Shaming, stringency, and shirking: Evidence from food-safety inspections

John Bovay*

October 12, 2022

Abstract

This paper examines the responses of chicken producers to public disclosure of discrete quality information (or categorization) regarding *Salmonella* in chicken carcasses. I demonstrate that producers exert effort to attain better categorization and shirk when failing to meet the required thresholds. Public disclosure mitigates this shirking effect. However, some producers shirk even under public disclosure when the threshold for disclosure is too stringent. The results suggest that the most effective quality disclosure policies would either disclose continuous (non-categorical) information or impose fines or other sanctions on producers attaining the poorest quality.

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

This paper has benefitted from thoughtful reading and comments by Mary Evans, Peyton Ferrier, Mike Ollinger, Will Townes, Wei Zhang, and Chen Zhen, and helpful discussions with Aaron Adalja, Elinor Benami, Josh Beverly, Anne Byrne, Katie Harris-Lagoudakis, Robert Heilmayr, Pierre Mérel, Ariel Ortiz-Bobea, Brad Rickard, Ben Scharadin, Jay Shimshack, Mike Williams, and seminar participants at Auburn University, USDA, the 2020 SEA meeting, the 2021 ASSA meeting, and the 2021 SAEA meeting.

Shaming, stringency, and shirking: Evidence from food-safety inspections

Moral hazard is common in consumer product settings whenever producers have more information about the quality of their products than consumers do. Regulators have responded to this market failure through various regulatory approaches including direct regulation of product quality (e.g., through FDA's drug approval process) and indirect solutions like information disclosure (e.g., FTC's energy efficiency labeling requirements). Information disclosure regulations might require the provision of either continuous or discrete information about product quality. Discrete quality information (e.g., traffic-light labels) might be more easily understood by consumers but may also discourage producers from attaining quality scores that greatly surpass the thresholds associated with each labeled category (Shewmake and Viscusi, 2015; Ito and Sallee, 2018). Furthermore, if producers see thresholds as unattainable, they may make very little effort to improve along the relevant quality dimensions. Thus, in designing an information disclosure requirement, regulators face a tradeoff between eliminating the moral hazard stemming from the information asymmetry and providing actionable information to consumers.

In this paper, I explore a unique context in which producers faced a series of regulatory regimes targeting product quality through mandatory disclosure of discrete quality ratings, a type of policy sometimes referred to as "naming and shaming". The context is a series of three regulatory changes undertaken by the U.S. Department of Agriculture (USDA) regarding disclosure of information about *Salmonella* in chicken carcasses at slaughter establishments.[1] This paper documents the effects of categorization, publication of information about categories, and a later tightening of categorization and disclosure criteria on outcomes of tests for *Salmonella*.

My results provide evidence that establishments respond to incentives created by the categorization and disclosure program, sometimes by shirking or attaining worse food-safety outcomes. When an establishment exceeds the maximum number of positive samples required to attain a given performance category, its subsequent test performance worsens. However, this shirking effect is mitigated by the publication of information about test results. Furthermore, when establishments have more leeway with respect to the thresholds, their performance on *Salmonella* tests worsens. I also show that a tightening of standards in 2011 had a bifurcating effect wherein establishments with middling performance (prior to the policy change) tended to improve while establishments that performed poorly (prior to the policy change) tended to perform even worse, suggesting another form of shirking.

This paper 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

[1] *Salmonella* is a genus of bacteria typically present in the intestines of birds and other animals. Meat and poultry can become contaminated with *Salmonella* during slaughter if they come into contact with feces or the digestive tract or through cross contamination (Rasschaert et al., 2008). *Salmonella* in poultry is a major cause of food-borne illness in the United States, with an economic cost of up to $3.6 billion per year according to the following estimates and calculations. Hoffmann et al. (2015) report that *Salmonella* is the pathogen with the greatest economic cost of associated food-borne illnesses, causing up to $9.49 billion (in 2013 dollars) in losses from illnesses, hospitalizations, and deaths per year (at the upper end of the authors' 90% credible interval). Painter et al. (2013) estimate that 10.1 to 29.2% of the cases of illness caused by *Salmonella enterica* are attributed to poultry. $.292 \times \$9.49$ billion $= \$2.77$ billion in 2013 dollars, or $3.56 billion in July 2022 dollars. Scharff (2020) provides a similar estimate.

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, 2009; Bederson et al., 2018; Dai and Luca, 2020), drinking water (Bennear and Olmstead, 2008), and workplace safety violations (Johnson, 2020). Ollinger and Bovay (2020) find that, in the same context as this paper, public disclosure improved *Salmonella* test results. In section 5, I confirm the earlier finding but also show that a later tightening of disclosure standards resulted in worse average *Salmonella* test results, a result driven by the worst-performing establishments.

The design of the USDA *Salmonella* Verification Testing Program generates incentives for establishment operators to reduce effort around *Salmonella* control, i.e., to shirk. Under this program, USDA Food Safety and Inspection Service (FSIS) inspectors randomly pull chicken carcasses off the processing line to test them for *Salmonella*. From 2006 to 2015, if establishments exceeded certain numbers of positive samples within a "sample set" of 51 carcasses sampled over 51 consecutive operating days, they were designated "Category 2" or "Category 3". Under some policy regimes, Category 2 and 3 establishments have been listed on a public USDA website. Compared with a regime with disclosure of continuous information about *Salmonella* test results, this creates clear incentives for moral hazard, specifically a reduction in effort around controlling *Salmonella*. Under the discrete threshold disclosure system based on sample sets, 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).

This paper employs carcass-level data on *Salmonella* test results over 1999–2017 for all federally inspected chicken-slaughter establishments to test hypotheses about shaming and moral hazard.[2] 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, suggesting that establishment operators reduce effort related to controlling *Salmonella*. (2) When establishments fail to meet thresholds and are therefore subjected to public disclosure, there is no statistically significant change in *Salmonella* test performance. In other words, the shirking effect appears to be mitigated under public disclosure. There is little evidence that establishments shirked after

[2]For ease of exposition, data from May 2015 through December 2017 are analyzed only in Appendix C.

sustained good test performance ensured they would avoid public disclosure.

If chicken buyers have information about establishments' categorization status but not more detailed information about the results of FSIS *Salmonella* inspections, one would expect operators to exert effort to avoid poor categorization, but then shirk after poor performance makes the better categories unattainable. However, the apparent shirking is most widespread when categorization was not publicly disclosed; public disclosure mostly seems to have mitigated this type of shirking behavior. One potential explanation is that private buyers would have demanded additional information about the results of *Salmonella* inspections. That is, some buyers may have required that establishments provide documentation of their category status before purchasing chicken, even before category status was publicly disclosed by FSIS. Once FSIS publicly disclosed the categorical information, perhaps some buyers demanded additional, detailed information on inspection results that would have limited establishment operators' incentives to shirk.

Second, I document that when establishments have more leeway with respect to the thresholds, their performance on *Salmonella* tests worsens. 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 each policy change on average *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. A tightening of both categorization and disclosure standards in July 2011 had a bifurcating effect. Establishments that performed poorly prior to July 2011 tended to perform even worse after the tightening of standards. The results suggest a fourth type of moral hazard outcome not related to current performance with respect to the thresholds. Instead, it appears that some establishment operators exerted little effort to achieve the tighter thresholds, given their history of test performance. On the other hand, middling establishments for which the thresholds might have been more easily achievable responded to the incentives by improving performance. 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 remains relevant in legislation and policymaking today. In July 2020, bills were introduced into both chambers of the U.S. Congress to limit line speeds in chicken-slaughter establishments. According to some lawmakers, increased line speeds have negative implications for both worker health and food safety.[3] 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. In addition, in October 2021, FSIS formally announced a program to investigate future regulatory actions with the goal of reducing *Salmonella* in poultry by 25%.[4]

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 demon-

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

[4]See https://www.usda.gov/media/press-releases/2021/10/19/usda-launches-new-effort-reduce-salmonella-illnesses-linked-poultry.

strates 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 additional validation and robustness tests, and describe results on shaming and shirking for two additional policy regimes that were in place over 2015–2017.

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 226 federally inspected chicken-slaughter establishments.[5] 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. Buyers of chicken from chicken-slaughter establishments typically include grocery retail chains and restaurants, or distributors from whom retailers and restaurants buy. Often, chicken-slaughter establishments will produce chicken that retail consumers see as any of several different brands, including store brands.[6]

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. Starting in 2006, 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 figure 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 Category 1. The new category designations were conveyed to firms privately until March 28, 2008, when

[5]During the period covered in this paper (1999 to 2017), there were 301 federally inspected chicken-slaughter establishments, but 75 of these exited the industry or opted for state inspection during the period.

[6]For example, in 2014 the Foster Farms establishment located in Livingston, California produced chicken products for the FoodMaxx, Kroger, Safeway, Savemart, Sunland, and Valbest brands, in addition to the Foster Farms brand. See https://www.fsis.usda.gov/sites/default/files/import/Foster-Farms-recalled-products.pdf.

the names and locations of Category 2 and 3 establishments were posted publicly on the FSIS website.[7] 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. This standard remained in place through May 5, 2015.

As seen in figure 2, 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 changes in technology and buyer requirements for food safety were taking place concurrently with FSIS policy changes (Park et al., 2014; Page, 2018), a careful empirical approach is needed to identify the effects of disclosure policies on producer behavior with respect to *Salmonella* control.

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 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.[8]

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 3, I plot histograms of the number of positive samples per sample set for each of the four policy periods over 1999–2015. Establish-

[7]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.

[8]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.

ment 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.[9] 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 with one or two positive samples fewer than the threshold than with one or two positive samples more than 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 2/3 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.

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

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 after either (1) too many positive samples result in crossing a threshold into a worse category (Category 2 or 3) or (2) sufficiently many negative samples ensure a better categorization outcome (Category 1 or 2). Effects of crossing thresholds are analyzed separately for each policy regime because under each policy regime, establishment operators faced somewhat different incentives related to controlling *Salmonella*. In particular, the information that would be disclosed upon exceeding the 5-, 6-, and 12-positive-sample thresholds varied under the various policy regimes.

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

[9] An FSIS policy in place since 1998 states that inspectors must select a random chill tank, a random time, and a predetermined location for collecting the carcass samples, then identify a carcass at that location, then count five carcasses back or ahead, and collect that sixth carcass for sampling. See https://www.fsis.usda.gov/sites/default/files/media_file/2021-02/Salmonella_Analysis.pdf.

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 $[1, i-1]$.[10] 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 samples would not have affected categorization.[11]

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), $D_{0ijk} = \mathbf{1}\left\{leewayC_{ijk} \geq 0\right\}$, $D_{1ijk} = \mathbf{1}\left\{leewayC_{ijk} \geq 1\right\}$, $f(\cdot)$ 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 ε_{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.

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

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

3.2. Results: Effects of known categorization on Salmonella test outcomes

The estimates from the RD models strongly suggest that establishment operators relaxed efforts around *Salmonella* control when categorization outcomes were known to establishments but when the categorization would not result in disclosure. However, in most periods, when an establishment's categorization would result in disclosure, the shirking effect was not evident.

Panel A of table 1 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 are less likely when $leewayC \in [0, 1)$ than when $leewayC < 0$; positive coefficients on the $leewayC = 1$ cutoffs imply that positive test results are more likely when $leewayC \geq 1$ than when $leewayC \in [0, 1)$. Interpretations of specific results in table 1 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 1, panel A, columns 3 and 5). The sharp effects of crossing these thresholds suggests that operators exerted effort to stay below the thresholds and then substantially reduced effort once the thresholds were exceeded.

During the 2008–11 policy period, the names of both Category 2 and 3 establishments were posted on the FSIS website. The results in table 1, panel B, columns 1–4, show that the cutoff values of $leeway6$ and $leeway12$ had statistically insignificant effects on subsequent *Salmonella* test performance.[12]

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 six 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 1, panel A, column 7). So, similar to the 2006–08 period, establishment operators apparently exerted effort to attain Category 1 but relaxed after failing to attain that standard, despite categorization status not being published for Category 1 and 2 establishments. But as in the 2008–11 period, the cutoff values associated with the Category 2/3 (disclosure) threshold did not have statistically significant effects on *Salmonella* test performance (table 1, panel B, columns 5–6).

[12]The insignificant effects are robust to the polynomial and kernel choices, as seen in appendix tables B1 and B2.

I now summarize the results in table 1. First, when establishments fail to meet thresholds but are not subject to public disclosure, *Salmonella* test performance typically worsens (panel A, columns 3, 5, and 7). Second, in most cases, when establishments fail to meet thresholds that subject them to public disclosure, there is no statistically significant change in *Salmonella* test performance (panel B, columns 1, 3, and 5). Third, establishment operators do not relax efforts after sustained good performance on *Salmonella* tests ensures they will avoid public disclosure (panel B, columns 2, 4, and 6).

These results suggest strongly that before public disclosure was implemented, establishment operators paid attention to the thresholds and exerted effort to achieve better categorization, and then shirked after failing to achieve the targeted thresholds. Yet, after the thresholds began to trigger public disclosure, there was no statistically significant evidence of shirking. One possible explanation for these outcomes is that buyers may have demanded additional information about *Salmonella* test results, beyond what was publicly disclosed. Perhaps prior to public disclosure, buyers demanded information about suppliers' categorization, and perhaps after categorization was publicly disclosed, buyers began to seek additional information or give suppliers extra scrutiny that limited the moral hazard incentives. Whatever the reason, the shirking outcomes seen before public disclosure were not evident after the introduction of disclosure.

3.3. Validity of the RD design and robustness tests

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. Neither of these tests are appropriate in my setting because of unique features of the data, described below.

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 (i.e., 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 samples positive in each of the four policy periods. 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.[13] 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 $leewayC \in [0, 1]$,

[13]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.

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), it is unlikely that manipulation is done on the basis of predetermined covariates.[14] Furthermore, visual examination of the histograms of the number of positive samples per completed sample set in figure 3 suggests that manipulation through post-test fraud 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.[15]

Cattaneo et al. (2020b) recommend using local linear regressions in the running variable and triangular kernels as in table 1, but some readers may be interested in seeing whether the RD results are robust to alternative specifications. Because the alternative specifications are not the recommended best-practice models, emphasis should be placed on where the various models reach similar conclusions rather than on where they diverge. Appendix tables B1 and B2 use quadratic polynomials and Epanechnikov kernels, respectively, but are otherwise identical to the specifications in table 1. One consistent conclusion emerges from all three specifications: establishments' test results worsened after they failed to meet the Category 1 standard in 2006–08, before the introduction of public disclosure. This is a clear example of moral hazard or shirking, and public disclosure of category information appears to have reduced shirking.

In appendix table B3, I present results for regressions parallel to those in table 1 but using placebo cutoff values for the running variables ($leewayC$). The time periods and thresholds shown here represent

[14]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. Hence, neither the numerator nor the denominator of the running variable can be (precisely) controlled.

[15]In a private and candid conversation, an FSIS employee told me they did not believe establishment managers or FSIS inspectors would have been able to fraudulently manipulate test results or select individual "clean" carcasses for inspection. See also footnote 9.

the statistically significant estimates from table 1. The placebo cutoff values are three multiples of 0.05 in either direction from $c = 0$. In appendix table B3, four of the 18 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 estimates for $c = 0$ in table 1 and the expectations about incentives for shirking that motivate the analysis in this section. Given the large number of placebo thresholds tested, we can conclude that the placebos do not yield meaningful effect estimates.

In summary, the validity of my RD approach depends on institutional features that ensure the running variable is not manipulable. My key finding, that establishments' test performance worsened after sustained poor performance in the pre-disclosure period, is robust to alternative RD specifications, and regressions using placebo cutoffs do not raise concerns about the main findings.

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

In this section, I evaluate the relationship between proximity to thresholds and *Salmonella* test performance, when multiple category outcomes are still possible. The analysis demonstrates that *Salmonella* test outcomes were significantly worse in every policy period when establishments had more leeway with respect to the category thresholds.

4.1. Empirical approach

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 establishment operators, then *Salmonella* control efforts should increase when there is less leeway—when the value of $leewayC$ is smaller. 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 ε_{ijk} is the residual.

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.

4.2. Results: Proximity to thresholds and Salmonella test outcomes

Table 2 presents results from regressions of the form described by equation 3, which demonstrate the effect of proximity to the thresholds on *Salmonella* test outcomes. Table 2 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 2 report results for the regressions with the *leeway* variables defined with respect to the Category 1/2 and 2/3 thresholds, respectively.[16] 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 ($p < 0.001$; panel B, column 2). 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 slightly stronger than in the previous period. When the *leeway*6 value was 10 percentage points higher, the probability of a positive test result was 3.39 percentage points higher ($p < 0.001$; elasticity = 0.74; panel A, column 4), 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; panel B, column 4).

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. 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; panel A, column 6), 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; panel B, column 6).

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. 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; panel A, column 8). 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

[16] All discussion of results in table 2 references the even-numbered columns, as they are the preferred specifications.

was 0.96 percentage points higher ($p < 0.001$; elasticity = 0.84; panel B, column 8).

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 (in an elasticity sense) when the thresholds were associated with disclosure.

5. Effects of policy changes on *Salmonella* test outcomes

Regulators face tradeoffs when designing requirements that producers disclose information about product quality. Public disclosure may mitigate moral hazard, as seen in section 3. But if the thresholds associated with categorization and disclosure are so stringent that many producers cannot meet the thresholds at low cost, these producers may significantly reduce effort irrespective of their proximity to the thresholds—another type of moral hazard. In this section, I show that while the introduction of public disclosure in 2008 reduced the average share of samples positive, the tightening of standards in 2011 raised the average share of samples positive. The latter result is driven by the worst-performing establishments.

5.1. Empirical approach

Here, I use a regression discontinuity in time (RDiT) approach (Hausman and Rapson, 2018) to evaluate the effects of each policy change on average *Salmonella* test results. 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:

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

The running variable is the sample collection date and the three 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 three cutoffs c_j, $f(\cdot)$ is a polynomial function that can take on different values on either side of each cutoff, and ε_{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.

5.2. Results: Effects of policy changes

Panel A of table 3 presents results from the RDiT model described by equation 2 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. The other policy changes, in 2006 and 2011, 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.[17] There are a couple of different possible 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 may have exited around the time of the 2011 policy change. If these establishments had similarly poor performance before and after the standards change, keeping them as part of the analyzed sample would mask changes in average *Salmonella* outcomes. The other possibility is that many operators of worse-performing establishments remained active but may have given up on trying to meet the now more stringent standard necessary to avoid disclosure.

To explore the first of these two possible 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 positive for *Salmonella* during this policy period, as compared with 4.0 percent for all other establishments ($p < 0.0001$ 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 possible interpretation, that operators gave up on trying to meet the now more stringent standard, appears to be more plausible. Table 4 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.[18] 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 responded to the 2011 policy change differently depending on their food-safety records. 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

[17] 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.

[18] All results described in the rest of this section use the same data set as panel B of table 3, dropping all establishments that ever exited.

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 January 2018.

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. Whereas in prior periods, the incentive to shirk only had effects once establishments crossed the disclosure threshold, after 2011 some establishments reduced effort even before crossing the threshold—another form of moral hazard. 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 introduced new moral hazard incentives, worsened test outcomes, and more than offset the earlier improvement.

6. Summary and conclusion

Using carcass-level data on USDA inspections for *Salmonella* in chicken carcasses from 1999 to 2015, 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 failing to meet thresholds does not subject establishments to public disclosure, *Salmonella* test performance worsens following failures. Public disclosure of categorization mitigates the shirking effect. One possible explanation is that buyers demanded that potential suppliers provide additional information about *Salmonella* test results, beyond what was publicly disclosed. (In appendix C, I demonstrate that under the more stringent disclosure policy in place in 2015 and 2016, establishment operators also 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, the initial public disclosure policy in 2008 reduced the average rate of positive *Salmonella* samples by about 55 percent, but the subsequent tightening of standards in 2011 led some establishments to considerably decrease efforts around *Salmonella* control and increased the average rate of positive samples by 140 percent. The worst-performing establishments drove the overall decline in performance after the 2011 tightening of standards, a result I attribute to another form of moral hazard or shirking.

The empirical results provide some insights about the design of information disclosure policies, especially disclosure of discrete (categorical) information. As has been demonstrated in other contexts, inspected entities have incentives to achieve better categorization but may shirk and achieve worse quality if they do not meet categorical thresholds. In this particular context, shirking was apparent when categorical information was conveyed privately to slaughter establishments but not when the categorical information was posted publicly. Thus, one policy lesson is that if categorization is used, the categorization

outcomes should be made public.

A second policy lesson is that disclosing categorical information about quality does not incentivize all producers to make effort to improve quality. The tightening of standards in 2011 resulted in worse average *Salmonella* test outcomes. Some establishment operators apparently judged the new non-disclosure standard too stringent to attain and gave up on trying. In some settings, disclosing continuous (rather than discrete or categorical) information about quality or imposing financial penalties or other sanctions for very poor performance may be necessary to incentivize quality improvements.

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, and there is some uncertainty about the sample sets I reconstructed for this analysis. Nonetheless, the paper shows convincingly that slaughter establishments responded to both well-designed and perverse incentives created by the FSIS testing and disclosure system.

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.

Bederson, Benjamin B, Ginger Zhe Jin, Phillip Leslie, Alexander J Quinn, and Ben Zou. 2018. "Incomplete disclosure: Evidence of signaling and countersignaling." *American Economic Journal: Microeconomics*, 10(1): 41–66.

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: Cambridge University Press.

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

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.

Jin, Ginger Zhe, and Phillip Leslie. 2009. "Reputational incentives for restaurant hygiene." *American Economic Journal: Microeconomics*, 1(1): 237–67.

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.

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.

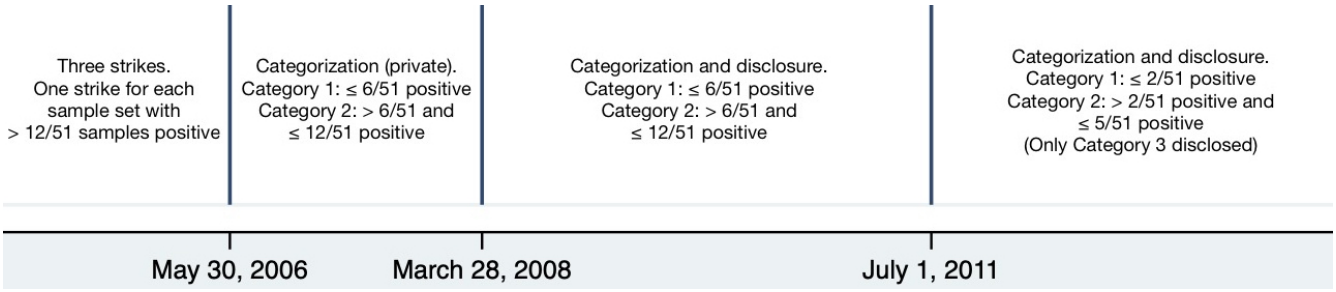
Rasschaert, Geertrui, Kurt Houf, C Godard, C Wildemauwe, M Pastuszczak-Frak, and Lieven De Zutter. 2008. "Contamination of carcasses with Salmonella during poultry slaughter." *Journal of Food Protection*, 71(1): 146–152.

Scharff, Robert L. 2020. "Food attribution and economic cost estimates for meat-and poultry-related illnesses." *Journal of Food Protection*, 83(6): 959–967.

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

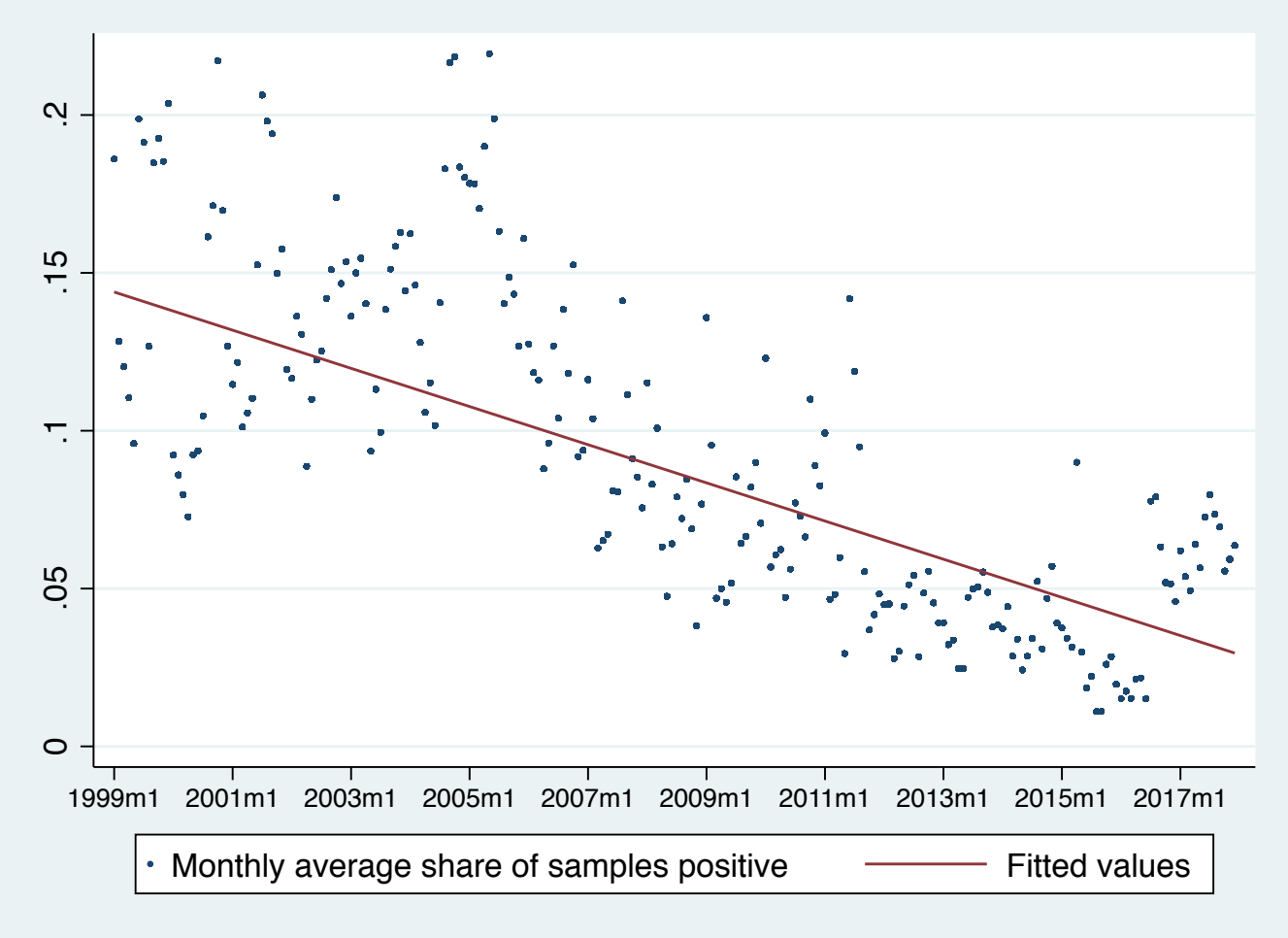
USDA. 2019. "2017 Census of Agriculture." https://www.nass.usda.gov/Publications/AgCensus/2017/.

Figure 1: Policy regimes and dates of implementation



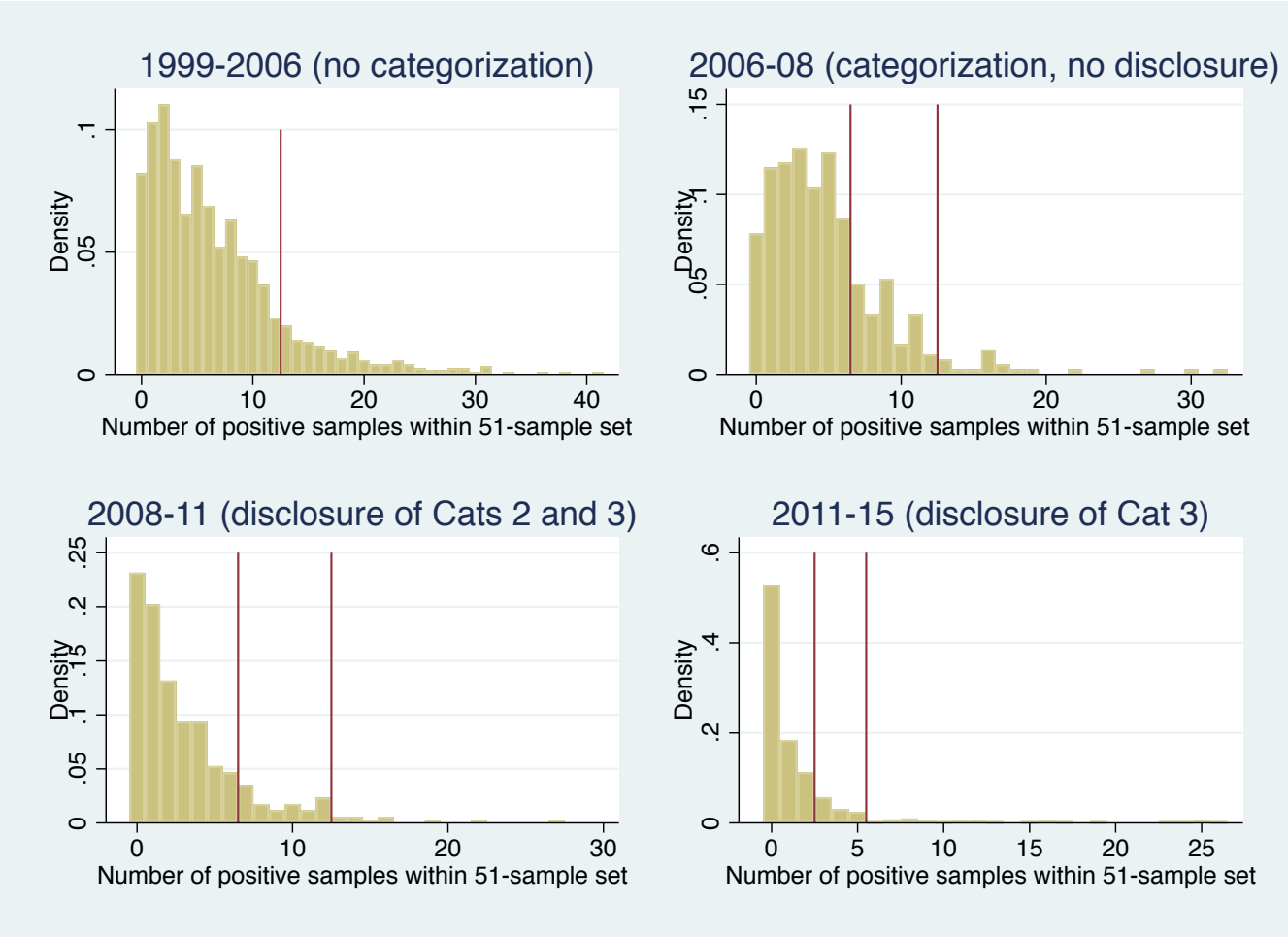
Notes: FSIS *Salmonella* testing began prior to 1999 and is still ongoing. Additional, later, policy changes are discussed in Appendix C.

Figure 2: Monthly average rate of *Salmonella* samples positive, with fitted OLS regression



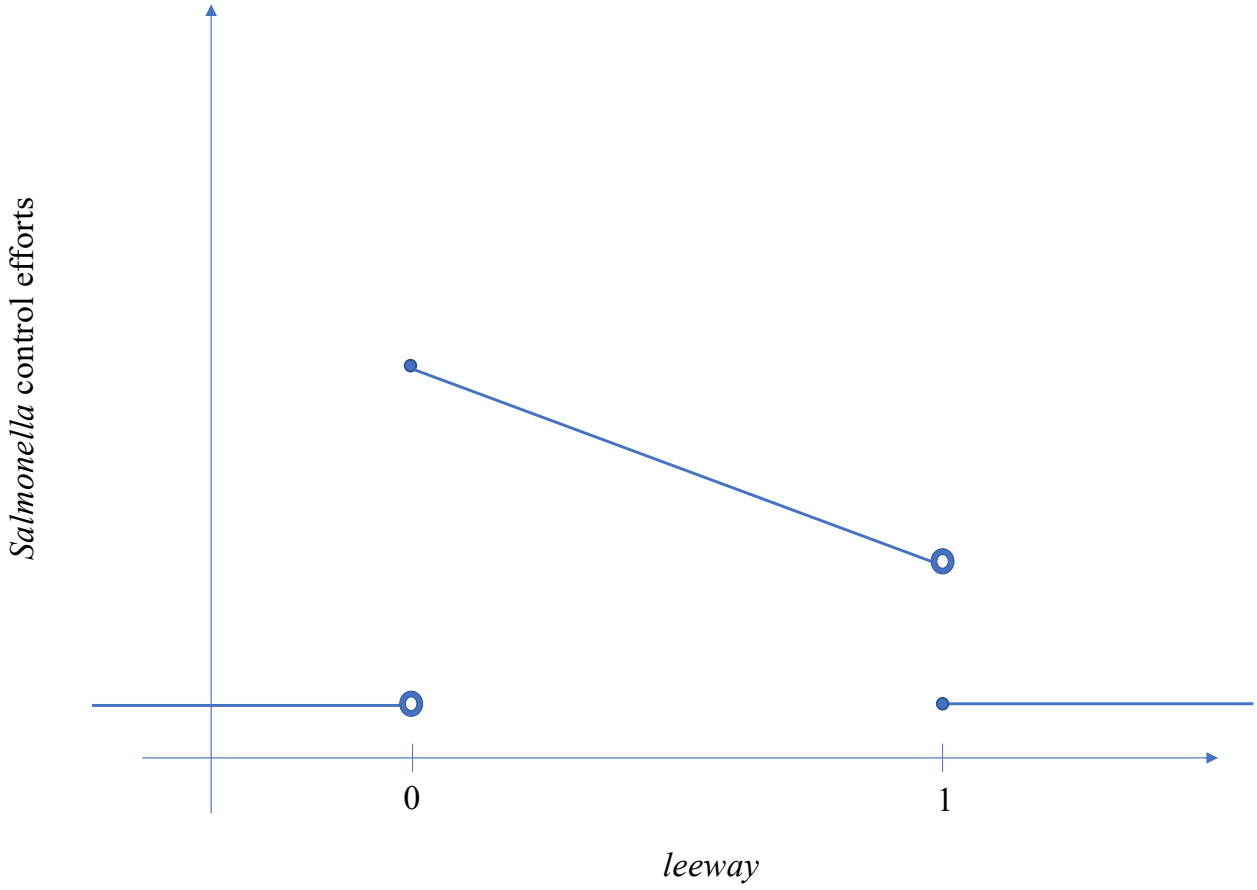
Note: OLS regression is fitted to monthly average data.

Figure 3: 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).

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: 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	
Years	1999 to 2006		2006 to 2008				2011 to 2015	
RD cutoff (c)	0	1	0	1	0	1	0	1
Max. # pos. samples (C)	12	12	6	6	12	12	2	2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$LeewayC \geq c$	0.043	−0.021	−0.061	0.046	−0.079	0.026	−0.089	0.006
Robust p-value	0.656	0.709	0.000	0.373	0.001	0.412	0.005	0.467
95% CI (lower limit)	−0.11	−0.24	−0.21	−0.09	−0.54	−0.06	−0.14	−0.01
(upper limit)	0.07	0.16	−0.06	0.03	−0.14	0.14	−0.03	0.02
Observations	23969	5594	13520	14287	10397	2396	17174	19160
Left bandwidth	0.53	0.13	1.29	0.99	0.75	0.17	0.14	0.95
Right bandwidth	0.13	2.33	0.31	2.84	0.47	2.13	0.95	1.05

Panel B: Cutoffs associated with disclosure

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

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 2: 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 3: Effects of policy changes on average *Salmonella* test outcomes

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

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 4: 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: Robustness and placebo tests

This appendix provides the results of various robustness and placebo tests described in the text.

Effects of known categorization: Robustness and placebo tests

As described in section 3, the key RD results presented in table 1 use linear polynomials with triangular kernels as recommended by Cattaneo et al. (2020b). In appendix tables B1 and B2, 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. Although the results in table B1 and B2 are not entirely consistent with those in table 1, the finding that establishment operators shirked prior to public disclosure (i.e., over 2006–08) is robust to alternative specifications.

Table B3 presents the results of RD models that use placebo cutoffs near the $c = 0$ cutoffs that yield significant estimates in table 1. As discussed in the main body of the paper, only one of the 18 cutoffs in table B3 is statistically significant and with the correct sign (i.e., a negative sign), and this estimate is only marginally significant ($p = 0.058$). In conclusion, the placebo cutoffs do not raise concerns about the validity of the main results.

Effects of policy changes: robustness tests

Here I present the results of robustness tests relevant to the RDiT design discussed in section 5. 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 B4, essentially conform with the results in panel B of table 3: the introduction of public disclosure in 2008 led to a 4.5 percentage point decrease in the share of samples positive, while the tightening of standards in 2011 led to a 6.3 percentage point increase. In this specification, the 2006 introduction of the categorization system is also estimated to have led to a statistically significant 3.4 percentage point increase in the share of samples positive. The result for 2006 is of the same sign as the insignificant result shown for that year in table 3, but is of larger magnitude.

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 3. 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 5.5 to 6.7 percentage point increase) than the main specification. Panel B of appendix table B4 shows results for the RDiT regression

with all observations within 7 days of the policy changes removed. In all of the donut specifications, the 2006 policy change is estimated to have insignificant effects on the share of samples positive.

Third, I drop all observations belonging to sample sets that span two policy periods. Under the policy regimes in place through 2015, 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. 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 5.1 percentage point decrease in the share of samples positive (though not statistically significant), while the 2011 tightening of standards led to a 13.1 percentage point increase ($p = 0.004$). The 2006 policy change had an insignificant effect.

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

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 appendix table B5, I use the dates of the relevant Federal Register announcements 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 appendix table B5, I use placebo dates 120, 240, 360, and 480 days before the actual policy changes. Under the null hypothesis, with 12 placebo cutoff values, one placebo would be expected to have $p \leq 0.083$. In appendix table B5, the lowest p-value is 0.094. We can therefore conclude that the placebo effects are the consequence of random variation and that the estimated effects of the policy changes in table 3 are valid.

Table B1: Effects of known categorization on *Salmonella* outcomes—robustness to alternative specifications

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

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

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

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

Notes: See notes to table 1.

Table B2: Effects of known categorization on *Salmonella* outcomes—robustness to alternative specifications

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

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

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

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

Notes: See notes to table 1.

Table B3: 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						
$leewayC \geq 0$	-0.056	-0.005	0.039	0.007	0.001	-0.001
Robust p-value	0.255	0.736	0.967	0.571	0.691	0.944
Panel B: 2006 to 2008, $C = 12$ positive samples						
$leewayC \geq 0$	-0.001	0.088	0.160	0.045	0.177	0.091
Robust p-value	0.810	0.889	0.698	0.440	0.042	0.109
Panel C: 2011 to 2015, $C = 2$ positive samples						
$leewayC \geq 0$	-0.073	-0.002	-0.004	0.000	0.011	0.139
Robust p-value	0.058	0.707	0.701	0.720	0.015	0.036

Notes: This table presents results of regressions paralleling those in table 1 with statistically significant results but for placebo cutoffs not associated with any change in disclosure status. Each panel reports results for three placebo cutoffs on either side of the actual cutoff ($c = 0$) according to $c \pm 0.05n$, where $n = \{1, 2, 3\}$. 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 control for sample collection date, test number within sample set, and the share of samples positive in the establishment's prior sample set.

Table B4: Effects of policy changes on average *Salmonella* test outcomes: Robustness tests

Policy introduced	Categorization (private)	Public disclosure	Public disclosure w/ tighter standards
Date of implementation (c)	5/30/2006	3/28/2008	7/1/2011
	(1)	(2)	(3)
Panel A: Observations collapsed by sample collection date			
$t \geq c$	0.034	−0.045	0.063
Robust p-value	0.046	0.015	0.001
95% CI (lower limit)	0.00	−0.07	0.03
(upper limit)	0.08	−0.01	0.12
Observations	381	326	380
Left bandwidth	372	275	284
Right bandwidth	175	200	260
Panel B: "Donut" approach: Drop all observations within 7 days of policy changes			
$t \geq c$	0.024	−0.057	0.055
Robust p-value	0.294	0.039	0.088
95% CI (lower limit)	−0.03	−0.11	−0.01
(upper limit)	0.10	−0.00	0.14
Observations	15236	7183	5414
Left bandwidth	366	204	199
Right bandwidth	220	237	233
Panel C: Drop all observations belonging to sample sets that span policy periods			
$t \geq c$	0.019	−0.051	0.131
Robust p-value	0.450	0.141	0.004
95% CI (lower limit)	−0.05	−0.13	0.05
(upper limit)	0.11	0.02	0.25
Observations	12190	3984	3125
Left bandwidth	342	170	161
Right bandwidth	259	195	245

Notes: See notes to table 3.

Table B5: 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)
Panel A: Cutoffs c = Federal Register announcement dates			
$t \geq c$	−0.032	−0.029	0.014
Robust p-value	0.622	0.859	0.263
95% CI (lower limit)	−0.13	−0.09	−0.01
(upper limit)	0.08	0.07	0.05
Observations	9527	3326	5731
Left bandwidth	354	176	165
Right bandwidth	89	60	138
Panel B: Cutoffs c = 120 days before policy changes			
$t \geq c$	−0.015	0.022	−0.008
Robust p-value	0.877	0.334	0.915
95% CI (lower limit)	−0.07	−0.04	−0.04
(upper limit)	0.08	0.12	0.05
Observations	11627	2664	2944
Left bandwidth	426	144	174
Right bandwidth	117	38	120
Panel C: Cutoffs c = 240 days before policy changes			
$t \geq c$	−0.018	0.053	0.002
Robust p-value	0.187	0.236	0.925
95% CI (lower limit)	−0.07	−0.04	−0.09
(upper limit)	0.01	0.16	0.08
Observations	27891	4213	7277
Left bandwidth	1190	96	294
Right bandwidth	237	104	113
Panel D: Cutoffs c = 360 days before policy changes			
$t \geq c$	0.019	0.029	0.030
Robust p-value	0.573	0.094	0.111
95% CI (lower limit)	−0.08	−0.01	−0.01
(upper limit)	0.15	0.08	0.08
Observations	15410	4474	6121
Left bandwidth	523	68	235
Right bandwidth	146	76	112
Panel E: Cutoffs c = 480 days before policy changes			
$t \geq c$	−0.026	0.032	−0.006
Robust p-value	0.226	0.204	0.988
95% CI (lower limit)	−0.11	−0.02	−0.06
(upper limit)	0.03	0.08	0.06
Observations	13538	5298	6117
Left bandwidth	545	66	238
Right bandwidth	160	130	108

Notes: For additional details on the regression specifications, see notes to table 3.

Appendix C: Analysis of additional policy regimes in place over 2015–2017

For clarity and ease of exposition, the body of the paper analyzes *Salmonella* test outcomes and shirking only for the four policy periods in place from 1999 until May 5, 2016. The data set I obtained from FSIS by FOIA request covers two additional policy regimes. This appendix describes those policy regimes and analysis of shirking or moral hazard over these periods.

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. 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. The rolling-window system was introduced because FSIS officials recognized that under the sample-set system, establishment operators might increase efforts related to *Salmonella* control during the weeks that establishments were under scrutiny but shirk during all other weeks of the year.[19] Moreover, the rolling-window system seemed to be an effective way to mitigate shirking: each week, a new rolling window began, so the end-of-sample-set incentives to shirk might be countered by incentives to obtain good categorization in the coming year.

Shortly after the rolling-window system was introduced, FSIS began using a new chemical solution (neutralizing buffered peptone water) as part of the test procedure.[20] After this change, which was implemented on July 1, 2016, the share of positive test results rapidly rose, and in November 2016, FSIS suspended public disclosure of *Salmonella* category information for chicken-slaughter establishments but continued to sample carcasses for *Salmonella*. No date was given for the resumption of disclosure; on December 15, 2017, FSIS announced that disclosure would resume the following month. Thus, during the final period analyzed, there were no immediate consequences for poor test outcomes. Establishment operators may have anticipated that the tests might ultimately be incorporated into their categorization, but they would not have known this for certain.[21]

Tables C1, C5, and C6 in this appendix present the results of regression models equivalent to those in tables 1, 2 and 3, covering the periods 2015–16 (rolling windows) and 2016–17 (disclosure hiatus). Tables C2, C3, C4, C7, and C8 present robustness and placebo tests equivalent to those in appendix B, covering the periods 2015–16 and 2016–17.

Effects of known categorization on Salmonella test outcomes, 2015–17

Results for RD models equivalent to those shown in table 1 are shown in table C1. During the 2015–16 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

[19] See https://www.federalregister.gov/documents/2015/01/26/2015-01323/changes-to-the-salmonella-and-campylobacter-verification-testing-program-proposed-performance, page 3945.

[20] See https://www.govinfo.gov/content/pkg/FR-2018-11-09/pdf/2018-24540.pdf.

[21] My data set also includes test results for December 15, 2017 to January 25, 2018, but because of the small number of observations for the last period I have not analyzed data from that period.

the FSIS website. Similar to the 2008–11 period, establishment operators apparently exerted effort to meet the Category 1 standard but then reduced effort once exceeding the threshold. Establishments were 5.0 percentage points more likely to have positive samples after failing to meet the Category 1 standard for the soonest-ending window (table C1, panel A, column 1). At the same time, establishments with sufficiently good performance that they were guaranteed to meet the Category 1 standard in the soonest-ending window were 3.3 percentage points more likely to have positive samples, which is evidence that they shirked after a period of sustained good test performance (table C1, panel A, column 2). This form of shirking was not evident during earlier periods. In addition, during this period, establishments appear to have reduced effort after failing to meet the Category 2 standard and therefore becoming subject to information disclosure. Establishments were 14.1 percentage points more likely to have positive samples after failing to meet the Category 2 standard for the soonest-ending window during 2015–16 (table C1, Panel B, column 1).

Under the hiatus in disclosure (2016–17), crossing thresholds associated with any of the categories had statistically insignificant effects on *Salmonella* test outcomes.[22]

Table C4, presents results for regressions parallel to those in table C1 using placebo cutoff values for the running variables (leewayC). Similar to table B3, the thresholds shown here are placebo cutoffs near the statistically significant estimates from table C1. Specifically, 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 table C4, three of the 15 RD coefficients are statistically significant with p ¡ 0.1, and two of these have the "correct" sign in the sense that they are consistent with the estimates for $c = 0$ in table C1 and the expectations about incentives for shirking that motivate the analysis. The small p-values of these two coefficients gives some pause, but they are the only two coefficients with the correct sign and $p < 0.3$. Moreover, when considering the results in table C4 together with those in table B3, only three of the 33 coefficients have the correct sign and $p < 0.3$. In conclusion, the placebo tests do not raise significant concerns about the conclusions drawn from table C1.

Proximity to thresholds and Salmonella test outcomes, 2015–17

Table C5 presents results of regressions that demonstrate the positive correlations between *leeway*2 (*leeway*5) for the soonest-ending window and the likelihood of positive *Salmonella* test results. The regressions are similar to those in table 2, except that instead of using sample sets to calculate the values of the running variable *leeway* and the regressor for share of samples positive, these regressions use the soonest-ending window. In 2015–16, when the *leeway*2 value was 10 percentage points higher, the probability of a positive *Salmonella* test result was 2.00 percentage points higher ($p = 0.012$; elasticity = 0.87; panel A, column 2). Under the disclosure hiatus, there was no statistically significant relationship between *leeway*2 and *Salmonella* test results. When the *leeway*5 value was 10 percentage points higher, the probability of a positive *Salmonella* test result was 3.03 percentage points higher in 2015–16 ($p < 0.001$;

[22]The insignificant effects are robust to the polynomial and kernel choices, as seen in tables C2 and C3.

elasticity = 0.88; panel B, column 2) and 4.16 percentage points higher in 2016–17 ($p = 0.018$; elasticity = 0.55; panel B, column 4).

Although the correlation between *leeway*2 and positive *Salmonella* test results lessened under the disclosure hiatus from 2016 to 2017, the correlation between *leeway*5 and test results increased in this period, relative to 2011–15 and 2015–16. In other words, establishment operators appear to have relaxed efforts around *Salmonella* control when they had more leeway with respect to the Category 2/3 threshold, and did so more in 2016–17 than during the earlier periods when the same threshold applied.

Effects of policy changes, 2015 and 2016

Table C6 presents results of RDiT regressions for the policy changes in 2015 and 2016. These policy changes had insignificant effects on average *Salmonella* test outcomes under the main specifications. When collapsing the data set and using the daily share of samples positive as the dependent variable (rather than carcass-level test results), the 2015 introduction of rolling windows is estimated to have decreased the share of samples positive by 2.6 percentage points, evidence of the effectiveness of the rolling-windows system (table C7, panel A, column 1). The additional robustness tests and placebo tests presented in tables C7 and C8 do not raise concerns about the validity of the main result. In conclusion, the 2015 introduction of rolling windows may have improved average test results, but the estimated effects are not as robust as those presented in table 3, which shows that the introduction of public disclosure in 2008 reduced the share of samples positive by about 55 percent.

Table C1: Effects of known categorization on *Salmonella* outcomes, 2015–17

Panel A: Cutoffs not associated with disclosure				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
RD cutoff (c)	0	1	0	1
Max. # pos. samples (C)	2	2	2	2
	(1)	(2)	(3)	(4)
$LeewayC \geq c$	−0.050	0.033	−0.029	−0.217
Robust p-value	0.030	0.006	0.127	0.212
95% CI (lower limit)	−0.09	0.01	−0.07	−0.56
(upper limit)	−0.00	0.06	0.01	0.12
Observations	8405	8076	4390	2619
Left bandwidth	1.15	1.00	2.50	1.00
Right bandwidth	0.21	1.00	0.67	1.00
Panel B: Cutoffs associated with disclosure				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
RD cutoff (c)	0	1	0	1
Max. # pos. samples (C)	5	5	5	5
	(1)	(2)	(3)	(4)
$LeewayC \geq c$	−0.141	0.061	0.004	−0.107
Robust p-value	0.027	0.643	0.880	0.580
95% CI (lower limit)	−0.25	−0.02	−0.05	−0.59
(upper limit)	−0.03	0.05	0.06	1.06
Observations	6977	8772	2365	4145
Left bandwidth	0.78	1.00	3.67	0.67
Right bandwidth	0.47	1.77	1.00	2.02

Notes: 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, 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 C2: Effects of known categorization on *Salmonella* outcomes

Panel A: Cutoffs not associated with disclosure, quadratic polynomials, triangular kernels				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
RD cutoff (c)	0	1	0	1
Max. # pos. samples (C)	2	2	2	2
	(1)	(2)	(3)	(4)
$LeewayC \geq c$	−0.049	−0.027	−0.024	−0.317
Robust p-value	0.046	0.216	0.309	0.952
95% CI (lower limit)	−0.09	−0.07	−0.06	−0.20
(upper limit)	−0.00	0.02	0.02	0.22
Observations	9096	8076	5662	2602
Left bandwidth	3.08	1.00	5.63	0.67
Right bandwidth	0.29	1.00	0.62	1.00
Panel B: Cutoffs associated with disclosure, quadratic polynomials, triangular kernels				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
RD cutoff (c)	0	1	0	1
Max. # pos. samples (C)	5	5	5	5
	(1)	(2)	(3)	(4)
$LeewayC \geq c$	−0.114	−0.005	−0.013	0.231
Robust p-value	0.016	0.006	0.606	0.602
95% CI (lower limit)	−0.21	0.03	−0.10	−1.56
(upper limit)	−0.02	0.18	0.06	2.70
Observations	7522	10680	2471	5869
Left bandwidth	2.77	1.00	4.95	1.00
Right bandwidth	0.65	3.65	0.67	3.54

Notes: See notes to table 1.

Table C3: Effects of known categorization on *Salmonella* outcomes

Panel A: Cutoffs not associated with disclosure, linear polynomials, Epanechnikov kernels				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
RD cutoff (c)	0	1	0	1
Max. # pos. samples (C)	2	2	2	2
	(1)	(2)	(3)	(4)
$LeewayC \geq c$	−0.050	0.030	−0.030	−0.198
Robust p-value	0.024	0.012	0.117	0.208
95% CI (lower limit)	−0.09	0.01	−0.07	−0.51
(upper limit)	−0.01	0.05	0.01	0.11
Observations	8293	8076	4381	2619
Left bandwidth	1.20	1.00	2.33	1.00
Right bandwidth	0.18	1.00	0.67	1.00
Panel B: Cutoffs associated with disclosure, linear polynomials, Epanechnikov kernels				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
RD cutoff (c)	0	1	0	1
Max. # pos. samples (C)	5	5	5	5
	(1)	(2)	(3)	(4)
$LeewayC \geq c$	−0.123	0.062	0.003	−0.122
Robust p-value	0.051	0.685	0.970	0.587
95% CI (lower limit)	−0.22	−0.03	−0.06	−0.58
(upper limit)	0.00	0.05	0.06	1.02
Observations	6978	8772	2176	3091
Left bandwidth	0.93	1.00	2.67	1.00
Right bandwidth	0.48	1.62	0.67	1.83

Notes: See notes to table 1.

Table C4: Placebo effects of known categorization on *Salmonella* outcomes

RD cutoff (c)	-0.15	-0.1	-0.05	0.1	0.15	0.2
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 2015 to 2016, $C = 2$ positive samples						
$leewayC \geq 0$	0.013	-0.017	-0.071	0.015	-0.011	0.004
Robust p-value	0.865	0.954	0.938	0.213	0.306	0.310
RD cutoff (c)	-0.2	-0.15	-0.1	0.05	0.1	0.15
	(1)	(2)	(3)	(4)	(5)	(6)
Panel B: 2015 to 2016, $C = 5$ positive samples						
$leewayC \geq 0$	-0.058	0.191	-0.138	-0.141	-0.026	0.003
Robust p-value	0.544	0.035	0.000	0.487	0.002	0.847
RD cutoff (c)	0.35	0.4	0.45			
	(1)	(2)	(3)			
Panel C: 2015 to 2016, $C = 2$ positive samples						
$leewayC \geq 1$	-0.009	-0.013	0.013			
Robust p-value	0.253	0.299	0.387			

Notes: This table presents results of regressions paralleling those in table 1 with statistically significant results but for placebo cutoffs not associated with any change in disclosure status. Each panel uses the nearest placebo cutoffs to the actual cutoff ($c = 0$ in panels A and B; $c = 1$ in panel C) 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. Because the only possible value of $leeway2$ greater than 1 is 2, there are no usable placebo cutoffs above $c = 1$ for panel 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 control for sample collection date.

Table C5: Effects of proximity to category thresholds on *Salmonella* test outcomes, 2015–17

	Rolling windows, 2015 to 2016		Disclosure hiatus, 2016 to 2017	
	(1)	(2)	(3)	(4)
Panel A				
*Leeway*2 (soonest-ending window)	0.258	0.200	0.261	−0.121
	(0.081)	(0.079)	(0.11)	(0.15)
Share of samples positive, soonest-ending window		−2.942		−14.07
		(1.02)		(6.50)
Observations	7787	7604	1863	1863
Elasticity	1.12	0.87	0.72	−0.34
Panel B				
*Leeway*5 (soonest-ending window)	0.368	0.303	0.660	0.416
	(0.069)	(0.063)	(0.13)	(0.17)
Share of samples positive, soonest-ending window		−2.581		−8.067
		(0.48)		(5.30)
Observations	7117	6934	971	971
Elasticity	3.84	3.16	0.88	0.55

Notes: This table represents the results of similar regressions to those shown in table 2, for the 2015–16 policy period during which sample sets were replaced with overlapping sampling windows, and the 2016–17 hiatus in public disclosure. 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 the even-numbered columns also control for the share of samples positive in the soonest-ending window. 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 C6: Effects of policy changes on average *Salmonella* test outcomes

Policy introduced	Rolling windows	Disclosure hiatus
Date of implementation (c)	5/6/2015	11/20/2016
	(1)	(2)
Panel A: All establishments included		
$t \geq c$	−0.015	0.005
Robust p-value	0.388	0.819
95% CI (lower limit)	−0.05	−0.03
(upper limit)	0.02	0.04
Observations	11935	5734
Left bandwidth	392	98
Right bandwidth	165	128
Panel B: Establishments that ever exited excluded		
$t \geq c$	−0.015	0.005
Robust p-value	0.393	0.803
95% CI (lower limit)	−0.05	−0.03
(upper limit)	0.02	0.04
Observations	13650	5795
Left bandwidth	512	99
Right bandwidth	167	129

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 C7: Effects of policy changes on average *Salmonella* test outcomes: Robustness tests

Policy introduced	Rolling windows	Disclosure hiatus
Date of implementation (c)	5/6/2015	11/20/2016
	(1)	(2)
Panel A: Observations collapsed by sample collection date		
$t \geq c$	−0.026	0.145
Robust p-value	0.066	0.115
95% CI (lower limit)	−0.06	−0.04
(upper limit)	0.00	0.38
Observations	444	144
Left bandwidth	390	81
Right bandwidth	183	98
Panel B: "Donut" approach: Drop all observations within 7 days of policy changes		
$t \geq c$	−0.024	0.006
Robust p-value	0.143	0.874
95% CI (lower limit)	−0.06	−0.04
(upper limit)	0.01	0.05
Observations	14017	4631
Left bandwidth	513	83
Right bandwidth	186	108
Panel C: Drop all observations belonging to sample sets that span policy periods		
$t \geq c$	−0.013	0.005
Robust p-value	0.424	0.785
95% CI (lower limit)	−0.05	−0.03
(upper limit)	0.02	0.04
Observations	11372	5790
Left bandwidth	389	98
Right bandwidth	170	130

Notes: See notes to table 3.

Table C8: Effects of policy changes on average *Salmonella* test outcomes: Placebo cutoff dates

	Rolling windows (1)	Disclosure hiatus (2)
Panel A: Cutoffs c = Federal Register announcement dates		
$t \geq c$	−0.014	
Robust p-value	0.617	
95% CI (lower limit)	−0.04	
(upper limit)	0.02	
Observations	10789	
Left bandwidth	440	
Right bandwidth	99	
Panel B: Cutoffs c = 120 days before policy changes		
$t \geq c$	−0.013	0.016
Robust p-value	0.894	0.435
95% CI (lower limit)	−0.05	−0.02
(upper limit)	0.04	0.05
Observations	9699	5456
Left bandwidth	357	72
Right bandwidth	119	150
Panel C: Cutoffs c = 240 days before policy changes		
$t \geq c$	−0.020	−0.007
Robust p-value	0.069	0.573
95% CI (lower limit)	−0.07	−0.03
(upper limit)	0.00	0.02
Observations	8811	4492
Left bandwidth	295	72
Right bandwidth	83	122
Panel D: Cutoffs c = 360 days before policy changes		
$t \geq c$	−0.006	0.004
Robust p-value	0.547	0.589
95% CI (lower limit)	−0.03	−0.03
(upper limit)	0.02	0.04
Observations	8271	3253
Left bandwidth	232	46
Right bandwidth	115	102
Panel E: Cutoffs c = 480 days before policy changes		
$t \geq c$	−0.008	−0.012
Robust p-value	0.726	0.157
95% CI (lower limit)	−0.04	−0.04
(upper limit)	0.03	0.01
Observations	12195	5227
Left bandwidth	316	84
Right bandwidth	161	164

Notes: Panel A does not include column (2) because the hiatus in disclosure was not preceded by a Federal Register announcement. For additional details on the regression specifications, see notes to table 3.