 positive samples out of the 51 samples in the set. In this case, the incentives for good *Salmonella* control as they relate to categorization and public disclosure could not be captured by using the number of positive samples as the running variable. In addition, an RD design with the number of positive samples as the running variable would not reflect the differential effects on effort of positive samples near the beginning of a sample set relative to positive samples near the end. For example, incentives differ when an establishment has 5 positive samples among the first 10, and when it has 5 positive samples among the first 50.

Given these considerations, the running variable used in the RD approach described in this section is the share of the remaining samples (within the sample set) that may be positive if the establishment is to achieve a given categorization (either Category 1 or 2). I term this variable $leewayC$, or leeway with respect to category threshold $C$, and formally define it as

$$leewayC_{ijk} = \frac{C - \sum_{l=1}^{i-1} Y_{ljk}}{52 - i}, \tag{1}$$

where $C \in \{2, 5, 6, 12\}$ is the maximum number of samples permitted to be positive within a sample set, to achieve the given category; $i$ is the test number within sample set $j$ at establishment $k$; and $\sum_{l=1}^{i-1} Y_{ljk}$ is a count of the number of positive observations within sample set $j$ at establishment $k$, within the interval $[1, i-1]$.[12] The denominator $52 - i$ is a count of the total number of observations that still need to be collected to complete the sample set, including $i$. I exclude any observations with $i > 51$, as these extra

[12]Figure 4 helps provide some intuition for the empirical approaches in this section and section 4.

samples would not have affected categorization.[13]

Thus, I use the following regression equation for the RD model to investigate the effects of crossing category thresholds on *Salmonella* test results:

$$Y_{ijk} = \alpha + \beta_0 D_{0ijk} + \beta_1 D_{1ijk} + f\left(leewayC_{ijk}\right) + \gamma_1 t_{ijk} + \gamma_2 i + \gamma_3 s_{j-1,k} + \varepsilon_{ijk}, \tag{2}$$

where $Y_{ijk}$ is a binary variable representing the results of test $i$ for *Salmonella* within sample set $j$ at establishment $k$ (positive = 1), $D_{0ijk} = \mathbf{1}\left\{leewayC_{ijk} \geq 0\right\}$, $D_{1ijk} = \mathbf{1}\left\{leewayC_{ijk} \geq 1\right\}$, $f\left(\cdot\right)$ is a polynomial function that can take on different values on either side of each cutoff ($c \in \{0, 1\}$); $t_{ijk}$ is the sample collection date; $s_{j-1,k}$ is establishment $k$'s share of samples positive in sample set $j-1$, and $\varepsilon_{ijk}$ is the residual.

Following Calonico et al. (2014), Cattaneo et al. (2020b), and Cattaneo et al. (2020c), I use sharp RD analysis with local linear regressions, triangular kernel weighting, bandwidths chosen to minimize mean squared errors on either side of both cutoffs, and robust nonparametric confidence intervals. In appendix tables, I also provide results using quadratic polynomials for the running variable and linear polynomials with Epanechnikov kernels to demonstrate the robustness of significant results from the main specifications. However, it should be noted that Cattaneo et al. (2020b) recommend using linear polynomials with triangular kernels.

### *3.2. Validity of the RD design*

In most contemporary studies that use RD approaches (see Lee and Lemieux, 2010; Calonico et al., 2014; Cattaneo et al., 2020b), two empirical tests are used to allay concerns that the running variable may be manipulated by agents (in this case, establishment managers or FSIS inspectors). One test shows that the running variable is smooth around the cutoff(s), that is, as-good-as-randomly distributed on either side of the cutoff(s) within a narrow band. This is typically tested using a density test as described by McCrary (2008); a recent update is proposed by Cattaneo et al. (2018). The second test shows that baseline covariates are also randomly distributed around the cutoff value(s) of the running variable by running an RD model on the baseline covariates.

[13] As discussed in Appendix A, FSIS inspectors sometimes collected more than 51 samples but the extra samples were not used for categorization.

Given that the running variable used in the regressions in this section is a ratio with some values (especially 0 and 1) much more common than others, density tests may yield spurious rejections of the null hypothesis (smoothness). To demonstrate this, I simulate 10,000 values of the $leewayC$ variables for each test $i \in \{1, \ldots, 51\}$ according to a Bernoulli distribution with the probability of a positive sample equal to the mean share of positives in each of the four policy periods when sample sets were used. The `rddensity` test proposed by Cattaneo et al. (2018) suggests that the running variable has discontinuous density at the cutoffs ($p < 0.001$) in nearly all cases using both the simulated and real data.[14] For another comparison of smoothness in the running variable, I use $t$-tests to compare the ratios of the number of observations with $leewayC = 0$ and $leewayC = 1$, over the number of observations with $leewayC < 0$ and $leewayC \in [0, 1]$, across my real and simulated data. I find that the real data are somewhat smoother than the simulated data at $leewayC = 0$ ($p = 0.097$) and almost exactly as smooth at $leewayC = 1$. Given that the running variable is inherently lumpy even in the simulated data, I conclude that the distribution of the running variable is as good as random around the cutoffs.

The second common way to test for manipulation of the running variable is to run an RD model on baseline covariates. A finding that the baseline covariates are discontinuous at the cutoffs may imply that agents are able to manipulate their status with respect to the cutoffs and that manipulation ability is somehow correlated with baseline characteristics of establishments. Because the running variable used in the regressions in this section is a ratio that takes on certain values much more frequently than other values, RD estimates of the effects of the actual cutoffs and many placebo cutoffs on the baseline covariates are statistically significant across many policy periods. I suggest that the unusual nature of the running variable makes a manipulation test based on baseline covariates inappropriate. Instead, I rely on a practical approach suggested by Eggers et al. (2015) and de la Cuesta and Imai (2016) to argue that manipulation is unlikely. Since agents cannot determine the values of their running variables with "extreme precision" (de la Cuesta and Imai, 2016),[15] it is unlikely that manipulation is done on the basis of predetermined covariates. Furthermore, visual examination of the histograms of the number of positive samples per completed sample set in figure 1 suggests that manipulation through post-test fraud

[14] For some of the cutoff and policy-period combinations, the `rddensity` test does not produce estimates using the simulated data because there are not enough observations on one side of one threshold.

[15] Recall that the denominator of the running variable is test number within the sample set, which cannot be controlled by the establishment managers. Furthermore, establishments had relatively poor ability to precisely control their share of positive tests and stay below the disclosure thresholds.

is also unlikely. When disclosure was in place (starting in 2008), the density of cumulative positive tests per sample set was clustered well below the disclosure thresholds, with no discontinuity just below the thresholds. The increased density of cumulative positive tests further below the thresholds suggest that establishment managers exerted (legitimate) effort to stay below the thresholds, and not that fraudulent behavior helped them stay below the thresholds.

### *3.3. Results: Effects of known categorization on Salmonella test outcomes*

The results of the RD models, shown in table 2, strongly suggest that establishment operators relaxed efforts around *Salmonella* control when categorization outcomes were known, especially when disclosure was not possible. Panel A of table 2 shows estimates of the RD coefficients at the $leewayC = 0$ and $leewayC = 1$ cutoffs for the thresholds $C$ associated with regulation or categorization but not with disclosure, and panel B shows estimates of the same RD coefficients for the thresholds $C$ associated with disclosure. The RD coefficients reflect the discontinuous effect of the running variable as it increases in value and passes each of the cutoffs. So, the interpretation of the coefficients is as follows: negative coefficients on the $leewayC = 0$ cutoffs imply that positive test results were less likely when $leewayC \in [0, 1)$ than when $leewayC < 0$; positive coefficients on the $leewayC = 1$ cutoffs imply that positive test results were more likely when $leewayC \geq 1$ than when $leewayC \in [0, 1)$. Interpretations of specific results in table 2 follow.

During the initial 1999–2006 period, when the category system had not yet been introduced and FSIS did not impose sanctions until establishments failed to meet the 12/51 threshold on three consecutive sample sets, crossing the $leeway12 = 0$ and $leeway12 = 1$ thresholds had no effect on subsequent *Salmonella* test performance.

During the 2006–08 period, when categorization was known only to the establishment (no disclosure), establishments had worse results after crossing the thresholds that ensured Category 2 and 3 outcomes. In particular, establishments were 6.1 percentage points more likely to have positive *Salmonella* test outcomes after failing to meet the 6/51 threshold necessary to be denoted Category 1, and 7.9 percentage points more likely to have positive samples after failing to meet the Category 2 standard (see table 2, panel A, columns 3 and 5).

During the 2008–11 policy period, the names of both Category 2 and 3 establishments were posted

on the FSIS website. My results, in table 2, panel B, columns 1–4, show that the cutoff values of *leeway*6 and *leeway*12 had statistically insignificant effects on subsequent *Salmonella* test performance.[16] This result is surprising, given that test results worsened after establishments failed to meet the Category 1 and 2 thresholds in the previous period, when establishment names were not disclosed. One possible explanation is that operators may have been especially diligent about *Salmonella* control during this period, for fear of increasingly stringent regulations down the line.

During the 2011–15 policy period, the thresholds associated with Category 2 and 3 were tightened so that Category 1 consisted of establishments with two or fewer positive samples out of 51 and Category 3 consisted of establishments with 6 or more. Under these new, more stringent thresholds, only the names of Category 3 establishments were publicly disclosed. During 2011–15, establishments were 8.9 percentage points more likely to have positive samples after failing to attain Category 1 status (table 2, panel A, column 7). As in the 2008–11 period, the cutoff values associated with the Category 3 threshold did not have statistically significant effects on *Salmonella* test performance.

During the 2015–18 period, sample sets were no longer used and establishments with more than 9.8 percent of samples positive during any 52-week window ending within the last three months were listed as Category 3 on the FSIS website. Table 2, panel A, column 9 shows that establishments were 3.3 percentage points more likely to have positive samples after failing to meet the Category 1 standard for the soonest-ending window.

Table 2, Panel B, column 8 shows that during the 2015–18 period, establishments were 2.2 percentage points more likely to have positive samples after being assured of meeting the Category 2 standard for the soonest-ending window. Since the rolling-window system to determine disclosure status ensured that a failed test result would affect categorization for the next 15 months, the soonest-ending window should not have been the only goalpost. Yet the results suggest that the continuous-sampling regime that replaced sample sets did not fully eliminate moral hazard as it was designed to do.

I now summarize the results in table 2. First, when establishments fail to meet thresholds but are not subject to public disclosure, *Salmonella* test performance worsens (panel A, columns 3, 5, 7, and 9). Second, when establishments fail to meet thresholds therefore subjecting them to public disclosure, there is no statistically significant change in *Salmonella* test performance (panel B, columns 1, 3, 5, and 7).

[16]The insignificant effects are robust to the polynomial and kernel choices.

Third, in most cases, establishment operators do not relax efforts after sustained good performance on *Salmonella* tests ensured they would avoid public disclosure. The exception was during the 2015–18 period, when the categorization system was revised to reduce the potential for moral hazard (panel B, column 8).

### *3.4. Robustness tests*

The RD results reported in tables A1 and A2 are for models that use quadratic polynomials and Epanechnikov kernels, respectively, but are otherwise identical to table 2. Again, keep in mind that Cattaneo et al. (2020b) recommend using local linear regressions in the running variable and triangular kernels, so the emphasis should be on where the various models reach similar conclusions, and not on where they diverge. I now review the statistically significant results in tables A1 and A2.

Table A1 uses triangular kernels and quadratic polynomials. Three of the 18 RD coefficients in table A1 are statistically significant with $p < 0.03$. Similar to the result in table 2, failing to meet the Category 1 standard in the 2006–08 period increased the likelihood of a positive test result by 5.1 percentage points (panel A, column 3); failing to meet the Category 1 standard in 2015–18 increased the likelihood of a positive test result by 4.5 percentage points (panel A, column 9); and establishments were 3.9 percentage points more likely to have positive samples in 2015–18 if they were certain to meet the Category 2 standard and avoid disclosure (panel B, column 8).

Table A2 uses linear polynomials and Epanechnikov kernels; three of the 18 RD coefficients are statistically significant with $p < 0.01$ and one is marginally significant. According to these results, failing to meet the regulatory standard (12 positive samples) in 1999–2006 increased the likelihood of a positive sample by 2.2 percentage points (panel A, column 1). This result suggests that during this period, the regulatory standard did not create particular incentives for moral hazard and that establishment operators exerted effort to improve *Salmonella* test performance after failures. (Recall that FSIS did not impose any kind of sanctions during this period until establishments failed to meet the standards on three consecutive sample sets.) Similar to results in both tables 2 and A1, failing to meet the Category 1 standard in the 2006–08 period increased the likelihood of a positive test result by 5.8 percentage points (panel A, column 3); and failing to meet the Category 1 standard in 2015–18 increased the likelihood of a positive sample by 3.3 percentage points (panel A, column 9). Finally, consistent with the other two

sets of results, establishments were 2.0 percentage points more likely to have positive samples in 2015–18 if they were certain to meet the Category 2 standard and avoid disclosure (panel B, column 8). The last effect is weaker than in the other two sets of results ($p = 0.085$).

Two RD coefficients are consistently estimated to be statistically significant with $p < 0.05$ (and have the same sign) across all three sets of tables. These results both suggest moral hazard or shirking in specific policy periods. To recap, establishments' test results worsened after they failed to meet the Category 1 standard in 2006–08, before the introduction of public disclosure. They also worsened after good test performance ensured they would avoid disclosure during the 2015–18 period, even though the sample-set system had been replaced during that period with categorization based on rolling windows to encourage establishment operators to maintain good safety regardless of recent test results.

### *3.5. Results for placebo cutoffs*

In table A3, I present results for regressions parallel to those in table 2 using placebo cutoff values for the running variables ($leewayC$). The time periods and thresholds shown here represent the statistically significant estimates from table 2. The placebo cutoff values are three multiples of 0.05 in either direction from $c = 0$; and the nearest multiples of 0.05 to $c = 1$ for which optimal bandwidths (in the sense of minimizing mean squared errors) could be computed using the `rdms` command in Stata (Cattaneo et al., 2020c).

In panels A through D of table A3, five of the 24 RD coefficients are statistically significant with $p < 0.1$, but only one of these has the "correct" sign in the sense that it is consistent with the estimate for $c = 0$ in table 2 and the expectations about incentives for shirking that motivate the analysis in this section. While the number of statistically significant placebo coefficients is higher than one would expect, it is reassuring that all but one have incorrect signs (whereas all five significant coefficients in table 2 have the correct signs).

One of the placebo cutoffs shown in panel E of appendix table A3 ($leeway5 = 1.3$) yields an even larger coefficient estimate (3.9 percentage points) than the result in table 2, column 8. This placebo result is more precisely estimated than the one in table 2 and is robust to polynomial and kernel choices. While this placebo result appears to raise doubts about the main result, it is not an actual concern. Recall that the regressions for 2015–18 in table 2 use the soonest-ending window to calculate the value

of the leeway variable but that each observation was associated with up to 52 windows. Conditioning on establishments meeting the Category 1 or 2 standard for the soonest-ending window, crossing the $leeway5 = 1$ cutoff for the second-soonest-ending window leads to a 5.1 percentage point rise in the likelihood of subsequent positive samples (see appendix table A4).[17] Therefore, the RD coefficient on the placebo cutoff $leeway5 = 1.3$ provides additional support for the finding that establishments relaxed efforts when they were certain to avoid having their names listed as Category 3 on the FSIS website.

## 4. Proximity to regulatory thresholds and *Salmonella* test outcomes

The previous section demonstrates that in some cases, crossing regulatory thresholds with a series of positive or negative *Salmonella* test results leads to significantly worse results on subsequent *Salmonella* tests. In this section, I use a series of regressions to evaluate the relationship between proximity to thresholds, when multiple category outcomes are still possible, and *Salmonella* test performance. As in the previous section, the dependent variable is the binary *Salmonella* test result. The key explanatory variable in these regressions is again $leewayC$. Larger values of $leewayC$ indicate that a larger share of remaining samples could test positive for *Salmonella*. Therefore, if the *Salmonella* category assignment matters to producers, then *Salmonella* control efforts should increase when the value of $leewayC$ is smaller within the $[0, 1)$ interval. To estimate the relationship between $leewayC$ and test outcomes when multiple category outcomes are possible, I use only observations with $leewayC \in [0, 1)$.

I estimate the relationship between $leewayC$ and *Salmonella* test outcomes under each policy regime using a series of linear probability models, according to equation 3:

$$Y_{ijk} = \alpha + \beta leewayC_{ijk} + \gamma_1 i + \gamma_2 s_{ijk} + u_{km(ijk)} + \varepsilon_{ijk}, \tag{3}$$

where $Y_{ijk}$ is a binary variable representing the results of test $i$ for *Salmonella* within sample set $j$ at establishment $k$ (positive = 1); $s_{ijk}$ is the share of samples positive within the current sample set (over tests $1, \ldots, i-1$); $u_{km(ijk)}$ represents establishment–month–year fixed effects; and $\varepsilon_{ijk}$ is the residual. For the 2015–18 period, I cannot use the test number regressor ($k$) because sample sets were not used

[17] In my data set, $leeway5_1 = 4/3$ corresponds to $leeway5_2 = 1$ about two thirds of the time (where the subscripts indicate the soonest-ending and second-soonest-ending windows). So, if the placebo cutoff $leeway5_1 = 1.3$ is estimated to increase the rate of positive *Salmonella* tests by 3.9 percentage points, this corresponds roughly with the estimated effect of crossing the $leeway5_2$ cutoff (5.1 percentage points $\times$ 2/3 = 3.4 percentage points).

during this period and windows had varying numbers of observations. Note that in the 2015–18 period, the analysis uses the *leewayC* variable that reflects the soonest-ending window.

Admittedly, there are some shortcomings in the identification strategy described here, given that $leewayC_{ijk}$ is (mechanically and empirically) negatively correlated with the share of samples positive $s_{ijk}$ and positively correlated with the test number $i$. However, it is essential to control for recent test results at each establishment, given that average test results vary widely across establishments. Establishment operators cannot (precisely) control any of these three regressors, so *leewayC* is plausibly exogenous. By including $s_{ijk}$ and $i$ as regressors, I can tease out effects of proximity to the threshold on *Salmonella* control efforts. Moreover, my empirical results are generally consistent whether or not I include $s_{ijk}$ as a regressor.

Table 3 presents results from regressions of the form described by equation 3, which demonstrate the effect of proximity to the thresholds on *Salmonella* test outcomes, for the policies in place from 1999 to 2015. Table 3 demonstrates that in all periods, when the value of *leewayC* was larger, carcasses were more likely to test positive for *Salmonella*. In other words, establishments controlled *Salmonella* better when it was necessary to ensure a better categorization outcome. These results hold regardless of whether the policy of public disclosure of Category 2 and 3 outcomes was in place. I now review the results in more detail.

Panels A and B of table 3 report results for the regressions for the policies in place from 1999 to 2015 with respect to the thresholds equivalent to the maximum number of positive samples associated with Categories 1 and 2, respectively.[18] From 1999 to 2006, when categorization had not yet been introduced but 12 positive samples out of 51 was a regulatory requirement, *Salmonella* test outcomes were worse when establishments were closer to both the 6- and 12-positive-sample thresholds. When the *leeway*12 value was 10 percentage points higher, the probability of a positive test result was 4.88 percentage points higher ($p < 0.001$). The elasticity of the share of samples positive with respect to *leeway*12 was 1.39, calculated using the mean share of samples positive and the mean value of *leeway*12.

From 2006 to 2008, when categorization was reported privately, the relationship between proximity to the 12-positives threshold and *Salmonella* test outcomes was stronger (columns 3 and 4). When the *leeway*6 value was 10 percentage points higher, the probability of a positive test result was 3.39 percentage

[18] All discussion of results in tables 3 and 4 references the even-numbered columns, as they are the preferred specifications.

points higher ($p < 0.001$; elasticity = 0.74), and when the *leeway*12 value by 10 percentage points higher, the probability of a positive test result was 3.82 percentage points higher ($p < 0.001$; elasticity = 1.52).

Public disclosure of the names of both Category 2 and 3 establishments from 2008–11 further strengthened the relationship between proximity to the thresholds and test results (columns 5 and 6). During this period, when the *leeway*6 value was 10 percentage points higher, the probability of a positive test result was 2.28 percentage points higher ($p < 0.001$; elasticity = 0.86), and when the *leeway*12 value was 10 percentage points higher, the probability of a positive test result was 4.11 percentage points higher ($p < 0.001$; elasticity = 2.46).

Over 2011–15, the standards were tightened and only the names of Category 3 establishments were posted. Correspondingly, the relationship between the leeway value associated with the Category 1/2 threshold and test outcomes was weaker over 2011–15 (columns 7 and 8). When the *leeway*2 value was 10 percentage points higher, the probability of a positive test result was 0.91 percentage points higher ($p < 0.001$; elasticity = 0.43). The relationship between the leeway value associated with the Category 2/3 threshold and test outcomes was also highly significant but much weaker than in the 2006–08 and 2008–11 periods: when the *leeway*5 value was 10 percentage points higher, the probability of a positive test result was 0.96 percentage points higher ($p < 0.001$; elasticity = 0.84).

As previously discussed, from 2015–18, categorization and disclosure were based on results over 52-week rolling windows. In table 4, I present results of regressions that use *leeway*2 and *leeway*5 for the soonest-ending window; and in addition, results of regressions that use the *leeway* values for all (i.e, up to 52) windows associated with any given test result. In table 4, I do not report coefficient estimates for the windows after the soonest-ending window because they are mostly insignificant. In the full specifications, there is no statistically significant relationship between *leeway*2 and *Salmonella* test outcomes. However, when the *leeway*5 value for the soonest-ending window was 10 percentage points higher, the probability of a positive *Salmonella* test result was 2.58 percentage points higher (p=0.013; elasticity = 1.82) holding the *leeway* values for all other windows constant.

What should we take away from all of these results? To put it most simply, incentives matter. *Salmonella* test results were better when they needed to be. Proximity to thresholds mattered whether or not there was a threat of public disclosure, but the relationship between proximity and test outcomes tended to be stronger when the thresholds were associated with disclosure when considering the rela-

tionships as elasticities. Lastly, in the 2015–18 periods, *Salmonella* test results were strongly associated with the Category 2 disclosure threshold for the soonest-ending window, even though each individual test result would be used to determine categorization for up to 15 months into the future. Most of the results outlined above are consistent with the findings in the previous section, which suggested that *Salmonella* test results generally worsened when Category 1 status was impossible, and that over 2015–18, test results worsened when attaining Category 2 (rather than Category 3) was guaranteed.

## 5. Effects of changes in categorization and disclosure policies on average *Salmonella* test outcomes

Lastly, I evaluate the effects of the series of policy changes on average *Salmonella* test results. Here, I use a regression discontinuity in time (RDiT) approach (Hausman and Rapson, 2018). This section builds on the results of Ollinger and Bovay (2020), who evaluate the effects of the 2006 and 2008 FSIS policy changes on *Salmonella* test performance of chicken-slaughter establishments using annual data on the establishments' average share of samples positive. Ollinger and Bovay (2020) find that the introduction of the categorization system in 2006 led to a 6 to 10 percentage-point reduction in the share of samples positive and that public disclosure of the names of Category 2 and 3 establishments in 2008 led to a 3 to 5 percentage-point reduction. By using carcass-level test result data and an RDiT approach, the analysis in this section refines the earlier analysis of Ollinger and Bovay (2020).

As in section 3, I use sharp RD analysis with local linear regressions, triangular kernel weighting, bandwidths chosen to minimize mean squared errors on either side of each cutoff, and robust nonparametric confidence intervals (Calonico et al., 2014; Cattaneo et al., 2020b,c). The regression equation is as follows:

$$Y_{ikt} = \alpha + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + f(t) + \varepsilon_{ikt}. \tag{4}$$

The running variable is the sample collection date and the four dates of policy changes are the cutoffs. The binary dependent variable $Y_{ikt}$ is the *Salmonella* test outcome for sample $i$ at establishment $k$ on date $t$ (positive = 1), $D_{jt} = \mathbf{1}\{t \geq c_j\}$ for each of the four cutoffs $c_j$, $f(\cdot)$ is a polynomial function that can take on different values on either side of each cutoff, and $\varepsilon_{ikt}$ is the residual. The RD bandwidths are

selected separately for each date of policy change to minimize mean squared error on each side of each cutoff date, as recommended by Cattaneo et al. (2020b). As discussed by Hausman and Rapson (2018), tests for smoothness in density of the running variable are inappropriate to establish the validity of RDiT designs.

### *5.1. Results: Effectiveness of policy changes*

Panel A of table 5 presents results from the RDiT model described by equation 4 using all observations from all establishments. The results suggest that the introduction of public disclosure in 2008 led to a 5.1 percentage point reduction in the probability of positive *Salmonella* samples. Given that 9.2 percent of samples tested positive for *Salmonella* during the 177 days before the policy change (i.e., the MSE-optimal bandwidth), the introduction of public disclosure reduced *Salmonella* levels by 55 percent. Other policy changes, in 2006, 2011, and 2015, had statistically insignificant effects on average test outcomes.

Including observations from establishments that were active in earlier periods but not in later periods may bias the results in panel A if, for example, establishments with worse food safety were more likely to exit the industry for reasons unrelated to FSIS inspections and disclosure policies. Panel B drops all establishments that were listed as "inactive" at the time the data set was created. In this way, panel B achieves better balance of (unobserved) covariates than panel A. The results in panel B suggest again that the introduction of public disclosure in 2008 led to a large (4.8 percentage point; 55 percent) reduction in the probability of positive *Salmonella* samples, but that the subsequent tightening of the thresholds in 2011 led to an even larger (6.8 percentage point; 139 percent) increase.[19] There are a couple of different interpretations of the estimated increase in positive *Salmonella* samples starting in 2011, when removing establishments that ever exited. One is that many establishments with worse performance exited around the time of the 2011 policy change. The other is that many operators of worse-performing establishments remained active but gave up on trying to meet the now more stringent standard necessary to avoid disclosure.

To explore the first of these two interpretations, I query the data and find that ten establishments exited during the 2011–15 policy period. On average, these establishments had 8.8 percent of samples test

[19]Panel B uses different bandwidths than panel A, again by minimizing mean squared error on each side of each cutoff date. Percent changes are again calculated using the share of samples positive within the MSE-optimal bandwidth before the policy changes as the baselines.

positive for *Salmonella* during this policy period, as compared with 4.0 percent for all other establishments ($p < 0.0000$ for $t$-test for difference in means). However, only three of the ten ever reached the 6-sample threshold necessary to be listed as Category 3 during the 2011–15 period. So, while the establishments that exited during 2011–15 had worse *Salmonella* test results on average, it is not clear that establishments exited because of the increased stringency that began in 2011.

The latter interpretation, that operators gave up on trying to meet the now more stringent standard, appears to be more plausible. Table 6 shows the estimated RDiT effect of the 2008 and 2011 policy changes, splitting the samples by establishment-level average *Salmonella* test results over 2006–08 and 2008–11, respectively.[20] The 2008 policy change is estimated to have reduced the share of samples positive for establishments at each performance level, although the effect is only statistically significant for those with average test results equivalent to Category 1. Establishments with different safety records responded to the 2011 policy change differently. Establishments that had an average of more than 5 out of 51 (about 9.8 percent) positive samples during the 2008–11 period (corresponding to the 2011–15 Category 3 threshold) had a 17.7 percentage point (111 percent) increase in the likelihood of positive samples at the time of the 2011 policy change. Meanwhile, establishments with average test results during 2008–11 that would place them in the new Category 2 (more than 2, and no more than 5 positive samples out of 51) had a 3.9 percentage point decrease in positive samples at the time of the policy change. As stated above, the overall effect was to greatly increase the share of samples positive, by 6.8 percentage points or about 139 percent, among establishments that remained active through 2018.

How did the complete series of policy changes affect average *Salmonella* levels in chicken carcasses? Adding up the RDiT coefficients in panel A or panel B of table 5 (regardless of statistical significance) suggests that the four policy changes cumulatively increased the share of samples testing positive by between 0.6 and 2.9 percentage points. If we only consider the statistically significant coefficients in tables 5 and A5, the net effect of the four policy changes is somewhere between a 5.1 percentage-point decrease and a 12.7 percentage-point increase. Of these, the best estimate is probably that in panel B of table 5: that the 2008 policy change decreased the likelihood of positive *Salmonella* samples by 4.8 percentage points and the 2011 policy change increased the likelihood by 6.8 percentage points, a net increase of 2.0 percentage points (23 percent relative to the 2006–08 baseline).

[20] All results described in the rest of this section use the same data set as panel B of table 5, dropping all establishments that ever exited.

### *5.2. Effectiveness of policy changes: robustness tests*

For RDiT approaches to analysis of policy changes, Hausman and Rapson (2018) recommend a few additional robustness tests. First, as recommended by Cattaneo et al. (2020a) for RD designs where the data have many "mass points", I collapse the data set and use the daily share of samples positive, across all establishments, as my dependent variable. The results, in panel A of table A5, essentially conform with the results in panel B of table 5 above: the introduction of public disclosure in 2008 led to a 4.3 percentage point decrease in the share of samples positive, while the tightening of standards in 2011 led to a 6.4 percentage point increase. The 2015 policy change is also estimated to have led to a 2.4 percentage point decrease, although this result is marginally significant ($p = .096$).

Second, I employ a "donut" approach as recommended by Barreca et al. (2011) to ensure that *Salmonella* sampling dates were not subject to manipulation around the dates of the policy changes, which might have occurred if sampling dates were misreported or establishments briefly shut down before or after policy changes. These results are again similar to the main results in table 5 above. The donut specifications, removing all observations within 1 to 7 days on both sides of policy changes, yield somewhat larger estimated effects of the 2008 policy change (a 4.9 to 5.8 percentage point decrease in the share of samples positive) and somewhat smaller estimated effects of the 2011 policy change (a 6.3 to 6.7 percentage point increase) than the main specification. Panel B of table A5 shows results for the RDiT regression with all observations within 7 days of the policy changes removed. For the 7-day donut specification only, the 2015 policy change is estimated to have decreased the share of samples positive by 2.5 percentage points ($p = 0.096$).[21]

Third, I drop all observations belonging to sample sets that span two policy periods. Under each policy regime, category status was assigned on the basis of sample sets as they were completed; incomplete sample sets were not reset at the time of the policy changes. The exception is the 2015–18 period, under which the "sample set" concept was not used. When I drop observations from sample sets that span policy periods, the estimated RDiT effects change somewhat: the introduction of disclosure in 2008 resulted in a 5.4 percentage point decrease in the share of samples positive (though not statistically significant), while the 2011 tightening of standards led to a 12.7 percentage point increase ($p = 0.002$).

While the various specifications yield somewhat different point estimates, the sign and magnitude

[21] For the smaller donuts, the 2015 effect is insignificant with $p > 0.13$.

of the estimates are fairly consistent. The introduction of mandatory disclosure in 2008 resulted in a significant improvement in average *Salmonella* test results, roughly a 55 percent reduction in the share of samples positive. Perversely, though, the tightening of standards in 2011 resulted in a significant worsening of test results, more than doubling the share of samples positive. I documented above that establishments' responses to the 2011 policy change are correlated with *Salmonella* test performance in the prior period.

As another robustness test, I use several sets of placebo dates of policy changes. Each policy change was preceded by an announcement in the Federal Register about the scheduled policy change. In Panel A of table A6, I use the dates of each Federal Register announcement as the cutoffs. I find that *Salmonella* test results did not change discontinuously at the dates of the announcements. In Panels B through E of table A6, I use placebo dates 120, 240, 360, and 480 days before the actual policy changes. Under the null hypothesis, with 20 placebo cutoff values, one placebo would be expected to have $p < 0.05$ and two would be expected to have $p < 0.10$. In table A6, the lowest $p$-value is 0.094. We can therefore be confident in the estimated effects of the policy changes presented above.

To recap, the introduction of public disclosure in 2008 decreased the rate of positives by about 55 percent. When only considering establishments that remained active until 2018, the tightening of standards in 2011 more than doubled the rate of positives, a result driven by the worst-performing establishments. It is clear that while the initial public disclosure policy was successful in improving the average rate of positive *Salmonella* samples, the next policy change worsened test outcomes and more than offset this improvement. The RDiT approach using carcass-level data provides convincing evidence that the series of policy changes over 2006–15 had little effect on average *Salmonella* test results.

## 6. Summary and conclusion

Using carcass-level data on USDA inspections for *Salmonella* in chicken carcasses from 1999 to 2018, I demonstrate several ways in which chicken-slaughter establishments responded to incentives created by the inspection, categorization, and disclosure policies. First, using a regression discontinuity approach, I demonstrate that when establishments fail to meet categorization thresholds, *Salmonella* test performance worsens only if the failures do not subject them to public disclosure. Under the more stringent

disclosure policy in place from 2015 to 2018, establishment operators relaxed efforts after sustained good test performance ensured they would avoid public disclosure. Second, I document that when two or more categorization outcomes are possible and establishments have more leeway with respect to the thresholds, their performance on *Salmonella* tests worsens. Third, while the initial public disclosure policy in 2008 reduced the average rate of positive *Salmonella* samples by about 55 percent, the net effect of the series of policy changes was to increase the average rate of positive samples by about 23 percent. I demonstrate that the worst-performing establishments had much worse performance after the 2011 tightening of standards.

There are some limitations to this study, naturally. The formal tests for manipulability of the running variable in the RD models on categorization fail because of the lumpy nature of the running variable. Also, the RD coefficient estimates in section 3 are not always robust to different specifications, such as using quadratic polynomials in the running variable or Epanechnikov kernels. The identification strategy used in section 4 to study the relationship between leeway and test results when two or more categories were possible may not permit causal claims. There are some drawbacks to the data set I obtained from FSIS, too. It has very few time-varying covariates that can be used in any of the regressions. Lastly, there is some uncertainty about the sample sets I reconstructed for this analysis. Nonetheless, the paper shows convincingly that slaughter establishments responded to the perverse incentives created by the FSIS testing and disclosure system. Moreover, it shows that the series of policy changes from 2006 to 2015 had little net impact on *Salmonella* in chicken carcasses, when controlling for exit from the industry.

## References


**Barreca, Alan I, Melanie Guldi, Jason M Lindo, and Glen R Waddell.** 2011. "Saving babies? Revisiting the effect of very low birth weight classification." *The Quarterly Journal of Economics*, 126(4): 2117–2123.

**Bennear, Lori S, and Sheila M Olmstead.** 2008. "The impacts of the 'right to know': Information disclosure and the violation of drinking water standards." *Journal of Environmental Economics and Management*, 56(2): 117–130.

**Calonico, Sebastian, Matias D Cattaneo, and Rocío Titiunik.** 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica*, 82(6): 2295–2326.

**Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik.** 2020a. *A Practical Introduction to Regression Discontinuity Designs: Extensions.*: Cambridge University Press.

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

**Cattaneo, Matias D, Michael Jansson, and Xinwei Ma.** 2018. "Manipulation testing based on density discontinuity." *The Stata Journal*, 18(1): 234–261.

**Cattaneo, Matias D, Rocío Titiunik, and Gonzalo Vazquez-Bare.** 2020c. "Analysis of regression discontinuity designs with multiple cutoffs or multiple scores." *Stata Journal*, forthcoming.

**Dai, Weijia, and Michael Luca.** 2020. "Digitizing disclosure: The case of restaurant hygiene scores." *American Economic Journal: Microeconomics*, 12(2): 41–59.

**de la Cuesta, Brandon, and Kosuke Imai.** 2016. "Misunderstandings about the regression discontinuity design in the study of close elections." *Annual Review of Political Science*, 19 375–396.

**Dee, Thomas S, Will Dobbie, Brian A Jacob, and Jonah Rockoff.** 2019. "The causes and consequences of test score manipulation: Evidence from the new york regents examinations." *American Economic Journal: Applied Economics*, 11(3): 382–423.

**Dranove, David, and Ginger Zhe Jin.** 2010. "Quality disclosure and certification: Theory and practice." *Journal of Economic Literature*, 48(4): 935–63.

**Eggers, Andrew C, Anthony Fowler, Jens Hainmueller, Andrew B Hall, and James M Snyder Jr.** 2015. "On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races." *American Journal of Political Science*, 59(1): 259–274.

**Fulponi, Linda.** 2006. "Private voluntary standards in the food system: The perspective of major food retailers in OECD countries." *Food Policy*, 31(1): 1–13.

**Hausman, Catherine, and David S Rapson.** 2018. "Regression discontinuity in time: Considerations for empirical applications." *Annual Review of Resource Economics*, 10 533–552.

**Hoffmann, Sandra Ann, Bryan Maculloch, and Michael Batz.** 2015. "Economic burden of major foodborne illnesses acquired in the United States." Economic Information Bulletin No. 140, United States Department of Agriculture, Economic Research Service.

**Houde, Sébastien.** 2018. "Bunching with the stars: How firms respond to environmental certification." Working paper, ETH Zurich, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3222513.

**Ito, Koichiro, and James M Sallee.** 2018. "The economics of attribute-based regulation: Theory and evidence from fuel economy standards." *Review of Economics and Statistics*, 100(2): 319–336.

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

**Johnson, Matthew S.** 2020. "Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws." *American Economic Review*, 110(6): 1866–1904.

**Lee, David S, and Thomas Lemieux.** 2010. "Regression discontinuity designs in economics." *Journal of economic literature*, 48(2): 281–355.

**MacDonald, James M.** 2015. "Trends in agricultural contracts." *Choices*, 30(3): 1–6.

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

**Ollinger, Michael, and John Bovay.** 2020. "Producer response to public disclosure of food-safety information." *American Journal of Agricultural Economics*, 102(1): 202–219.

**Ollinger, Michael, Danna Moore, and Ram Chandran.** 2004. "Meat and poultry plants' food safety investments." U.S. Department of Agriculture Economic Research Service Technical Bulletin No. 1911.

**Olmstead, Alan L, and Paul W Rhode.** 2015. *Arresting Contagion: Science, Policy, and Conflicts over Animal Disease Control.*: Harvard University Press.

**Page, Elina Tselepidakis.** 2018. "Trends in food recalls: 2004-13."Technical report, United States Department of Agriculture, Economic Research Service.

**Painter, John A, Robert M Hoekstra, Tracy Ayers, Robert V Tauxe, Christopher R Braden, Frederick J Angulo, and Patricia M Griffin.** 2013. "Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008." *Emerging Infectious Diseases*, 19(3): 407–415.

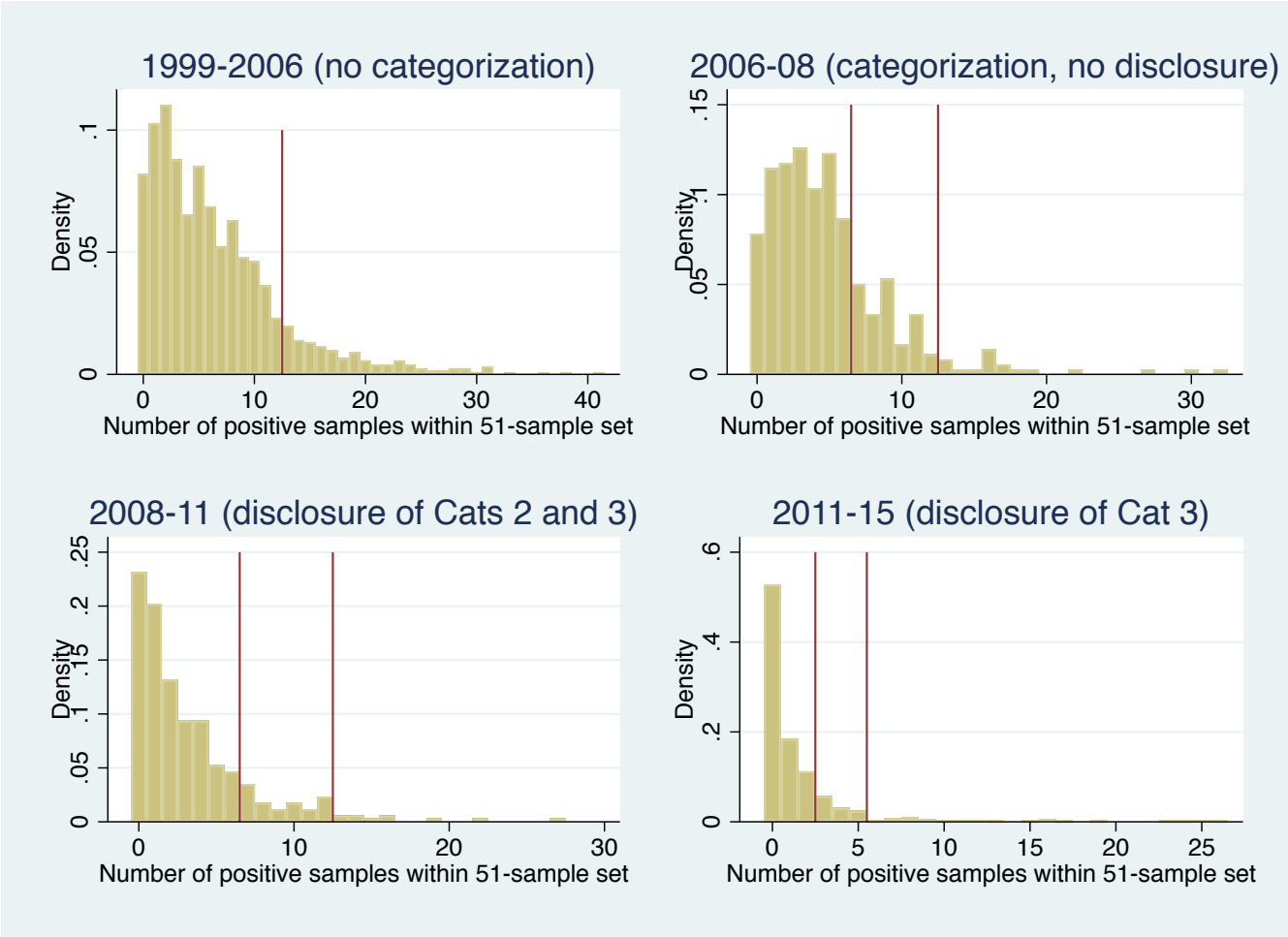
**Park, Si Hong, Muhsin Aydin, Anita Khatiwara, Maureen C Dolan, David F Gilmore, Jennifer L Bouldin, Soohyoun Ahn, and Steven C Ricke.** 2014. "Current and emerging technologies for rapid detection and characterization of Salmonella in poultry and poultry products." *Food Microbiology*, 38 250–262.

**Scallan, E., R. M. Hoekstra, F. J. Angulo, R. V. Tauxe, M.-A. Widdowson, S. L. Roy, J. L. Jones, and P. M. Griffin.** 2011. "Foodborne illness acquired in the United States—major pathogens." *Emerging Infectious Diseases*, 17(1): 7–15.

**Shewmake, Sharon, and W Kip Viscusi.** 2015. "Producer and consumer responses to green housing labels." *Economic Inquiry*, 53(1): 681–699.

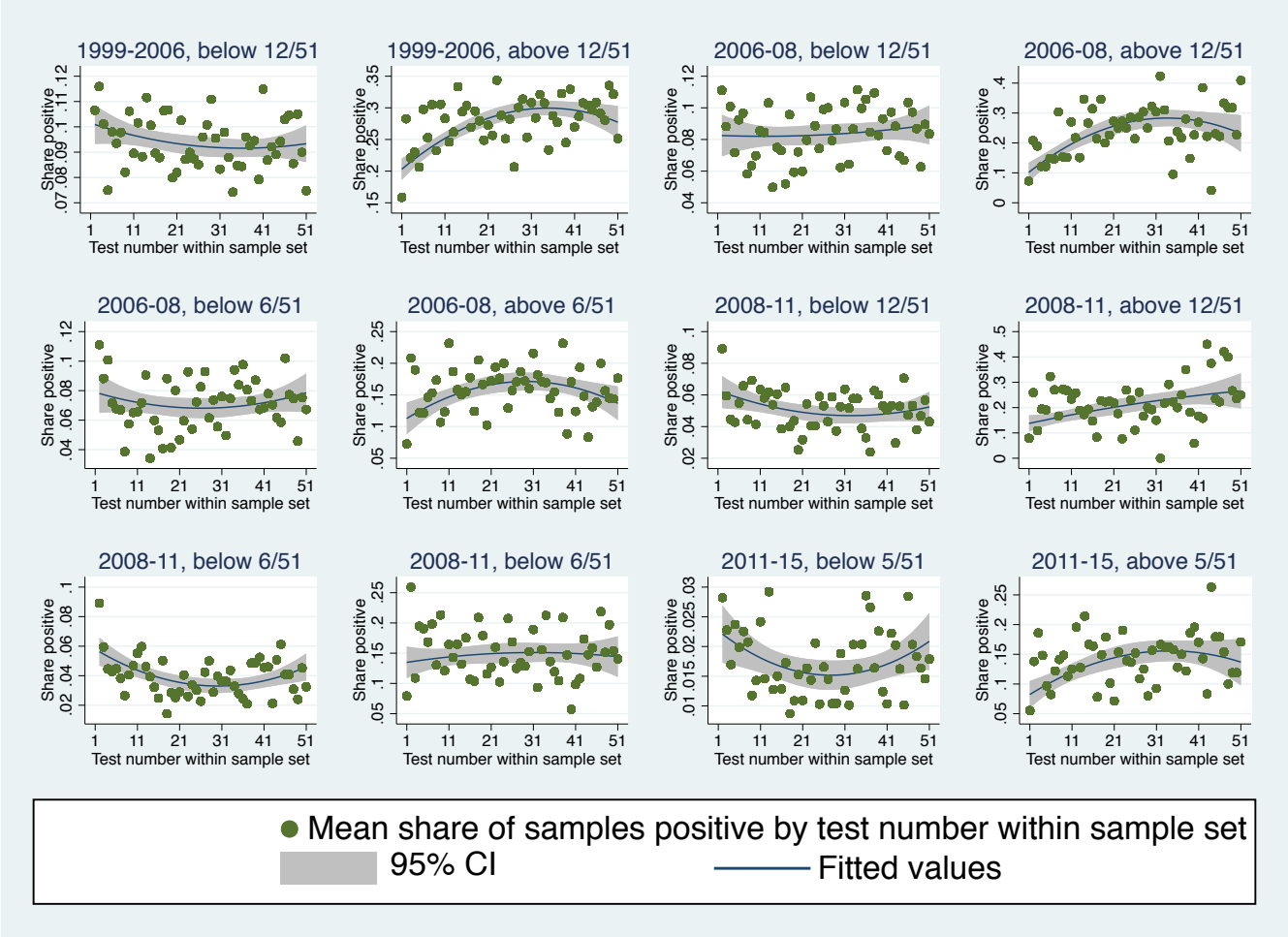
**USDA.** 2019. "2017 Census of Agriculture." https://www.nass.usda.gov/Publications/AgCensus/2017/.

Figure 1: Histograms of the number of positive samples per sample set, by policy period



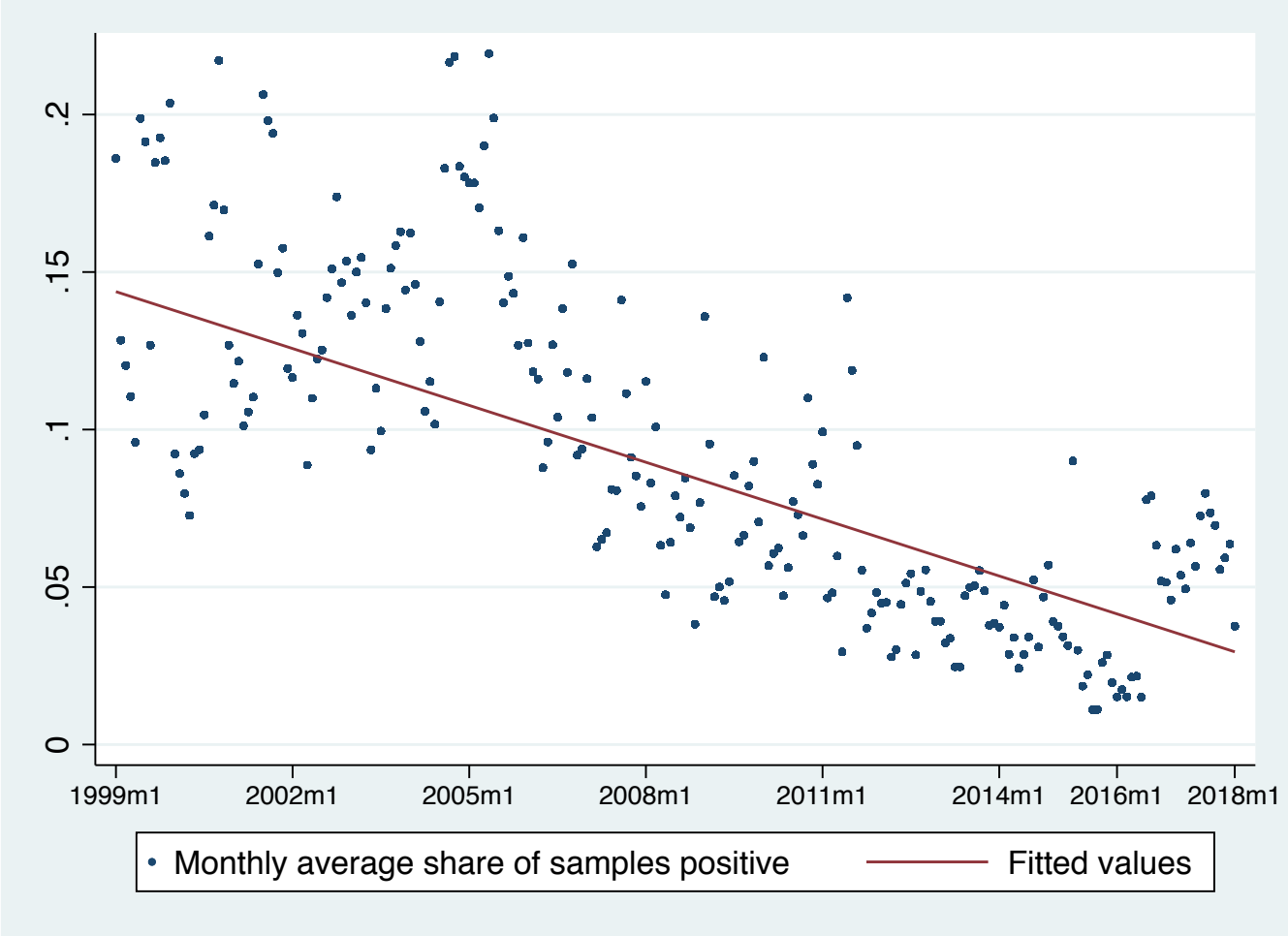

*Notes:* Each panel represents the density of the number of positive samples per 51-sample set, for each policy period. Vertical lines represent the regulatory threshold (until 2006) and the category thresholds (starting in 2006).
*Source:* Generated by the author using data from FSIS.

Figure 2: Mean share of samples positive, by test number within sample sets and by policy period



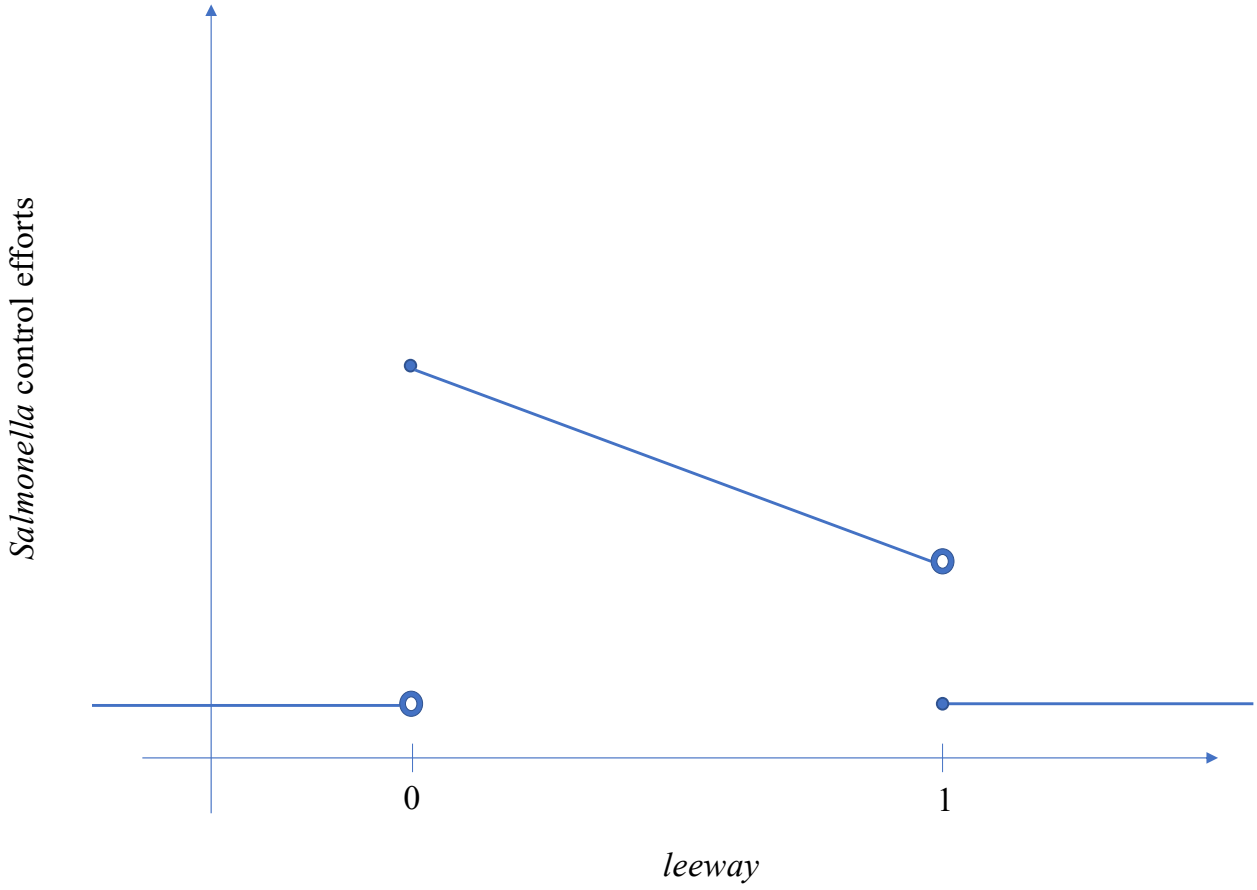

*Notes:* Each pair or quartet of graphs is broken down by establishments' current status relative to the relevant thresholds. That is, the lefthand graph in each pair shows test observations from establishments that had a lower percentage of tests positive than the threshold, in the current sample set at the time of the test; the righthand graph shows observations from establishments that had a higher percentage of tests positive than the threshold. Each dot represents the average share of samples positive for a given test number within sample sets. The solid curves are fitted quadratics, and the shading represents 95% confidence intervals.
*Source:* Generated by the author using data from FSIS.

Figure 3: Monthly average rate of *Salmonella* samples positive, with fitted OLS regression




*Note:* OLS regression is fitted to monthly average data.
*Source:* Generated by the author using data from FSIS.

Figure 4: Motivating the analysis of moral hazard




*Notes:* This figure is intended to explain the incentives for establishments to control *Salmonella* relative to the *leeway* variables. When $leeway \geq 1$, incentives to control *Salmonella* are weak, because the establishment may have 100% of remaining samples test positive and still be categorized the same way. When $leeway < 0$, incentives are also weak because even if none of the remaining samples test positive, the establishment will still fail to achieve the threshold associated with the better categorization. When $0 \leq leeway < 1$, incentives decrease with *leeway* because with more leeway, establishments may have a higher share of remaining samples test positive and still achieve the threshold associated with the better categorization.

Table 1: Summary statistics: FSIS *Salmonella* test outcomes by policy period

| FSIS policy regime | Dates in place | Share positive | Observations |
|---|---|---|---|
| Three-strikes system | 1/4/1999 to 5/29/2006 | 0.130 | 71,449 |
| Categories (privately reported) | 5/30/2006 to 3/27/2008 | 0.100 | 20,406 |
| Categories 2 and 3 published | 3/28/2008 to 6/30/2011 | 0.068 | 21,478 |
| Category 3 published, tighter standard | 7/1/2011 to 5/5/2015 | 0.042 | 35,083 |
| Categories assigned based on continuous sampling | 5/6/2015 to 1/25/2018 | 0.045 | 24,206 |

*Notes:* Average, unweighted, share of samples testing positive for *Salmonella* during each policy period. My data set begins on January 4, 1999 and ends on January 25, 2018; FSIS *Salmonella* testing began earlier and is still ongoing.

Table 2: Effects of known categorization on *Salmonella* outcomes

*Panel A: Cutoffs not associated with disclosure*

| Policy regime | No categorization | | Categorization (private) | | | | Public disclosure w/ tighter standards | | Rolling windows | |
|---|---|---|---|---|---|---|---|---|---|---|
| Years | 1999 to 2006 | | 2006 to 2008 | | | | 2011 to 2015 | | 2015 to 2018 | |
| RD cutoff ($c$) | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Max. # pos. samples ($C$) | 12 | 12 | 6 | 6 | 12 | 12 | 2 | 2 | 2 | 2 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| $LeewayC \geq c$ | 0.043 | −0.021 | −0.061 | 0.046 | −0.079 | 0.026 | −0.089 | 0.006 | −0.033 | 0.010 |
| Robust $p$-value | 0.656 | 0.709 | 0.000 | 0.373 | 0.001 | 0.412 | 0.005 | 0.467 | 0.002 | 0.385 |
| 95% CI (lower limit) | −0.11 | −0.24 | −0.21 | −0.09 | −0.54 | −0.06 | −0.14 | −0.01 | −0.06 | −0.01 |
| (upper limit) | 0.07 | 0.16 | −0.06 | 0.03 | −0.14 | 0.14 | −0.03 | 0.02 | −0.01 | 0.03 |
| Observations | 23969 | 5594 | 13520 | 14287 | 10397 | 2396 | 17174 | 19160 | 14595 | 10904 |
| Left bandwidth | 0.53 | 0.13 | 1.29 | 0.99 | 0.75 | 0.17 | 0.14 | 0.95 | 4.22 | 0.24 |
| Right bandwidth | 0.13 | 2.33 | 0.31 | 2.84 | 0.47 | 2.13 | 0.95 | 1.05 | 1.00 | 1.00 |

*Panel B: Cutoffs associated with disclosure*

| Policy regime | Public disclosure | | | | Public disclosure w/ tighter standards | | Rolling windows | |
|---|---|---|---|---|---|---|---|---|
| Years | 2008 to 2011 | | | | 2011 to 2015 | | 2015 to 2018 | |
| RD cutoff ($c$) | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Max. # pos. samples ($C$) | 6 | 6 | 12 | 12 | 5 | 5 | 5 | 5 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| $LeewayC \geq c$ | 0.024 | 0.015 | 0.155 | −0.024 | −0.013 | 0.001 | −0.031 | 0.022 |
| Robust $p$-value | 0.436 | 0.200 | 0.660 | 0.582 | 0.672 | 0.270 | 0.243 | 0.019 |
| 95% CI (lower limit) | −0.13 | −0.09 | −0.21 | −0.10 | −0.11 | −0.04 | −0.07 | 0.01 |
| (upper limit) | 0.06 | 0.02 | 0.33 | 0.05 | 0.07 | 0.01 | 0.02 | 0.08 |
| Observations | 11056 | 13813 | 9228 | 2495 | 8378 | 21453 | 9244 | 12129 |
| Left bandwidth | 0.80 | 0.29 | 0.24 | 0.48 | 0.66 | 1.00 | 3.17 | 1.00 |
| Right bandwidth | 1.00 | 2.45 | 0.20 | 1.90 | 0.14 | 2.23 | 0.39 | 1.83 |

*Notes:* Each pair or quartet of columns represents regressions using observations from the policy regimes beginning and ending in the indicated years. For sample sets that span the dates of policy change, all observations are included in the later policy period. All regressions are local linear RD regressions with triangular kernels, using *leewayC* as the running variable, as described in the text. Bandwidths, robust $p$-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo et al., 2020c), clustering on establishment using nearest-neighbor estimation for the variance-covariance estimator. Bandwidths are chosen to minimize mean squared error on either side of each cutoff. All regressions control for sample collection date, test number within sample set, and the share of samples positive in the establishment's prior sample set.

Table 3: Effects of proximity to category thresholds on *Salmonella* test outcomes, 1999–2015

| Policy regime | No categorization | | Categorization (private) | | Public disclosure | | Tightened standards | |
|---|---|---|---|---|---|---|---|---|
| Years | 1999 to 2006 | | 2006 to 2008 | | 2008 to 2011 | | 2011 to 2015 | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *Panel A* | | | | | | | | |
| *Leeway*6 (1999 to 2011) or *Leeway*2 (2011 to 2015) | 0.430*** | 0.358*** | 0.413*** | 0.339*** | 0.256*** | 0.228*** | 0.108*** | 0.0915*** |
| | (0.027) | (0.026) | (0.041) | (0.042) | (0.029) | (0.030) | (0.018) | (0.017) |
| Test number, current sample set | -0.000944** | -0.000768* | -0.00197*** | -0.00155** | -0.00226*** | -0.00221*** | -0.000286 | -0.000223 |
| | (0.00037) | (0.00041) | (0.00052) | (0.00061) | (0.00052) | (0.00058) | (0.00023) | (0.00025) |
| Share of samples positive, current sample set | | -0.718*** | | -0.915*** | | -0.712*** | | -0.623*** |
| | | (0.036) | | (0.092) | | (0.097) | | (0.083) |
| Observations | 49073 | 47868 | 15386 | 15056 | 15392 | 15051 | 23972 | 23448 |
| Elasticity | 0.70 | 0.59 | 0.90 | 0.74 | 0.97 | 0.86 | 0.50 | 0.43 |
| *Panel B* | | | | | | | | |
| *Leeway*12 (1999 to 2011) or *Leeway*5 (2011 to 2015) | 0.563*** | 0.488*** | 0.461*** | 0.382*** | 0.448*** | 0.411*** | 0.110*** | 0.0965*** |
| | (0.029) | (0.030) | (0.043) | (0.048) | (0.051) | (0.053) | (0.016) | (0.016) |
| Test number, current sample set | -0.00589*** | -0.00553*** | -0.00580*** | -0.00495*** | -0.00647*** | -0.00641*** | -0.00118*** | -0.00113*** |
| | (0.00049) | (0.00054) | (0.00073) | (0.00091) | (0.00092) | (0.0010) | (0.00030) | (0.00033) |
| Share of samples positive, current sample set | | -0.720*** | | -0.905*** | | -0.671*** | | -0.658*** |
| | | (0.036) | | (0.087) | | (0.092) | | (0.082) |
| Observations | 50796 | 49591 | 14652 | 14322 | 14381 | 14040 | 24086 | 23562 |
| Elasticity | 1.60 | 1.39 | 1.84 | 1.52 | 2.68 | 2.46 | 0.96 | 0.84 |

*Notes:* Panel A demonstrates the effects of proximity to the Category 1 thresholds (i.e., leeway) on *Salmonella* test outcomes; Panel B the effects of proximity to the Category 2 thresholds. Horizontally, each pair of columns represents regressions using observations from the policy regimes beginning and ending in the indicated years. For sample sets that span the dates of policy change, observations are included as part of the later policy period. All regressions use establishment–month–year fixed effects. Standard errors, clustered by establishment, are given in parentheses. Elasticities reported are the elasticities of the share of samples positive with respect to $leewayC$, calculated using the mean share of samples positive and the mean value of $leewayC$. Observations are included only if $leewayC \in [0, 1)$.

Table 4: Effects of proximity to category thresholds on *Salmonella* test outcomes, 2015–18

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| *Panel A* | | | | |
| *Leeway*2 (soonest-ending window) | 0.262*** | -0.0691 | 0.194*** | -0.0903 |
| | (0.066) | (0.055) | (0.065) | (0.055) |
| Share of samples positive, soonest-ending window | | | -3.031*** | -2.477*** |
| | | | (1.04) | (0.88) |
| Control for *Leeway*2 values for all other windows | No | Yes | No | Yes |
| Observations | 9877 | 9877 | 9694 | 9694 |
| Elasticity | 0.98 | -0.26 | 0.73 | -0.34 |
| *Panel B* | | | | |
| *Leeway*5 (soonest-ending window) | 0.483*** | 0.299*** | 0.409*** | 0.258** |
| | (0.070) | (0.10) | (0.069) | (0.10) |
| Share of samples positive, soonest-ending window | | | -2.654*** | -2.314*** |
| | | | (0.50) | (0.45) |
| Control for *Leeway*5 values for all other windows | No | Yes | No | Yes |
| Observations | 8160 | 8160 | 7977 | 7977 |
| Elasticity | 3.42 | 2.11 | 2.90 | 1.82 |

*Notes:* This table represents the results of similar regressions to those shown in table 3, for the 2015–18 policy period during which sample sets were replaced with overlapping sampling windows. Panel A demonstrates the effects of proximity to the Category 1 threshold (*leeway*2) on *Salmonella* test outcomes; Panel B the effects of proximity to the Category 2 threshold (*leeway*5). The main variables of interest are *leeway*2 and *leeway*5 for the soonest-ending window, but columns (2) and (4) also control for the *leeway* values for all other windows to which a given observation belongs. All regressions use establishment–month–year fixed effects. Standard errors, clustered by establishment, are given in parentheses. Elasticities reported are the elasticities of the share of samples positive with respect to *leewayC*, calculated using the mean share of samples positive and the mean value of *leewayC*. Observations are included only if $leewayC \in [0, 1)$.

Table 5: Effects of policy changes on average *Salmonella* test outcomes

| Policy introduced | Categorization (private) | Public disclosure | Public disclosure w/ tighter standards | Rolling windows |
|---|---|---|---|---|
| Date of implementation ($c$) | 5/30/2006 | 3/28/2008 | 7/1/2011 | 5/6/2015 |
| | (1) | (2) | (3) | (4) |
| *Panel A: All establishments included* | | | | |
| $t \geq c$ | 0.020 | −0.051 | 0.058 | −0.021 |
| Robust $p$-value | 0.506 | 0.008 | 0.108 | 0.195 |
| 95% CI (lower limit) | −0.04 | −0.10 | −0.02 | −0.06 |
| (upper limit) | 0.07 | −0.02 | 0.16 | 0.01 |
| Observations | 17230 | 8537 | 6271 | 13328 |
| Left bandwidth | 386 | 177 | 252 | 384 |
| Right bandwidth | 183 | 267 | 202 | 243 |
| *Panel B: Establishments that ever exited excluded* | | | | |
| $t \geq c$ | 0.031 | −0.048 | 0.068 | −0.022 |
| Robust $p$-value | 0.211 | 0.018 | 0.026 | 0.168 |
| 95% CI (lower limit) | −0.02 | −0.09 | 0.01 | −0.05 |
| (upper limit) | 0.10 | −0.01 | 0.15 | 0.01 |
| Observations | 16746 | 7912 | 5555 | 15139 |
| Left bandwidth | 371 | 194 | 204 | 492 |
| Right bandwidth | 265 | 271 | 232 | 248 |

*Notes:* This table reports the results of RD in time regressions that use the dates of policy implementation as the cutoffs ($c$). All regressions are local linear RD regressions with triangular kernels, using the sample collection date as the running variable, as described in the text. Bandwidths, robust $p$-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo et al., 2020c). Bandwidths are chosen to minimize mean squared error on either side of each cutoff.

Table 6: Heterogeneous effects of policy changes on average *Salmonella* test outcomes

| Average pre-period *Salmonella* test performance equivalent to | Category 1 (1) | Category 2 (2) | Category 3 (3) |
|---|---|---|---|
| *2008 policy change (c = 3/28/2008)* | | | |
| $t \geq c$ | −0.038 | −0.057 | −0.047 |
| Robust $p$-value | 0.028 | 0.159 | 0.737 |
| 95% CI (lower limit) | −0.08 | −0.13 | −0.30 |
| (upper limit) | −0.00 | 0.02 | 0.21 |
| Observations | 5222 | 2592 | 389 |
| Left bandwidth | 207 | 244 | 183 |
| Right bandwidth | 232 | 371 | 259 |
| *2011 policy change (c = 7/1/2011)* | | | |
| $t \geq c$ | 0.037 | −0.039 | 0.177 |
| Robust $p$-value | 0.275 | 0.081 | 0.030 |
| 95% CI (lower limit) | −0.04 | −0.10 | 0.02 |
| (upper limit) | 0.13 | 0.01 | 0.38 |
| Observations | 5549 | 3505 | 1632 |
| Left bandwidth | 266 | 210 | 240 |
| Right bandwidth | 487 | 358 | 222 |

*Notes:* This table reports the results of RD in time regressions that use the dates of policy implementation as the cutoffs ($c$). All regressions are local linear RD regressions with triangular kernels, using the sample collection date as the running variable, as described in the text. Bandwidths, robust $p$-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo et al., 2020c). For the 2008 policy change, column (1) uses observations from establishments with an average of no more than 11.8 percent positive samples (equivalent to $\leq 6/51$) during the 2006–08 period; column (2) uses observations from establishments with more than 11.8 percent but no more than 23.5 percent (equivalent to $\leq 12/51$) during 2006–08; column (3) uses observations from establishments with more than 23.5 percent positive samples. For the 2011 policy change, column (1) uses observations from establishments with an average of no more than 3.9 percent positive samples (equivalent to $\leq 2/51$) during the 2008–11 period; column (2) uses observations from establishments with more than 3.9 percent but no more than 9.8 percent (equivalent to $\leq 5/51$) during 2008–11; column (3) uses observations from establishments with more than 9.8 percent positive samples.

## Appendix A: Details on data-cleaning procedure

The data set I obtained from FSIS does not include any indication of the sample-set groupings that were used to determine regulatory compliance and category designation over 1999–2015, and FSIS did not provide further guidance on this issue. Inspection of the data reveals clear patterns of 51 samples being collected over a short period, followed by a gap (often, approximately one year) before another set of 51 samples. However, it is clear that inspectors often collected slightly more and occasionally slightly fewer than 51 samples. FSIS personnel confirmed that inspectors were supposed to collect samples until *results* from 51 tests were available, which explains the frequent appearance of 52 to 56 samples over a brief period, followed by a gap. FSIS also sometimes terminated collection before reaching 51 samples, if a threshold was certain to be exceeded. After some preliminary data cleaning to eliminate duplicate observations, I assign observations into sample sets by identifying lengthy gaps between observations while maximizing the number of sample sets with 51 observations. Specifically, I identify the start of a new sample set as occurring when the gap between observations was at least $x$ times as long as the average gap over the previous 51 observations, where $x$ is chosen for each policy period as the integer that maximizes the number of sample sets with 51 observations. This method generates sample sets with lengths reasonably close to the expected length: at least 80% of all sample sets in each of the regulatory periods have 50 to 56 observations. To eliminate noise that would be generated through mis-assigning observations to sample sets, for the main analysis of sections 3 and 4, I only include observations from sample sets of length $[n, ..., N]$, where $n$ and $N$ are the minimum and maximum sample-set lengths such that at least 1% of sample sets have lengths $n$ and $N$. Note again that the 51-sample sets were eliminated effective May 6, 2015.

## Appendix B: Validation and robustness tests

The following tables provide the results of various validation and robustness tests described in the text.

Table A1: Effects of known categorization on *Salmonella* outcomes

*Panel A: Cutoffs not associated with disclosure, quadratic polynomials, triangular kernels*

| Policy regime | No categorization | | Categorization (private) | | | | Public disclosure w/ tighter standards | | Rolling windows | |
|---|---|---|---|---|---|---|---|---|---|---|
| Years | 1999 to 2006 | | 2006 to 2008 | | | | 2011 to 2015 | | 2015 to 2018 | |
| RD cutoff ($c$) | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Max. # pos. samples ($C$) | 12 | 12 | 6 | 6 | 12 | 12 | 2 | 2 | 2 | 2 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| *LeewayC* $\geq c$ | 0.032 | −0.120 | −0.051 | −0.043 | −0.116 | 0.100 | −0.029 | −0.012 | −0.045 | −0.071 |
| Robust $p$-value | 0.547 | 0.108 | 0.023 | 0.383 | 0.129 | 0.127 | 0.279 | 0.434 | 0.010 | 0.298 |
| 95% CI (lower limit) | −0.18 | −0.29 | −0.15 | −0.07 | −0.33 | −0.03 | −0.07 | −0.04 | −0.08 | −0.21 |
| (upper limit) | 0.09 | 0.03 | −0.01 | 0.17 | 0.04 | 0.25 | 0.02 | 0.02 | −0.01 | 0.06 |
| Observations | 34800 | 8041 | 14501 | 14364 | 8942 | 3521 | 18821 | 19160 | 14737 | 10904 |
| Left bandwidth | 1.54 | 0.51 | 2.69 | 1.00 | 1.72 | 0.39 | 0.47 | 0.95 | 4.35 | 1.00 |
| Right bandwidth | 0.26 | 5.17 | 0.44 | 4.13 | 0.31 | 5.66 | 0.18 | 1.05 | 0.29 | 1.00 |

*Panel B: Cutoffs associated with disclosure, quadratic polynomials, triangular kernels*

| Policy regime | Public disclosure | | | | Public disclosure w/ tighter standards | | Rolling windows | |
|---|---|---|---|---|---|---|---|---|
| Years | 2008 to 2011 | | | | 2011 to 2015 | | 2015 to 2018 | |
| RD cutoff ($c$) | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Max. # pos. samples ($C$) | 6 | 6 | 12 | 12 | 5 | 5 | 5 | 5 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *LeewayC* $\geq c$ | 0.040 | −0.046 | 0.206 | −0.004 | −0.023 | −0.027 | −0.033 | 0.039 |
| Robust $p$-value | 0.784 | 0.175 | 0.258 | 0.855 | 0.254 | 0.161 | 0.207 | 0.000 |
| 95% CI (lower limit) | −0.09 | −0.03 | −0.15 | −0.09 | −0.20 | −0.01 | −0.08 | 0.08 |
| (upper limit) | 0.12 | 0.17 | 0.55 | 0.10 | 0.05 | 0.08 | 0.02 | 0.23 |
| Observations | 12765 | 13850 | 9237 | 3635 | 16802 | 21453 | 10426 | 17150 |
| Left bandwidth | 1.26 | 1.00 | 0.56 | 0.31 | 1.14 | 1.00 | 6.09 | 1.00 |
| Right bandwidth | 0.49 | 3.73 | 0.49 | 4.83 | 0.32 | 2.44 | 0.66 | 3.78 |

*Notes:* See notes to table 2.

Table A2: Effects of known categorization on *Salmonella* outcomes

*Panel A: Cutoffs not associated with disclosure, linear polynomials, Epanechnikov kernels*

| Policy regime | No categorization | | Categorization (private) | | | | Public disclosure w/ tighter standards | | Rolling windows | |
|---|---|---|---|---|---|---|---|---|---|---|
| Years | 1999 to 2006 | | 2006 to 2008 | | | | 2011 to 2015 | | 2015 to 2018 | |
| RD cutoff ($c$) | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Max. # pos. samples ($C$) | 12 | 12 | 6 | 6 | 12 | 12 | 2 | 2 | 2 | 2 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| $LeewayC \geq c$ | 0.022 | −0.021 | −0.058 | 0.049 | −0.056 | 0.019 | −0.030 | 0.006 | −0.033 | 0.011 |
| Robust $p$-value | 0.000 | 0.738 | 0.000 | 0.233 | 0.512 | 0.489 | 0.269 | 0.425 | 0.002 | 0.310 |
| 95% CI (lower limit) | −0.28 | −0.28 | −0.24 | −0.10 | −0.23 | −0.06 | −0.06 | −0.01 | −0.06 | −0.01 |
| (upper limit) | −0.09 | 0.20 | −0.08 | 0.02 | 0.12 | 0.13 | 0.02 | 0.02 | −0.01 | 0.03 |
| Observations | 34866 | 5509 | 13467 | 14251 | 4240 | 2276 | 17439 | 19160 | 14111 | 10904 |
| Left bandwidth | 0.58 | 0.12 | 0.88 | 0.97 | 0.78 | 0.17 | 0.28 | 0.95 | 3.74 | 1.00 |
| Right bandwidth | 0.53 | 2.21 | 0.32 | 2.67 | 0.27 | 1.82 | 0.13 | 1.05 | 0.21 | 1.00 |

*Panel B: Cutoffs associated with disclosure, linear polynomials, Epanechnikov kernels*

| Policy regime | Public disclosure | | | | Public disclosure w/ tighter standards | | Rolling windows | |
|---|---|---|---|---|---|---|---|---|
| Years | 2008 to 2011 | | | | 2011 to 2015 | | 2015 to 2018 | |
| RD cutoff ($c$) | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Max. # pos. samples ($C$) | 6 | 6 | 12 | 12 | 5 | 5 | 5 | 5 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| $LeewayC \geq c$ | 0.029 | 0.019 | 0.174 | −0.023 | −0.028 | 0.002 | −0.028 | 0.020 |
| Robust $p$-value | 0.479 | 0.108 | 0.507 | 0.360 | 0.393 | 0.187 | 0.284 | 0.085 |
| 95% CI (lower limit) | −0.12 | −0.09 | −0.19 | −0.09 | −0.14 | −0.05 | −0.07 | −0.00 |
| (upper limit) | 0.05 | 0.01 | 0.38 | 0.03 | 0.05 | 0.01 | 0.02 | 0.07 |
| Observations | 10639 | 13813 | 8236 | 2589 | 8817 | 21453 | 8995 | 12129 |
| Left bandwidth | 0.80 | 0.26 | 0.24 | 0.22 | 0.61 | 1.00 | 2.76 | 1.00 |
| Right bandwidth | 1.00 | 2.33 | 0.22 | 1.62 | 0.14 | 2.07 | 0.37 | 1.70 |

*Notes:* See notes to table 2.

Table A3: Placebo effects of known categorization on *Salmonella* outcomes

| RD cutoff ($c$) | $-0.15$ | $-0.1$ | $-0.05$ | 0.05 | 0.1 | 0.15 |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A: 2006 to 2008, $C = 6$ positive samples* | | | | | | |
| $LeewayC \geq c$ | 0.017 | 0.005 | $-0.165$ | 0.004 | 0.038 | $-0.030$ |
| Robust $p$-value | 0.629 | 0.328 | 0.232 | 0.903 | 0.099 | 0.305 |
| *Panel B: 2006 to 2008, $C = 12$ positive samples* | | | | | | |
| $LeewayC \geq c$ | $-0.001$ | 0.088 | 0.157 | 0.046 | 0.177 | 0.090 |
| Robust $p$-value | 0.811 | 0.887 | 0.698 | 0.436 | 0.039 | 0.110 |
| *Panel C: 2011 to 2015, $C = 2$ positive samples* | | | | | | |
| $LeewayC \geq c$ | $-0.090$ | 0.024 | $-0.004$ | 0.001 | 0.011 | 0.139 |
| Robust $p$-value | 0.074 | 0.364 | 0.711 | 0.702 | 0.015 | 0.035 |
| *Panel D: 2015 to 2018, $C = 2$ positive samples* | | | | | | |
| $LeewayC \geq c$ | 0.019 | 0.014 | $-0.119$ | $-0.065$ | 0.016 | $-0.011$ |
| Robust $p$-value | 0.752 | 0.747 | 0.515 | 0.199 | 0.200 | 0.312 |
| RD cutoff ($c$) | 0.7 | 0.75 | 0.8 | 1.3 | 1.35 | 1.4 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel E: 2015 to 2018, $C = 5$ positive samples* | | | | | | |
| $LeewayC \geq c$ | 0.011 | 0.008 | 0.001 | 0.039 | 0.022 | 0.013 |
| Robust $p$-value | 0.938 | 0.867 | 0.742 | 0.001 | 0.328 | 0.601 |

*Notes:* This table presents results of regressions paralleling those in table 2 with statistically significant results but for placebo cutoffs not associated with any change in disclosure status. Panels A through D use three cutoffs on either side of the actual significant cutoff ($c = 0$) according to $c \pm 0.05n$, where $n = \{1, 2, 3\}$. Panel E uses the nearest cutoffs to the actual significant cutoff ($c = 1$) that are multiples of 0.05, for which there are enough observations on either side of the placebo cutoffs to estimate the optimal bandwidths around $c$. Each panel represents regressions using observations from the policy regimes beginning and ending in the indicated years. For sample sets that span the dates of policy change, all observations are included in the later policy period. All regressions are local linear RD regressions with triangular kernels, using *leewayC* as the running variable. Bandwidths, robust $p$-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo et al., 2020c), although bandwidths and confidence intervals are suppressed in this table. Bandwidths are chosen to minimize mean squared error on either side of each cutoff. All regressions in panels A through D control for sample collection date, test number within sample set, and the share of samples positive in the establishment's prior sample set. All regressions in panel E control for sample collection date.

Table A4: Effects of known categorization on *Salmonella* outcomes: 2015–18, conditional on Category 1 or 2 status in sooner-ending windows

| Window # (relative to soonest-ending window = 1) | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| $LeewayC \geq 1$ | 0.022 | 0.051 | 0.008 | 0.029 | 0.015 | 0.017 | −0.007 | 0.003 | 0.066 | −0.170 |
| Robust $p$-value | 0.019 | 0.000 | 0.017 | 0.177 | 0.518 | 0.062 | 0.771 | 0.400 | 0.148 | 0.114 |
| 95% CI (lower limit) | 0.01 | 0.07 | 0.02 | −0.02 | −0.09 | −0.00 | −0.08 | −0.11 | −0.07 | −0.95 |
| (upper limit) | 0.08 | 0.23 | 0.16 | 0.14 | 0.05 | 0.10 | 0.10 | 0.29 | 0.46 | 0.10 |
| Observations | 12129 | 12193 | 11663 | 9949 | 7281 | 3379 | 1117 | 593 | 429 | 466 |
| Left bandwidth | 1.00 | 1.00 | 0.88 | 0.94 | 0.88 | 1.00 | 1.00 | 0.94 | 0.59 | 0.49 |
| Right bandwidth | 1.83 | 2.33 | 2.56 | 2.18 | 3.29 | 3.23 | 2.84 | 2.35 | 1.60 | 1.47 |

*Notes:* This table demonstrates that, conditioning on being certain to avoid public disclosure (i.e., conditioning on being below the Category 3 threshold) based on the soonest-ending window, avoiding public disclosure based on the second-soonest-ending window increased the likelihood of a positive sample by 5.1 percentage points. Conditioning on being certain to avoid public disclosure in the second-soonest-ending window, avoiding public disclosure based on the third-soonest-ending window increased the likelihood of a positive sample by 0.8 percentage points. The RD coefficients in the remaining columns—including all windows for which for which there are enough observations to perform the `rdrobust` calculations—have $p > 0.06$. That is, conditioning on the third-soonest-ending window, fourth-soonest-ending window, and so on, avoiding public disclosure further in the future did not have a discontinuous effect on the likelihood of additional positive samples. For additional details on the regression specifications, see notes to table 2.

Table A5: Effects of policy changes on average *Salmonella* test outcomes: Robustness tests

| Policy introduced | Categorization (private) | Public disclosure | Public disclosure w/ tighter standards | Rolling windows |
|---|---|---|---|---|
| Date of implementation ($c$) | 5/30/2006 | 3/28/2008 | 7/1/2011 | 5/6/2015 |
| | (1) | (2) | (3) | (4) |
| *Panel A: Observations collapsed by sample collection date* | | | | |
| $t \geq c$ | 0.021 | −0.043 | 0.064 | −0.024 |
| Robust $p$-value | 0.270 | 0.011 | 0.003 | 0.096 |
| 95% CI (lower limit) | −0.02 | −0.08 | 0.02 | −0.05 |
| (upper limit) | 0.06 | −0.01 | 0.11 | 0.00 |
| Observations | 359 | 367 | 403 | 477 |
| Left bandwidth | 385 | 273 | 413 | 401 |
| Right bandwidth | 132 | 263 | 167 | 212 |
| *Panel B: "Donut" approach: Drop all observations within 7 days of policy changes* | | | | |
| $t \geq c$ | 0.024 | −0.057 | 0.055 | −0.025 |
| Robust $p$-value | 0.294 | 0.039 | 0.088 | 0.096 |
| 95% CI (lower limit) | −0.03 | −0.11 | −0.01 | −0.06 |
| (upper limit) | 0.10 | −0.00 | 0.14 | 0.00 |
| Observations | 15236 | 7183 | 5414 | 15539 |
| Left bandwidth | 366 | 204 | 199 | 512 |
| Right bandwidth | 220 | 237 | 234 | 256 |
| *Panel C: Drop all observations belonging to sample sets that span policy periods* | | | | |
| $t \geq c$ | 0.023 | −0.054 | 0.127 | −0.019 |
| Robust $p$-value | 0.386 | 0.125 | 0.003 | 0.225 |
| 95% CI (lower limit) | −0.05 | −0.13 | 0.05 | −0.05 |
| (upper limit) | 0.12 | 0.02 | 0.24 | 0.01 |
| Observations | 12453 | 4244 | 3073 | 14312 |
| Left bandwidth | 342 | 184 | 160 | 456 |
| Right bandwidth | 264 | 196 | 248 | 247 |

*Notes:* See notes to table 5.

Table A6: Effects of policy changes on average *Salmonella* test outcomes: Placebo cutoff dates

| Policy introduced | Categorization (private) (1) | Public disclosure (2) | Public disclosure w/ tighter standards (3) | Rolling windows (4) |
|---|---|---|---|---|
| *Panel A: Cutoffs $c$ = Federal Register announcement dates* | | | | |
| $t \geq c$ | −0.032 | −0.029 | 0.014 | −0.014 |
| Robust $p$-value | 0.622 | 0.859 | 0.263 | 0.617 |
| 95% CI (lower limit) | −0.13 | −0.09 | −0.01 | −0.04 |
| (upper limit) | 0.08 | 0.07 | 0.05 | 0.02 |
| Observations | 9527 | 3326 | 5731 | 10789 |
| Left bandwidth | 354 | 176 | 165 | 440 |
| Right bandwidth | 89 | 60 | 138 | 99 |
| *Panel B: Cutoffs $c$ = 120 days before policy changes* | | | | |
| $t \geq c$ | −0.013 | 0.022 | −0.012 | −0.007 |
| Robust $p$-value | 0.891 | 0.329 | 0.850 | 0.929 |
| 95% CI (lower limit) | −0.06 | −0.04 | −0.04 | −0.03 |
| (upper limit) | 0.07 | 0.11 | 0.04 | 0.03 |
| Observations | 11503 | 2663 | 3006 | 9747 |
| Left bandwidth | 414 | 145 | 180 | 361 |
| Right bandwidth | 117 | 37 | 120 | 117 |
| *Panel C: Cutoffs $c$ = 240 days before policy changes* | | | | |
| $t \geq c$ | −0.018 | 0.053 | 0.002 | −0.010 |
| Robust $p$-value | 0.187 | 0.236 | 0.925 | 0.595 |
| 95% CI (lower limit) | −0.07 | −0.04 | −0.09 | −0.06 |
| (upper limit) | 0.01 | 0.16 | 0.08 | 0.04 |
| Observations | 27891 | 4213 | 7277 | 9093 |
| Left bandwidth | 1190 | 96 | 294 | 176 |
| Right bandwidth | 237 | 104 | 113 | 216 |
| *Panel D: Cutoffs $c$ = 360 days before policy changes* | | | | |
| $t \geq c$ | 0.019 | 0.029 | 0.030 | −0.006 |
| Robust $p$-value | 0.573 | 0.094 | 0.111 | 0.547 |
| 95% CI (lower limit) | −0.08 | −0.01 | −0.01 | −0.03 |
| (upper limit) | 0.15 | 0.08 | 0.08 | 0.02 |
| Observations | 15410 | 4474 | 6121 | 8271 |
| Left bandwidth | 523 | 68 | 235 | 232 |
| Right bandwidth | 146 | 76 | 112 | 115 |
| *Panel E: Cutoffs $c$ = 480 days before policy changes* | | | | |
| $t \geq c$ | −0.026 | 0.032 | −0.006 | −0.008 |
| Robust $p$-value | 0.226 | 0.204 | 0.988 | 0.726 |
| 95% CI (lower limit) | −0.11 | −0.02 | −0.06 | −0.04 |
| (upper limit) | 0.03 | 0.08 | 0.06 | 0.03 |
| Observations | 13538 | 5298 | 6117 | 12195 |
| Left bandwidth | 545 | 66 | 238 | 316 |
| Right bandwidth | 160 | 130 | 108 | 161 |

*Notes:* Panels A and B use includes a lagged dependent variable, because a Wooldridge test for autocorrelation in panel data suggests autocorrelation in the left bandwidth of the third placebo cutoff (see Hausman and Rapson, 2018, for discussion). For additional details on the regression specifications, see notes to table 5.