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

John Bovay*

January 20, 2024

Abstract

This paper examines the responses of chicken producers to public disclosure of quality information (or categorization) regarding *Salmonella* in chicken carcasses. Producers exert effort to attain better categorization and shirk when failing to meet the thresholds required for better categorization. Public disclosure reduces 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. Mail: 250 Drillfield Drive, 223 Hutcheson Hall, Blacksburg, VA 24061 United States.

1. Introduction

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 far better than the thresholds associated with each labeled category (Shewmake and Viscusi, 2015; Ito and Sallee, 2018; Barahona, Otero, and Otero, 2023). Furthermore, if producers see thresholds as unattainable, they may make very little effort to improve along the relevant quality dimensions, i.e., they may shirk. 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.

Public disclosure of food-safety outcomes may be an important policy solution to a global public-health problem: food-borne diseases cause about 600 million cases of illness and 420,000 deaths per year (World Health Organization, 2015). As the 2023 Netflix documentary "Poisoned" makes vividly clear, there are many ways that government regulators and food producers in the United States could take more action to ensure a safer food supply. Publicly disclosing information about producers' food-safety records could incentivize producers to improve their efforts related to food safety, but disclosure based on discrete quality thresholds may also incentivize shirking. By exploring tradeoffs between improved safety outcomes under disclosure, shirking in response to discrete thresholds, and stringency of disclosure thresholds (with corresponding positive and negative incentives), this paper provides new perspectives on how food safety, and more generally product quality, may be improved through public disclosure of outcomes.

This paper explores 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. I use carcass-level data on *Salmonella* test results over 1999–2017[1] for all federally inspected chicken-slaughter establishments[2] to test hypotheses about shaming and moral hazard. The analysis 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*.

Salmonella is a genus of bacteria some types of which can cause illness in humans. Chicken is the most widely consumed meat or poultry product in the United States (USDA ERS, 2023), and it is associated

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

[2]An establishment could also be referred to as a "facility", "plant", or "slaughterhouse". In this paper, I use the USDA term "establishment".

with more food-borne illnesses from *Salmonella* than any other food group (Interagency Food Safety Analytics Collaboration, 2021). In the United States, *Salmonella* is the pathogen associated with the most hospitalizations and deaths, and the second-most illnesses, of all pathogens that cause food-borne illness (Hoffmann, Maculloch, and Batz, 2015). *Salmonella* in poultry has an economic cost of up to $3.65 billion per year.[3]

The design of the USDA *Salmonella* Verification Testing Program generates incentives for establishment operators to reduce effort around *Salmonella* control. 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 May 2006 to May 2015, FSIS collected 51 carcass samples over 51 operating days and designated establishments as either Category 1, 2, or 3 (in descending order by test performance) depending on the number of positive samples within these "sample sets" of 51 carcasses. An important feature of this categorization scheme is that discrete thresholds determined assignment into categories, and there were no subcategories or continuous measures to differentiate establishments within categories. For example, under the initial categorization scheme, establishments with 6 or fewer positive samples (out of 51) were designated Category 1 and those with 7 to 12 were Category 2; those with 13 or more were Category 3. Starting in March 2008, the names and addresses of establishments in Categories 2 and 3 (the worse-performing establishments) were disclosed on a public website; starting in July 2011, categories were redefined so that Categories 1 and 2 were harder to attain, and only information about establishments in Category 3 was disclosed.[4]

Compared with a regime with disclosure of continuous information about *Salmonella* test results, the categorization and disclosure system creates clear moral hazard, specifically incentives to reduce 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, establishments with more leeway with respect to the thresholds are likely to have worse test performance.[5]

The paper provides evidence that establishments respond to incentives created by the categorization and disclosure program, sometimes by shirking or attaining worse food-safety outcomes. My results are summarized as follows.

[3]Hoffmann, Maculloch, and Batz (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 $3.65 billion in November 2023 dollars. Scharff (2020) provides a similar estimate.

[4]Additional policy changes took place in May 2015 and November 2016; see appendix C for discussion and analysis of these changes.

[5]In this paper, leeway with respect to a threshold is the maximum share of positive samples among remaining samples in the set that would allow the establishment to meet a threshold and attain the better of two possible categorizations. The concept is formally defined in section 5.

First, using a regression discontinuity (RD) approach with leeway as the running variable, I demonstrate that when categorization was effectively determined, establishments' performance on subsequent *Salmonella* tests was significantly worse than that of establishments without categorization determined, under most policy regimes. In particular, the RD results show that: (1) After establishments fail to meet categorization thresholds but these failures do not subject them to public disclosure, *Salmonella* test performance worsens for the remainder of the sample set, suggesting that establishment operators reduce effort related to controlling *Salmonella*. (2) In the initial public-disclosure period, when establishments failed to meet thresholds and were therefore subjected to public disclosure, there was no statistically significant change in *Salmonella* test performance. In other words, the shirking effect appears to be mitigated under public disclosure. (3) When the standard for disclosure was tightened in July 2011, establishments that failed to avoid disclosure again had worse performance on subsequent tests for the remainder of the sample set. (4) Prior to public disclosure, there was also some evidence of shirking after sustained good performance guaranteed a certain categorization outcome. Note that I do not observe inputs to production or safety protocols undertaken by establishments, so I am only able to analyze how incentives affected *Salmonella* test outcomes; some of the empirical results suggest or imply that establishments responded to incentives by reducing effort. There are several possible explanations for the specific patterns of shirking behavior, which are consistent with features of the industry and the evolution of *Salmonella* test performance over the period analyzed, as discussed in section 5.

Second, I document that when establishments have more leeway with respect to the thresholds, their performance on *Salmonella* tests worsens. The relationship between distance from 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 (RDiT) approach to demonstrate the effects of each policy change on average *Salmonella* test results. This analysis shows that the introduction of public disclosure in March 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 or shirking outcome not related to current performance with respect to the thresholds. 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 July 2011 was to increase overall *Salmonella* rates by about 140 percent.

This paper demonstrates that chicken producers responded to the incentives created by the inspection program by reducing provision of food safety when the stakes were low. The results bear resemblance to studies on responses to the introduction of quality ratings, such as Dranove et al. (2003), which found that hospitals responded by focusing on healthier patients and ignoring the sickest; Jacob (2005), which found that schools responded with gaming behavior such as finding ways to avoid reporting scores of poorly performing students; and Jacob and Levitt (2003) and Dee et al. (2019), which found that teachers responded by cheating on standardized tests. (In contrast, Pope (2019) finds that the release of teacher value-added ratings in Los Angeles resulted in improved test scores for the students of low-rated teachers and presents suggestive evidence that teachers reallocated their efforts after the ratings were released.)[6] Similarly, Houde (2018) finds evidence that the energy efficiency of refrigerators is bunched just below the threshold necessary to obtain Energy Star certification, Shewmake and Viscusi (2015) find that home builders strategically incorporate "green" features to achieve green certifications, and Barahona, Otero, and Otero (2023) find that food manufacturers reformulate products to avoid being disclosed as unhealthy, with evidence of bunching in nutrient levels. 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), Clean Air Act violations (Evans, 2016), toxic emissions (Campa, 2018), farm antibiotic use (Belay and Jensen, 2020), and workplace safety violations (Johnson, 2020).

There is also a parallel with taxation theory and empirical evidence suggesting that (repeated) taxable income tends to fall just below thresholds at which the marginal tax rate changes discontinuously (Saez, 2010). According to the standard model (Saez, 2010), the distribution of (reported) income should be smooth except when there are discontinuities in the marginal income tax rate. Discontinuities in the tax schedule, in other words, create incentives for self-reported workers to either earn less or report less than they would under a smooth tax schedule, which reduces tax revenues compared with a smooth schedule. Bunching in the distribution of (reported) income is especially evident when the marginal tax rate discontinuously increases from zero to some positive number, or from negative (i.e., a subsidy) to zero (Saez, 2010), or when there are discontinuities in the average tax rate (Kleven and Waseem, 2013). Although few studies explicitly assess social welfare implications of discontinuities in tax schedules, Sallee and Slemrod (2012) show that the social losses caused by automakers' bunching responses to the U.S. Gas Guzzler Tax over 1991–2009 were four times greater than the social gains that would have resulted from a smooth tax schedule.

The results on shirking bear some resemblance to the theory of effort under rank-order tournament incentive schemes (Lazear and Rosen, 1981). Specifically, Lazear and Rosen (1981) show that when workers (players) are heterogeneous in ability, some players underinvest in effort and others overinvest, depending on their expectations about their abilities and therefore their expected outcomes relative to other players in the tournament. More recent studies (e.g., Grant, 2016; Adams and Waddell, 2018; Lemus and Marshall, 2021) have evaluated how information that changes expectations about final outcomes at intermediate points in tournaments or competitions leads some competitors to change their levels or risk-taking or effort. However, the application under study in this article is different from applications of tournament theory because unlike broiler farms, chicken-slaughter establishments are not competing

[6] For a more thorough review of evidence on responses of hospitals and schools to quality disclosure, see Dranove and Jin (2010).

directly against a fixed pool of opponents.[7] Moreover, this article provides evidence that effort changes in response to meeting or failing to meet fixed thresholds, whereas the tournament theory literature presents theory and evidence on how effort depends on competitors' subjective expectations about winning or meeting some target score.

Over the years, there has been a rich discussion about the best form of government intervention to improve food safety, which can be characterized as a credence attribute (Caswell and Mojduszka, 1996) about which both producers and consumers have imperfect information (Antle, 2001)—although producers nearly always have better information (Golan et al., 2004; Pouliot and Wang, 2018). Shapiro (1983) shows that both mandating information provision about product quality and imposing minimum quality standards can be welfare improving. Economists generally agree that regulation of performance standards or quality outcomes is more efficient than regulating production processes (see, e.g., Antle 1996; Josling, Roberts, and Orden, 2004, p. 23; Bovay, 2023). The analysis in this paper provides evidence that mandatory disclosure of information related to performance standards changes producers' behavior, although the performance standards do not always improve safety.

USDA FSIS also inspects other types of meat and poultry products for *Salmonella* and other pathogens, and has implemented similar categorization and disclosure programs for many of these products. Evidence on disclosure, stringency, and shirking around results of tests for *Salmonella* in chicken should therefore be seen as a meaningful example that may hold lessons for food-safety regulatory issues in other types of meat and poultry because of the prevalence of *Salmonella* in chicken. The findings may inform ongoing policy development, as FSIS continues to refine its inspection and disclosure programs.

Section 2 provides additional background information on the chicken-slaughter industry and federal food-safety inspections. Section 3 describes the data and provides descriptive statistics. Section 4 presents a model that provides hypotheses about effort to ensure food safety under categorization and disclosure. Sections 5, 6, and 7 contain the empirical approaches and results. Section 5 demonstrates the effects of known categorization on *Salmonella* test outcomes using an RD design with leeway as the running variable. Section 6 explores the effects of distance from thresholds on *Salmonella* test outcomes when categorization is unknown. Ollinger and Bovay (2020) find that, in the same context as this paper, public disclosure in March 2008 improved *Salmonella* test results. Using an RDiT approach, section 7 confirms the earlier finding but also show that the July 2011 tightening of disclosure standards resulted in worse average *Salmonella* test results, a result driven by the worst-performing establishments. Section 8 concludes. Appendices provide a description of the data-cleaning procedure additional validation and

[7] The majority of farmers raising broilers under production contracts are paid by integrators (which own slaughter establishments) under a tournament incentive system (Knoeber, 1989; Knoeber and Thurman, 1994, 1995, see also https://www.govinfo.gov/content/pkg/FR-2023-11-28/pdf/2023-24922.pdf). There are notable differences between the setting of this paper and the upstream transactions that are based on rank-order tournaments. Buyers of slaughtered broilers generally do not have exclusive contracts with integrators, different from the exclusive relationships between integrators and growers. Instead, integrators sell to multiple buyers and compete with each other in overlapping markets (that depend on geography and other factors). In addition, to the extent that prices integrators receive for broilers may depend in part on *Salmonella* test results, the key criterion is whether establishments have met the Category 1 threshold; performance relative to other establishments within the same category is likely to matter less and is also unknown to producers.

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

2. Background on the chicken-slaughter industry, *Salmonella*, 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.[8] 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.[9]

Salmonella is typically present in the intestines of birds and other animals. Because chickens are coprophagic and are nearly always raised in crowded environments, live birds entering a slaughter establishment are likely to have pathogens from feces on their feathers, feet, and skin, which may spread to the meat of the same bird or other birds during slaughter (USDA FSIS, 2021).[10] *Salmonella* may even be spread from birds slaughtered one day to carcasses slaughtered the next day if cleaning and disinfecting procedures are insufficient (Zeng et al., 2021). In a poultry-slaughter establishment, the basic process is that birds are killed, then cleaned, trimmed, and chilled. The share of samples testing positive for *Salmonella* generally decreases as carcasses move through the processing line, from slaughter to chill tank (Boubendir et al., 2021), which demonstrates that additional processing steps generally reduce risk by improving hygiene rather than increasing risk because of cross-contamination. Rinsing and steaming the carcasses, using disinfectants such as peracetic acid and chlorine, and chilling can all reduce the risk that

[8] During the period covered in this paper (1999 to 2017), there were 300 federally inspected chicken-slaughter establishments, but 74 of these exited the industry or opted for state inspection during the period.

[9] 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.

[10] On-farm practices, including vaccination, feed supplements, hygiene, and replacement of bedding material, can also reduce the risk that live chickens carry *Salmonella* to the slaughter establishment. See Harshavardhan Thippareddi, Manpreet Singh, Todd Applegate, and Sudhir Yadav, "Spotlight: A Critical Look at Reducing the Risk of Salmonella from Poultry—Part 2", *Food Safety Magazine*, October 4, 2022, https://www.food-safety.com/articles/8029-spotlight-a-critical-look-at-reducing-the-risk-of-salmonella-from-poultrypart-2 and Julie Larson Bricher, "Blockchain believer", *Meatingplace*, November 2018, http://library.meatingplace.com/publication/?m=21516&i=537911&p=82&ver=html5.

carcasses contain *Salmonella* (Buncic and Sofos, 2012).[11] Many risk-reducing processes can be applied at different levels (e.g., length of time, concentration of disinfectants, temperature), all of which are associated with different costs. Some risk-reducing processes can be adjusted quickly, while others (such as better defeathering equipment and chill tanks with higher efficiency) require capital investments.[12]

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 with 6 or fewer positive samples in a 51-sample set were designated Category 1; those with 7 to 12 positive samples were designated Category 2; and establishments that failed to meet the regulatory standard, with 13 or more positive samples, were designated Category 3. The new category designations were conveyed to firms privately until March 28, 2008, when the names and locations of Category 2 and 3 establishments were posted publicly on the FSIS website.[13] 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 2 or fewer positive samples out of 51 were designated Category 1; those with 3 to 5 positive samples were designated Category 2; and those with 6 or more positive samples were designated Category 3. 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. Along similar lines, table 1

[11] See also Harshavardhan Thippareddi and Manpreet Singh, "A Critical Look at Reducing the Risk of Salmonella from Poultry, Part 3: Processing Controls", *Food Safety Magazine*, December 6, 2022, https://www.food-safety.com/articles/8183-a-critical-look-at-reducing-the-risk-of-salmonella-from-poultry-part-3-processing-controls. Rinsing chicken is not recommended in home or restaurant kitchens, because the main effect is to spread bacteria to the sink and other surfaces (Henley et al., 2016).

[12] Ollinger and Bovay (2018) find evidence suggesting that ground-beef producers are selectively attentive to *Salmonella* when producing beef to supply the National School Lunch Program, which imposes a zero-tolerance standard for *Salmonella* in that product.

[13] 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.

shows that the share of sample sets in Category 1 increased from about 60% before 2006 to 85% during the initial public disclosure period (2008 to 2011), and remained at nearly the same level even after the requirements for being included in Category 1 were made significantly more stringent in 2011. Correspondingly, the shares of sample sets in Categories 2 and 3 fell over time. 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.

The policy changes described above did not come as a surprise to regulated entities but were the last stages of gradual public dialogue between FSIS and the slaughter industry. The policy changes around testing chicken for *Salmonella* also were part of a broader effort by FSIS to improve food safety in the meat and poultry industry through testing and disclosure. For example, producers may have anticipated that disclosure of *Salmonella* test categories was inevitable, long before the Federal Register announcement two months prior to the beginning of disclosure. Since it is likely that some producers anticipated either tighter thresholds or disclosure of additional information before the policy changes, and may have adjusted operations accordingly, the effects of policies and policy changes on outcomes may be dampened, compared with a counterfactual in which policy changes came as complete surprises.

The safety of poultry processing remains relevant in policymaking today. In October 2021, FSIS formally announced a program to investigate future regulatory actions with the goal of reducing *Salmonella* in poultry by 25%.[14] According to the proposed regulatory framework, chicken flocks will be tested for *Salmonella* before entering slaughter establishments to help establishment operators take appropriate risk-reducing actions within the establishment.[15] FSIS will also modify process control requirements for chicken-slaughter establishments and is considering declaring *Salmonella* an adulterant (when it is in high levels, or of certain serotypes), which would allow FSIS to enforce a final product standard for *Salmonella* in chicken. FSIS also announced in 2022 that it will begin measuring the number of *Salmonella* cells present in poultry samples rather than just testing for the presence or absence of *Salmonella*.[16]

3. 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

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

[15]See https://www.fsis.usda.gov/sites/default/files/media_file/documents/FINAL-Salmonella-Framework-10112022-508-edited.pdf.

[16]See https://www.fsis.usda.gov/news-events/news-press-releases/constituent-update-august-5-2022-0.

sets") used to determine regulatory compliance and category designations over 1999–2015.[17] 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.[18] The complete data set (covering 1999 to 2018) includes 172,571 non-duplicate observations for 300 establishments. For the period 1999 to 2015, there are 2,448 sample sets. Table 1 summarizes the number of sample sets in each category and the average share of samples positive in each policy period.

The basic data provide some evidence that establishment operators were attentive to the thresholds and may have adjusted their operations to avoid exceeding the thresholds. Figure 3 shows histograms of the number of positive samples per sample set for each of the four policy periods over 1999–2015. Establishment operators were unable to precisely manipulate the number of positive samples per set because the presence of *Salmonella* bacteria in chicken carcasses cannot be precisely controlled and because carcasses were pulled out of processing lines at random to be sampled.[19] 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 23.4% of sample sets had 3 or 4 positive samples, and 20.6% had 5 or 6, while only 8.5% 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.

4. A model of effort under categorization and disclosure

A simple model demonstrates how producers' decisions to exert effort related to *Salmonella* control may be a function of recent test results, categorization, and disclosure. Let the incremental profit of establishment

[17]For conciseness, throughout the rest of the paper, I generally refer only to the years in which policy regimes started and ended, rather than the precise dates of policy change described in section 2.

[18]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.

[19]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.

k as a function of effort e_{ik} preceding the sampling of carcass i be given by

$$\pi_{ik} = R\left(\sigma_{ik}\left(e_{ik}, \varepsilon_{1ik}\right), cat_k\left(e_{ik}, \varepsilon_{2ik}\right)\right) - C\left(e_{ik}\right), \tag{1}$$

where R represents incremental revenue, σ_{ik} is the share of carcass samples that will have tested positive after sample i is collected, ε_{1ik} and ε_{2ik} are (correlated) stochastic error terms, cat_k is the category that establishment k will be assigned at the end of the current sample set,[20] and C represents the costs of effort.

Suppressing subscripts, the expected incremental profit is then given by:

$$E\left[\pi\right] = E\left[R\left(\sigma\left(e\right), cat\left(e\right)\right)\right] - C\left(e\right). \tag{2}$$

Assume that incremental revenue is differentiable with respect to σ, σ is differentiable with respect to effort, and costs are differentiable with respect to effort. Finally, assume that $E\left[cat\right]$ is differentiable with respect to effort; even though *cat* is discrete, its expected value may be continuous. This last assumption is consistent with the notion of diminishing marginal returns to effort.

The derivative of expected incremental profit with respect to effort is:

$$\frac{\partial E\left[\pi\right]}{\partial e} = \frac{\partial R}{\partial \sigma}\frac{\partial \sigma}{\partial e} + \frac{\partial R}{\partial E\left[cat(e)\right]}\frac{\partial E\left[cat(e)\right]}{\partial e} - \frac{\partial C}{\partial e}. \tag{3}$$

If revenue gains are associated with a smaller share (σ) of carcass samples testing positive for *Salmonella*, and as long as σ decreases with effort e, then $\frac{\partial R}{\partial \sigma}\frac{\partial \sigma}{\partial e} > 0$. If revenue increases with better categorization outcomes (i.e., Category 1 or 2), and as long as effort leads to an improvement in expected categorization, then $\frac{\partial R}{\partial E[cat(e)]}\frac{\partial E[cat(e)]}{\partial e} > 0$. We should assume that costs are increasing in effort, $\frac{\partial C}{\partial e} > 0$.

The key term in equation (3) is the second term, $\frac{\partial R}{\partial E[cat(e)]}\frac{\partial E[cat(e)]}{\partial e}$. If categorization is not in place, then $\frac{\partial R}{\partial E[cat(e)]}\frac{\partial E[cat(e)]}{\partial e} = 0$, and establishments will choose effort to equate marginal incremental revenue and marginal costs, so that

$$\frac{\partial R}{\partial \sigma}\frac{\partial \sigma}{\partial e} = \frac{\partial C}{\partial e}. \tag{4}$$

In other words, in the absence of categorization, effort related to *Salmonella* control is optimal if the marginal benefits from reducing the share of samples positive equal the marginal cost.

Similarly, if there is no expectation that changes in the categorization outcome could result from changes in effort, then $\frac{\partial E[cat(e)]}{\partial e} = 0$, and establishments will choose effort as in equation (4). When categorization is in place, $\frac{\partial E[cat(e)]}{\partial e} = 0$ only if sufficiently good or poor performance over the set of carcass samples $1, \ldots, i-1$ guarantees a known categorization outcome. Thus, in this model, effort related to *Salmonella* control is greater after the introduction of categorization than before, if there are some benefits

[20] cat_k is not determined until (near) the end of the sample set but effort throughout may affect the categorization outcome.

from better categorization outcomes, and if categorization outcomes are not guaranteed already on the basis of good or poor performance. Because $\frac{\partial R}{\partial E[cat(e)]}\frac{\partial E[cat(e)]}{\partial e} > 0$, guaranteed categorization outcomes reduce effort.

If categorization outcomes are not known, then optimal effort depends on $\frac{\partial E[cat(e)]}{\partial e}$, i.e., how effort affects the likelihood that a worse categorization outcome will result. This depends on the relationship between effort and *Salmonella* test outcomes, of course, but it also depends on the number of samples remaining within the sample set and the number of samples that have tested positive so far within the sample set. For example, consider the policy in place from 2011 to 2015. If an establishment had more than 5 positive samples over a 51-sample set, the name of the establishment would be disclosed. If establishment a had zero positive samples among the first 45, additional (or reduced) effort would have been unlikely to affect ultimate categorization. In contrast, if establishment b had five positive samples among the first 45, additional or reduced effort could have had a strong effect on the expected categorization outcome. In this example, $\frac{\partial E[cat(e_b)]}{\partial e_b} < \frac{\partial E[cat(e_a)]}{\partial e_a} < 0$, and the returns to effort would be greater for establishment b, assuming that each establishment's effort has the same effect on the probability of a positive sample. Recalling the concept of leeway—the share of remaining samples within the sample set that may test positive if the establishment is to achieve a given categorization—establishment a has more leeway than establishment b.

Figure 4 illustrates the implications of the model for the relationship between expected *Salmonella* control efforts and leeway. When $leeway < 0$ or $leeway \geq 1$, categorization outcomes are known, and effort is chosen such that $\frac{\partial R}{\partial \sigma}\frac{\partial \sigma}{\partial e} = \frac{\partial C}{\partial e}$. When categorization outcomes are not known, effort is chosen such that $\frac{\partial R}{\partial \sigma}\frac{\partial \sigma}{\partial e} + \frac{\partial R}{\partial E[cat(e)]}\frac{\partial E[cat(e)]}{\partial e} = \frac{\partial C}{\partial e}$; all three terms are positive, so effort is expected to be larger than when categorization is known. When $leeway$ is smaller within the $[0, 1)$ interval, establishments are expected to exert more effort to control *Salmonella* because the returns to doing so are greater.

The simple model outlined in this section generates hypotheses about how the *Salmonella* testing, categorization, and disclosure program may affect effort, but effort cannot be directly observed. Instead, the empirical work outlined in the following sections demonstrates the effects of known categorization, distance from regulatory thresholds, and policy changes on test outcomes, which are related to unobservable effort.

5. 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. Based on the model outlined in the previous section, 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). Either of these outcomes causes $\frac{\partial R}{\partial E[cat(e)]}\frac{\partial E[cat(e)]}{\partial e} = 0$, within equation (3). 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.

5.1. Empirical approach

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

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

$$leeway\kappa_{ijk} = \frac{\kappa - \sum_{l=1}^{i-1} Y_{ljk}}{52 - i}, \tag{5}$$

where $\kappa \in \{2, 5, 6, 12\}$ is the maximum number of samples permitted to be positive within a sample set, to achieve the given category; i is the sample 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 $[1, i-1]$. The denominator $52 - i$ is a count of the total number of samples 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.[21]

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

$$Y_{ijk} = \alpha + \beta_0 D_{0ijk} + \beta_1 D_{1ijk} + f\left(leeway\kappa_{ijk}\right) + \varepsilon_{ijk}, \tag{6}$$

where Y_{ijk} is a binary variable representing the results of test i for *Salmonella* within sample set j at establishment k ($Y_{ijk} = 1$ when test i is positive), $D_{0ijk} = \mathbf{1}\left\{leeway\kappa_{ijk} < 0\right\}$, $D_{1ijk} = \mathbf{1}\left\{leeway\kappa_{ijk} \geq 1\right\}$,

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

$f(\cdot)$ is a polynomial function that can take on different values on either side of each cutoff ($c \in \{0, 1\}$); and ε_{ijk} is the residual. Following Calonico, Cattaneo, and Titiunik (2014b), Cattaneo, Idrobo, and Titiunik (2020b), and Cattaneo, Titiunik, and Vazquez-Bare (2020), 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. Note that within all sample sets, the value of $leeway\kappa$ starts at $0 < \kappa/51 < 1$ and that as more samples are taken, $leeway\kappa$ may decrease or increase and eventually cross the 0 or 1 thresholds, causing either $D_0 = 1$ or $D_1 = 1$. If $D_0 = 1$, then the better categorization outcome cannot be attained; if $D_1 = 1$, then the better categorization outcome is certain to be attained. Thus, positive values of either RD coefficient β_0 or β_1 imply that establishments shirk when crossing thresholds, consistent with the expectations outlined in section 4.

5.2. *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. There was not evidence of shirking during the initial period when categorization outcomes were publicly disclosed, but there was strong evidence of shirking when establishments failed to meet the more stringent disclosure threshold in place beginning in 2011.

RD plots for each regulatory threshold and each period are shown in figures 5 and 6. From these plots, it appears that the $leeway\kappa = 0$ threshold affected producers' behavior in most periods. Panel A of table 2 shows estimates of the RD coefficients at the $leeway\kappa = 0$ and $leeway\kappa = 1$ cutoffs for the thresholds κ associated with regulation or categorization but not with disclosure, and panel B shows estimates of the same RD coefficients for the thresholds κ associated with disclosure. Interpretations of specific results in table 2 follow. Note that shirking is suggested by discontinuous increases in the share of samples positive as the value of $leeway\kappa$ crosses 0 from above or crosses 1 from below, in figures 5 and 6, and that these increases correspond with positive coefficients in the top line of each column in table 2.

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, establishments were 3.4 percentage points more likely to have positive *Salmonella* test outcomes after failing to meet the 12/51 threshold (see table 2, panel A, column 1).

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 8.4 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 14.9 percentage points more likely to have positive samples after failing to meet the Category 2 standard (see table 2, panel A, columns 3 and 5). 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. In addition, during the 2006–08 period, establishments were 10.2 percentage points more likely to have positive test outcomes after good performance ensured they would avoid a Category 3 outcome (table 2, panel A, column 6). Thus, the results suggest that establishment operators shirked after either sustained poor performance or sustained poor performance ensured that categorization outcomes were known.

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 2, panel B, columns 1–4, show that crossing thresholds such that categorization outcomes were known had statistically insignificant effects on subsequent *Salmonella* test performance. Disclosure thus may have reduced establishment operators' incentives to shirk.

During the 2011–15 policy period, the thresholds associated with Category 2 and 3 were tightened so that Category 1 consisted of establishments with 2 or fewer positive samples out of 51 and Category 3 consisted of establishments with 6 or more. Under these new, more stringent thresholds, only the names of Category 3 establishments were publicly disclosed. During 2011–15, establishments were 7.8 percentage points more likely to have positive samples after failing to attain Category 1 status (table 2, panel A, column 7). They were also 15.8 percentage points more likely to have positive samples after failing to attain Category 2 status and ensuring disclosure (table 2, panel B, column 5). 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.

I now summarize the results in table 2. First, when establishments fail to meet thresholds but are not subject to public disclosure, *Salmonella* test performance typically worsens (panel A, columns 1, 3, 5, and 7). Second, during the initial public disclosure period, when establishments failed to meet thresholds that subjected them to public disclosure, there was no statistically significant change in *Salmonella* test performance (panel B, columns 1 and 3). Third, establishment operators relaxed efforts after sustained good performance on *Salmonella* tests ensured they would avoid a Category 3 outcome in the pre-disclosure period (panel A, column 6), but there was no evidence of shirking after sustained good performance in the disclosure period (panel B, columns 2, 4, and 6). Fourth, during the period with a more stringent threshold for avoiding public disclosure, test performance worsened after sustained poor performance ensured that the categorization outcome would be publicly disclosed (panel B, column 5).

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. This shirking behavior existed even though FSIS did not provide information about categorization and therefore the thresholds should have mattered little to producers and should have been unknown to buyers. In contrast, during the initial public disclosure period, there was no statistically significant evidence of shirking. These findings appear to be puzzling, but are consistent with three possible explanations.

First, buyers sometimes demand additional information about *Salmonella* test results, beyond what is publicly disclosed. A representative of a large-scale vertically integrated poultry producer that owns slaughter establishments indicated that in the period prior to public disclosure of category information, some buyers requested that producers disclose their categorization status (personal communication, October 2023). In addition, some buyers have required that chicken come from Category 1 establishments and that supplying establishments that fall out of Category 1 must review their food-safety practices monthly and take corrective action until returning to Category 1 (pers. comm., Oct. 2023). In the event that establishments move from Category 1 to Category 2 or 3, some buyers will also demand additional information about food safety, including the results of private (non-FSIS) tests for *Salmonella* in carcasses (pers. comm., Oct. 2023). If, prior to public disclosure, some buyers only requested categorical information but did not request information about the results of private tests or details about the number of positive samples within Category 2 or 3, then their suppliers would have had incentives to shirk once failing to meet the Category 1 standard. As more buyers began to request additional information, these incentives to shirk would have been lessened. In addition, it is possible that even in the absence of additional oversight from buyers, owners of establishments may have enforced stricter controls when the better categories were not achieved.[22]

A second possibility is that producer behavior was influenced by expectations about future changes to the testing and disclosure policies. For example, if producers anticipated that carcass-level test results would eventually be made available by FSIS, this would have dampened incentives to shirk.

A third possible reason that shirking was evident in the period prior to public disclosure (2006–08) but not during the initial public disclosure period (2008–11) is that improvements in performance on food safety tests meant that more establishments were in Category 1, and fewer in Categories 2 and 3, during the later period. To be specific, as seen in table 1, 75% of all sample sets placed establishments into Category 1 over 2006–08 but 85% of sample sets placed establishments into Category 1 over 2008–11. As Category 2 and 3 outcomes became less likely, the establishments in those categories may have been mainly selling to buyers that did not demand or expect Category 1 outcomes. If the customers of such lower-quality producers did not care about categorization outcomes, then crossing the threshold into Category 2 or 3 would not have changed incentives for producers. Thus, shirking would not have been as apparent.

Whatever the reason, public disclosure seems to have reduced the incentives to shirk until standards were tightened in 2011, after which shirking behavior was again evident for establishments crossing thresholds that ensured both Category 2 and 3 outcomes.

5.3. Validity of the RD design and robustness tests

In most contemporary studies that use RD approaches (see Lee and Lemieux, 2010; Calonico, Cattaneo, and Titiunik, 2014b; Cattaneo, Idrobo, and Titiunik, 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, Jansson, and Ma

[22]For some anecdotes about how producers and buyers have implemented their own tests for *Salmonella*, especially in the wake of major safety problems, see Lawrence Aylward, "Data driven", *Meat + Poultry*, December 10, 2015, https://www.meatpoultry.com/articles/19172-data-driven and Julie Larson Bricher, "Blockchain believer", *Meatingplace*, November 2018, http://library.meatingplace.com/publication/?m=21516&i=537911&p=82&ver=html5.

(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 $leeway\kappa$ 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, Jansson, and Ma (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.[23] For another comparison of smoothness in the running variable, I use t-tests to compare the ratios of the number of observations with $leeway\kappa = 0$ and $leeway\kappa = 1$, over the number of observations with $leeway\kappa < 0$ and $leeway\kappa \in (0, 1)$, across my real and simulated data. I find that the real data are significantly smoother than the simulated data at $leeway\kappa = 0$ ($p = 0.002$) and almost exactly as smooth at $leeway\kappa = 1$. 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.[24] 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.[25] 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

[23]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.

[24]The baseline covariates tested included sample collection date, the share of samples positive in the prior sample set, and sample number ($i = 1, \ldots, 51$) within sample set.

[25]Recall that the denominator of the running variable is sample number within the sample set, which cannot be controlled by the establishment managers or inspectors. 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.

establishment managers exerted (legitimate) effort to stay below the thresholds, and not that fraudulent behavior helped them stay below the thresholds.[26]

Appendix table B1 presents results for regressions parallel to those in table 2 but using placebo cutoff values for the running variables ($leeway\kappa$). The time periods and thresholds shown in table B1 represent the statistically significant estimates from table 2. The placebo cutoff values are three multiples of 0.05 in either direction from $c = 0$ or $c = 1$. In appendix table B1, several of the RD coefficients are statistically significant with $p < 0.1$, but only three of the 36 coefficients are statistically significant and have the positive sign that suggests shirking. 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, and regressions using placebo cutoffs do not raise concerns about the main findings.

6. Distance from regulatory thresholds and *Salmonella* test outcomes

In this section, I evaluate the relationship between distance from 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.

6.1. 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 $leeway\kappa$. Larger values of $leeway\kappa$ indicate that a larger share of remaining samples could test positive for *Salmonella*. In terms of equation (3) from the model in section 4, when $leeway\kappa$ is smaller, $\left|\frac{\partial E[cat(e)]}{\partial e}\right|$ is greater, and the returns to effort are greater. Therefore—assuming that *Salmonella* category assignment matters to establishment operators—*Salmonella* control efforts should increase when the value of $leeway\kappa$ is smaller. To estimate the relationship between $leeway\kappa$ and test outcomes when multiple category outcomes are possible, I use only observations with $leeway\kappa \in [0, 1)$.

I estimate the relationship between $leeway\kappa$ and *Salmonella* test outcomes under each policy regime using a series of linear probability models, according to equation 7:

$$Y_{ijk} = \alpha + \beta leeway\kappa_{ijk} + \gamma_1 i + \gamma_2 s_{ijk} + u_{jk} + \varepsilon_{ijk}, \tag{7}$$

[26] Makofske (2023) documents that in Las Vegas, food-service health inspectors underreported minor violations when those violations were likely to affect letter-grade outcomes. However, such manipulation by inspectors is unlikely to be feasible in the context of the FSIS *Salmonella* Verification Testing Program. 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 19.

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_{jk} represents establishment–sample-set fixed effects; and ε_{ijk} is the residual. Establishment–sample-set fixed effects control for factors that may affect test outcomes at an establishment over a narrow temporal window, such as fixed technology or biological factors common to the chickens supplied.

Admittedly, there are some shortcomings in the identification strategy described here, given that $leeway\kappa_{ijk}$ is (mechanically and empirically) negatively correlated with the share of samples positive s_{ijk} and positively correlated with the sample 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 $leeway\kappa$ is plausibly exogenous. Including s_{ijk} and i as regressors allows me to tease out effects of distance from the threshold on *Salmonella* control efforts. Moreover, my empirical results are consistent whether or not s_{ijk} is included as a regressor.

6.2. Results: Distance from thresholds and Salmonella test outcomes

Table 3 presents results from regressions of the form described by equation 7, which demonstrate the effect of distance from the thresholds on *Salmonella* test outcomes. Table 3 demonstrates that in all periods, when the value of $leeway\kappa$ was larger—that is, when establishments were further from thresholds—, carcasses were more likely to test positive for *Salmonella*.[27] In other words, establishments controlled *Salmonella* better when it was critical to do so to ensure a better categorization outcome. These results hold regardless of whether the policy of public disclosure of Category 2 and 3 outcomes was in place. I now review the results in more detail.

Panels A and B of table 3 report results for the regressions with the *leeway* variables defined with respect to the Category 1/2 and 2/3 thresholds, respectively.[28] 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 had more leeway with respect to both the 6- and 12-positive-sample thresholds. When the $leeway12$ value was 10 percentage points higher, the probability of a positive test result was 1.97 percentage points higher ($p < 0.001$; panel B, column 2). The elasticity of the share of samples positive with respect to $leeway12$ was 0.57, calculated using the mean share of samples positive and the mean value of $leeway12$.

From 2006 to 2008, when categorization was reported privately, the relationship between distance from the 12-positives threshold and *Salmonella* test outcomes was slightly stronger than in the previous period. When the $leeway6$ value was 10 percentage points higher, the probability of a positive test result was 1.30 percentage points higher ($p < 0.001$; elasticity = 0.28; panel A, column 4), and when the $leeway12$ value by 10 percentage points higher, the probability of a positive test result was 2.24 percentage points

[27] It is important to note that the results in this section are conditional on using establishment–sample-set fixed effects. Without using any fixed effects, the correlation between $leeway\kappa$ and subsequent positive samples was negative or insignificant in all periods before 2011, as suggested by figures 5 and 6.

[28] All discussion of results in table 3 references the even-numbered columns, as they are the preferred specifications.

higher ($p < 0.001$; elasticity = 0.90; panel B, column 4).

Public disclosure of the names of both Category 2 and 3 establishments from 2008–11 further strengthened the relationship between distance from 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 1.56 percentage points higher ($p < 0.001$; elasticity = 0.60; panel A, column 6), and when the *leeway*12 value was 10 percentage points higher, the probability of a positive test result was 2.79 percentage points higher ($p < 0.001$; elasticity = 1.70; 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 0.35 percentage points higher ($p = 0.004$; elasticity = 0.17; 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 was 0.78 percentage points higher ($p < 0.001$; elasticity = 0.71; 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. Distance from thresholds mattered whether or not there was a threat of public disclosure. Distance from the more lenient threshold also mattered much more than distance from the stringent threshold. Across all periods, the elasticity of the share of samples positive with respect to leeway was 2.8 to 4.2 times larger when considering the Category 2/3 threshold than the Category 1/2 threshold. The effect size increased when categorization and public disclosure were introduced, but the tightening of standards in 2011 reduced the effect. In short, the introduction of both categorization and public disclosure seems to have changed the extent to which establishment operators paid attention to the thresholds, and increased their efforts accordingly.

7. 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 5. 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 distance from 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.

7.1. 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 during a relatively narrow window around each policy change. As in section 5, 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, Cattaneo, and Titiunik, 2014b; Cattaneo, Idrobo, and Titiunik, 2020b; Cattaneo, Titiunik, and Vazquez-Bare, 2020). The regression equation is as follows:

$$Y_{ikt} = \alpha + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + f(t) + \varepsilon_{ikt}. \tag{8}$$

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, Idrobo, and Titiunik (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.

As noted by Winship and Morgan (1999) and Morgan and Winship (2015, p. 356), the key identifying assumption for the RDiT approach (which they refer to as interrupted time series) is that observations from before the cutoff point can be used to predict what the outcome variable would have been, in the absence of a treatment, in post-cutoff periods. One common reason that assumption may fail is anticipation effects—for example, before a policy change that imposes a stricter standard, establishment operators may adjust their operations (Baicker and Svoronos, 2019). If this occurred among chicken-slaughter establishments before any of the policy changes, then my RDiT estimates would understate the overall effects of the policy changes.[29] Moreover, the RDiT approach estimates differences in outcomes immediately before and after the dates of policy changes, so if responses to the policy changes were delayed, my estimates would again understate the total effects (Shadish, Cook, and Campbell, 2002).[30]

7.2. *Results: Effects of policy changes*

Panel A of table 4 presents results from the RDiT model described by equation 8 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

[29] With anticipation effects in mind, appendix B presents results for regulations that use the dates of Federal Register announcements about policy changes as the cutoffs; the announcement dates had insignificant effects on *Salmonella* test outcomes. Appendix B also presents various robustness and placebo tests for the RDiT analysis.

[30] The preceding comments about understated effects assume that all responses to a given policy change have the same effect sign.

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.[31] Figure 7 depicts RD plots that correspond with panel B of table 4. 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 5 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. 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 results during 2008–11 that would place them in the new Category 2 (3 to 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.

[31] 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.

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 increased incentives to shirk, worsened test outcomes, and more than offset the earlier improvement.

8. Summary and conclusion

Using carcass-level data on USDA inspections for *Salmonella* in chicken carcasses, this paper 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 when establishments failed to meet thresholds associated with better categorization outcomes that were not associated with disclosure, their performance on subsequent *Salmonella* tests within the same sample set worsened, a result I describe as shirking. There is also some evidence that prior to public disclosure, establishments shirked after sustained good performance. During the initial public disclosure period, there was no evidence of shirking after either good or poor performance resulted in known categorization outcomes. One of several possible explanations is that buyers demanded that potential suppliers provide additional information about *Salmonella* test results, beyond what was publicly disclosed, and that these demands limited incentives to shirk during the initial disclosure period. In addition, when the threshold that triggered disclosure was tightened in 2011, establishments were likely to shirk after poor performance ensured 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. Disclosure of discrete information may be more readily understood by buyers, particularly if the buyers are final consumers. 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. From a different perspective, inspected entities have incentives to just barely meet categorical thresholds (Makofske, 2023), but may shirk if thresholds are not met. 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 (until standards were tightened). 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, especially when intermediary buyers are expected to be the parties mainly interested in knowing the outcomes, 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. The identification strategy used in section 6 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 could 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

Adams, N.R., and G.R. Waddell. 2018. "Performance and risk taking under threat of elimination." *Journal of Economic Behavior & Organization* 156:41–54.

Antle, J.M. 2001. "Economic Analysis of Food Safety." *Handbook of Agricultural Economics* 1:1083–1136.

—. 1996. "Efficient food safety regulation in the food manufacturing sector." *American Journal of Agricultural Economics* 78:1242–1247.

Baicker, K., and T. Svoronos. 2019. "Testing the validity of the single interrupted time series design." Working paper No. 26080, National Bureau of Economic Research.

Barahona, N., C. Otero, and S. Otero. 2023. "Equilibrium effects of food labeling policies." *Econometrica* 91:839–868.

Barreca, A.I., M. Guldi, J.M. Lindo, and G.R. Waddell. 2011. "Saving babies? Revisiting the effect of very low birth weight classification." *The Quarterly Journal of Economics* 126:2117–2123.

Bederson, B.B., G.Z. Jin, P. Leslie, A.J. Quinn, and B. Zou. 2018. "Incomplete disclosure: Evidence of signaling and countersignaling." *American Economic Journal: Microeconomics* 10:41–66.

Belay, D.G., and J.D. Jensen. 2020. "'The scarlet letters': information disclosure and self-regulation: evidence from antibiotic use in Denmark." *Journal of Environmental Economics and Management* 104:102385.

Bennear, L.S., and S.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.

Boubendir, S., J. Arsenault, S. Quessy, A. Thibodeau, P. Fravalo, W.P. Thériault, S. Fournaise, and M.L. Gaucher. 2021. "Salmonella contamination of broiler chicken carcasses at critical steps of the slaughter process and in the environment of two slaughter plants: prevalence, genetic profiles, and association with the final carcass status." *Journal of Food Protection* 84:321–332.

Bovay, J. 2023. "Food safety, reputation, and regulation." *Applied Economic Perspectives and Policy* 45:684–704.

Buncic, S., and J. Sofos. 2012. "Interventions to control Salmonella contamination during poultry, cattle and pig slaughter." *Food Research International* 45:641–655.

Calonico, S., M.D. Cattaneo, and R. Titiunik. 2015. "Optimal data-driven regression discontinuity plots." *Journal of the American Statistical Association* 110:1753–1769.

—. 2014a. "Robust data-driven inference in the regression-discontinuity design." *The Stata Journal* 14:909–946.

—. 2014b. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82:2295–2326.

Campa, P. 2018. "Press and leaks: Do newspapers reduce toxic emissions?" *Journal of Environmental Economics and Management* 91:184–202.

Caswell, J.A., and E.M. Mojduszka. 1996. "Using informational labeling to influence the market for quality in food products." *American Journal of Agricultural Economics* 78:1248–1253.

Cattaneo, M.D., N. Idrobo, and R. Titiunik. 2020a. *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge: Cambridge University Press.

—. 2020b. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge: Cambridge University Press.

Cattaneo, M.D., M. Jansson, and X. Ma. 2018. "Manipulation testing based on density discontinuity." *The Stata Journal* 18:234–261.

Cattaneo, M.D., R. Titiunik, and G. Vazquez-Bare. 2020. "Analysis of regression discontinuity designs with multiple cutoffs or multiple scores." *The Stata Journal* 20:866–891.

Dai, W., and M. Luca. 2020. "Digitizing disclosure: The case of restaurant hygiene scores." *American Economic Journal: Microeconomics* 12:41–59.

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

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

Dranove, D., and G.Z. Jin. 2010. "Quality disclosure and certification: Theory and practice." *Journal of Economic Literature* 48:935–63.

Dranove, D., D. Kessler, M. McClellan, and M. Satterthwaite. 2003. "Is more information better? The effects of "report cards" on health care providers." *Journal of Political Economy* 111:555–588.

Eggers, A.C., A. Fowler, J. Hainmueller, A.B. Hall, and J.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* 59:259–274.

Evans, M.F. 2016. "The clean air act watch list: An enforcement and compliance natural experiment." *Journal of the Association of Environmental and Resource Economists* 3:627–665.

Golan, E.H., T. Roberts, E. Salay, J.A. Caswell, M. Ollinger, and D.L. Moore. 2004. "Food safety innovation in the United States: Evidence from the meat industry." USDA Economic Research Service Agricultural Economic Report No. 831.

Grant, D. 2016. "The essential economics of threshold-based incentives: Theory, estimation, and evidence from the Western States 100." *Journal of Economic Behavior & Organization* 130:180–197.

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

Henley, S.C., J. Gleason, J.J. Quinlan, et al. 2016. "Don't wash your chicken!: A food safety education campaign to address a common food mishandling practice." *Food Protection Trends* 36:43–53.

Hoffmann, S.A., B. Maculloch, and M. 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, S. 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.

Interagency Food Safety Analytics Collaboration. 2021. "Foodborne illness source attribution estimates for 2019 for *Salmonella*, *Escherichia coli* O157, *Listeria monocytogenes* and *Campylobacter* using multi-year outbreak surveillance data, United States." https://www.cdc.gov/foodsafety/ifsac/pdf/P19-2019-report-TriAgency-508.pdf.

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

Jacob, B.A. 2005. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89:761–796.

Jacob, B.A., and S.D. Levitt. 2003. "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *The Quarterly Journal of Economics* 118:843–877.

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

—. 2009. "Reputational incentives for restaurant hygiene." *American Economic Journal: Microeconomics* 1:237–67.

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

Josling, T.E., D. Roberts, and D. Orden. 2004. *Food regulation and trade: toward a safe and open global system*. Peterson Institute for International Economics.

Kleven, H.J., and M. Waseem. 2013. "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan." *The Quarterly Journal of Economics* 128:669–723.

Knoeber, C.R. 1989. "A real game of chicken: Contracts, tournaments, and the production of broilers." *The Journal of Law, Economics, and Organization* 5:271–292.

Knoeber, C.R., and W.N. Thurman. 1995. "Don't Count Your Chickens...: Risk and Risk Shifting in the Broiler Industry." *American Journal of Agricultural Economics* 77:486–496.

—. 1994. "Testing the theory of tournaments: An empirical analysis of broiler production." *Journal of Labor Economics* 12:155–179.

Korting, C., C. Lieberman, J. Matsudaira, Z. Pei, and Y. Shen. 2023. "Visual inference and graphical representation in regression discontinuity designs." *The Quarterly Journal of Economics*, pp. qjad011.

Lazear, E.P., and S. Rosen. 1981. "Rank-order tournaments as optimum labor contracts." *Journal of Political Economy* 89:841–864.

Lee, D.S., and T. Lemieux. 2010. "Regression discontinuity designs in economics." *Journal of Economic Literature* 48:281–355.

Lemus, J., and G. Marshall. 2021. "Dynamic tournament design: Evidence from prediction contests." *Journal of Political Economy* 129:383–420.

MacDonald, J.M. 2015. "Trends in agricultural contracts." *Choices* 30:1–6.

Makofske, M.P. 2023. "Disclosure policy design and regulatory agent behavior." *American Journal of Agricultural Economics*, forthcoming.

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

Morgan, S.L., and C. Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.

Ollinger, M., and J. Bovay. 2018. "Pass or Fail: Economic Incentives to Reduce *Salmonella* Contamination in Ground Beef Sold to the National School Lunch Program." *American Journal of Agricultural Economics* 100:414–433.

—. 2020. "Producer response to public disclosure of food-safety information." *American Journal of Agricultural Economics* 102:202–219.

Page, E.T. 2018. "Trends in food recalls: 2004-13." Economic Information Bulletin Number 191, United States Department of Agriculture, Economic Research Service.

Painter, J.A., R.M. Hoekstra, T. Ayers, R.V. Tauxe, C.R. Braden, F.J. Angulo, and P.M. Griffin. 2013. "Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998–2008." *Emerging Infectious Diseases* 19:407–415.

Park, S.H., M. Aydin, A. Khatiwara, M.C. Dolan, D.F. Gilmore, J.L. Bouldin, S. Ahn, and S.C. Ricke. 2014. "Current and emerging technologies for rapid detection and characterization of Salmonella in poultry and poultry products." *Food Microbiology* 38:250–262.

Pope, N.G. 2019. "The effect of teacher ratings on teacher performance." *Journal of Public Economics* 172:84–110.

Pouliot, S., and H.H. Wang. 2018. "Information, incentives, and government intervention for food safety." *Annual Review of Resource Economics* 10:83–103.

Saez, E. 2010. "Do taxpayers bunch at kink points?" *American Economic Journal: Economic Policy* 2:180–212.

Sallee, J.M., and J. Slemrod. 2012. "Car notches: Strategic automaker responses to fuel economy policy." *Journal of Public Economics* 96:981–999.

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

Shadish, W., T.D. Cook, and D.T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shapiro, C. 1983. "Premiums for high quality products as returns to reputations." *The Quarterly Journal of Economics* 98:659–679.

Shewmake, S., and W.K. Viscusi. 2015. "Producer and consumer responses to green housing labels." *Economic Inquiry* 53:681–699.

United States Department of Agriculture—Economic Research Service. 2023. "Meat supply and disappearance tables, recent." https://www.ers.usda.gov/webdocs/DataFiles/104360/MeatSDRecent.xlsx?v=2313.8, accessed June 12, 2023.

United States Department of Agriculture—Food Safety and Inspection Service. 2021. "A Generic HACCP Model for Poultry Slaughter." https://www.fsis.usda.gov/sites/default/files/media_file/2021-03/HACCP-Model-for-Poultry-Slaughter.pdf.

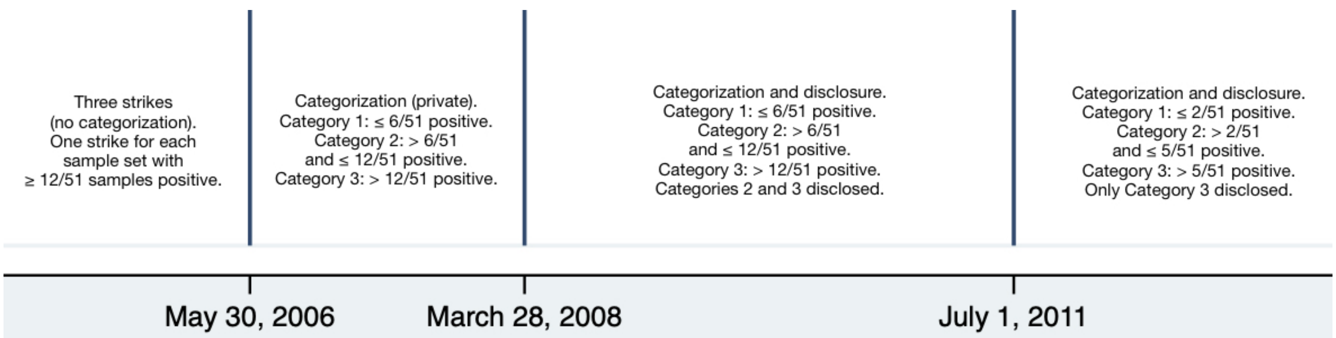
USDA. 2019. "2017 Census of Agriculture." https://www.nass.usda.gov/Publications/AgCensus/2017/.

Winship, C., and S.L. Morgan. 1999. "The estimation of causal effects from observational data." *Annual Review of Sociology* 25:659–706.

World Health Organization. 2015. *WHO estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015*. Geneva: World Health Organization.

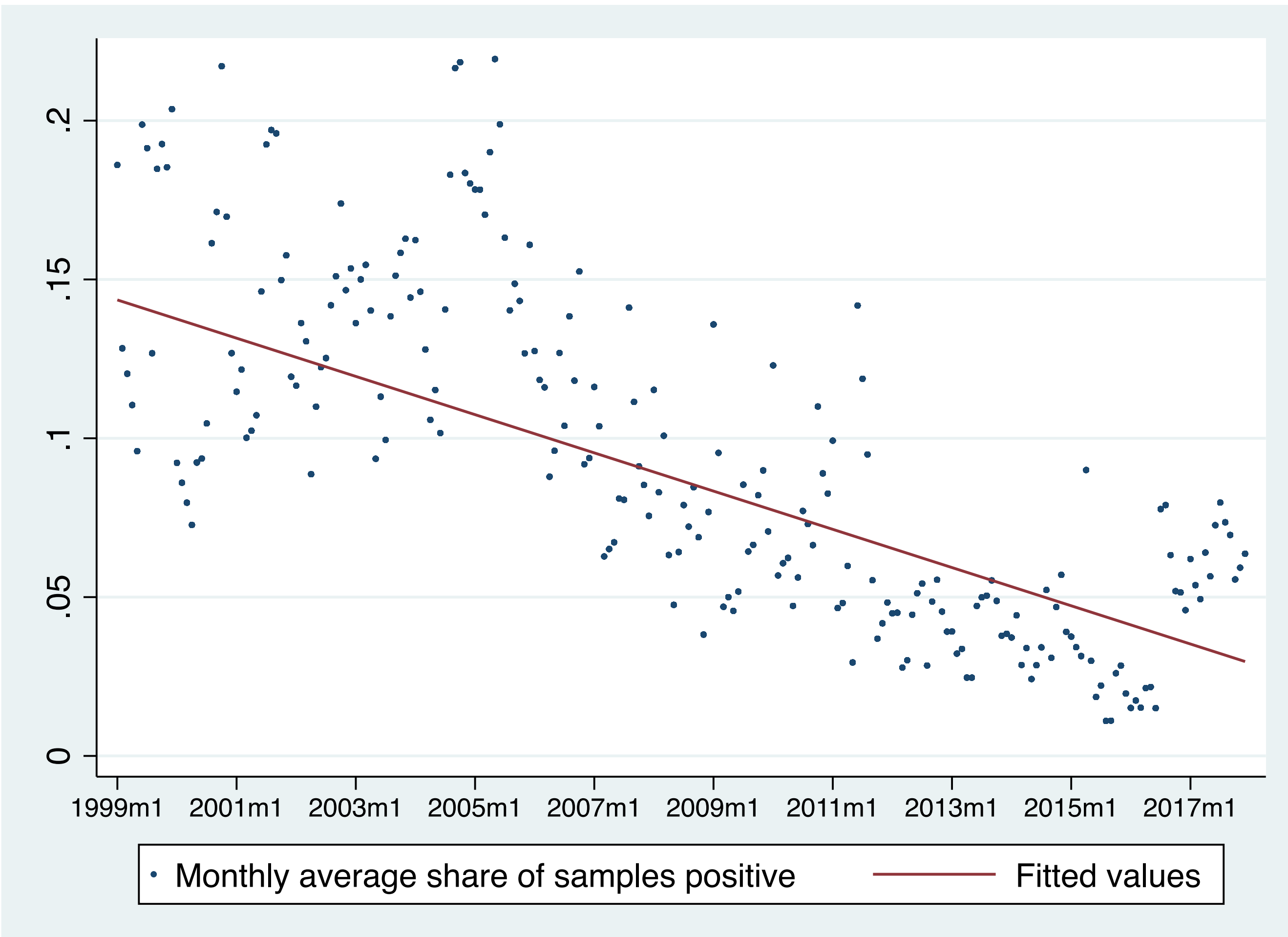
Zeng, H., K. De Reu, S. Gabriël, W. Mattheus, L. De Zutter, and G. Rasschaert. 2021. "*Salmonella* prevalence and persistence in industrialized poultry slaughterhouses." *Poultry Science* 100:100991.

Figure 1: Policy regimes and dates of implementation



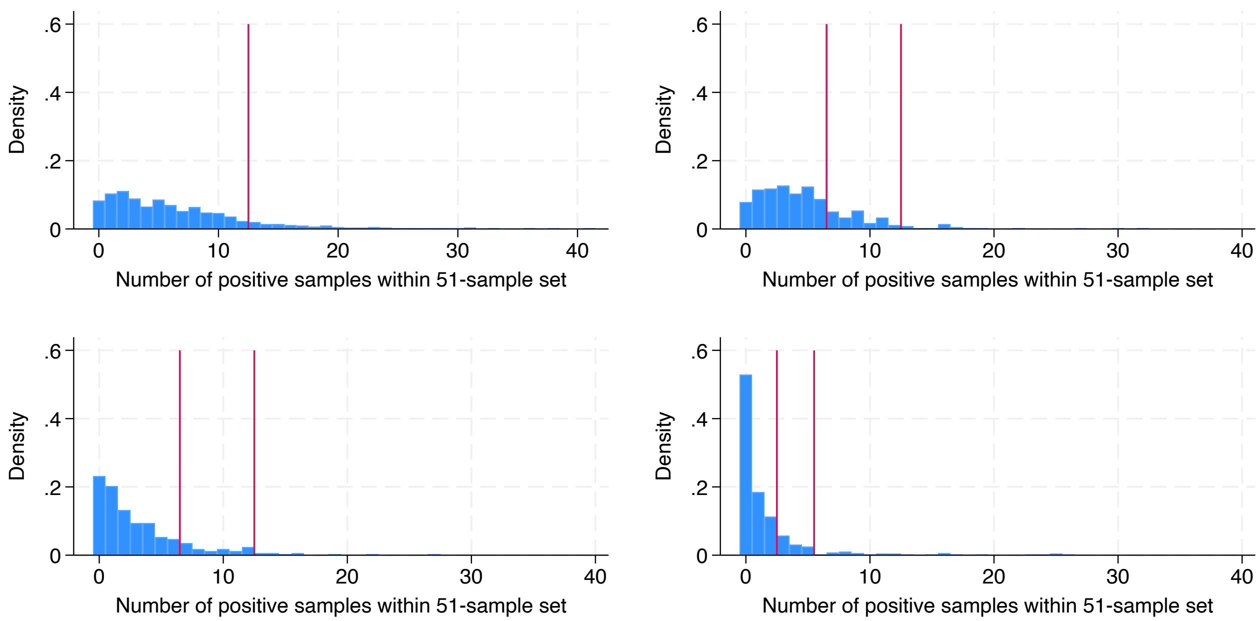
Notes: FSIS *Salmonella* testing began prior to 1999 and is still ongoing. Additional, later, policy changes are discussed in Appendix C.

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



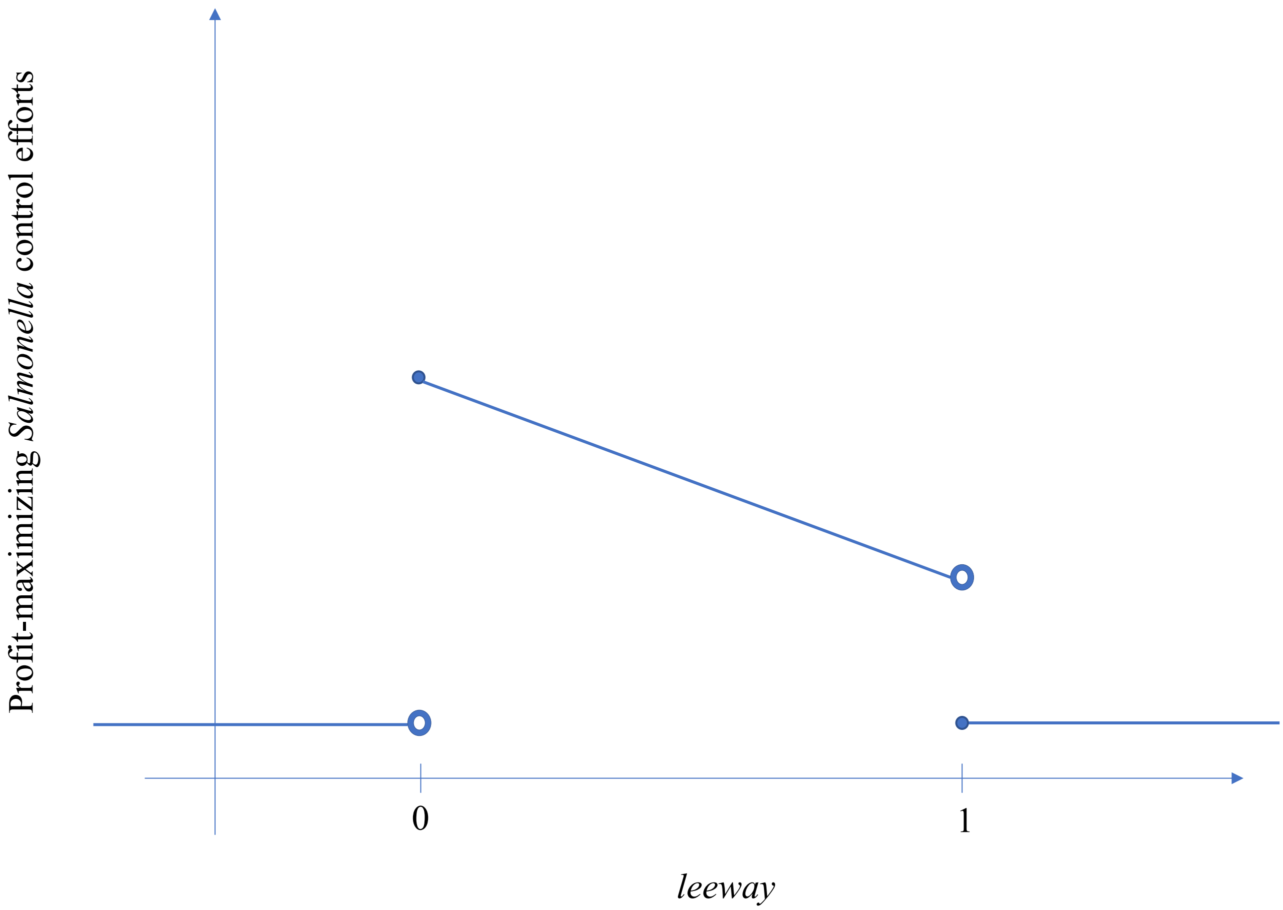
Note: OLS regression is fitted to monthly average data.

Figure 3: Histograms of the number of positive samples per sample set, by policy period



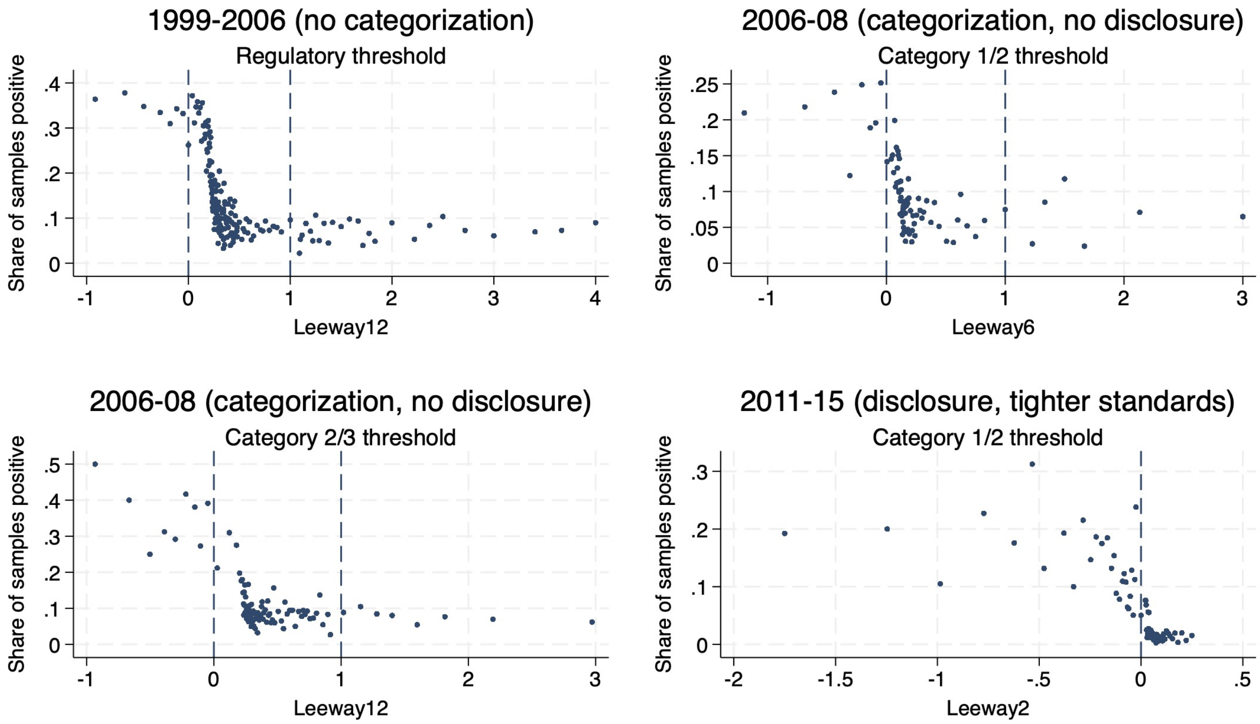
Notes: Each panel represents the density of the number of positive samples per 51-sample set, for each policy period. Vertical lines represent the regulatory threshold (until 2006) and the category thresholds (starting in 2006). Number of 51-sample-set observations per period: 1,204 (1999 to 2006); 357 (2006 to 2008); 341 (2008 to 2011); 488 (2011 to 2015).

Figure 4: Motivating the analysis of moral hazard



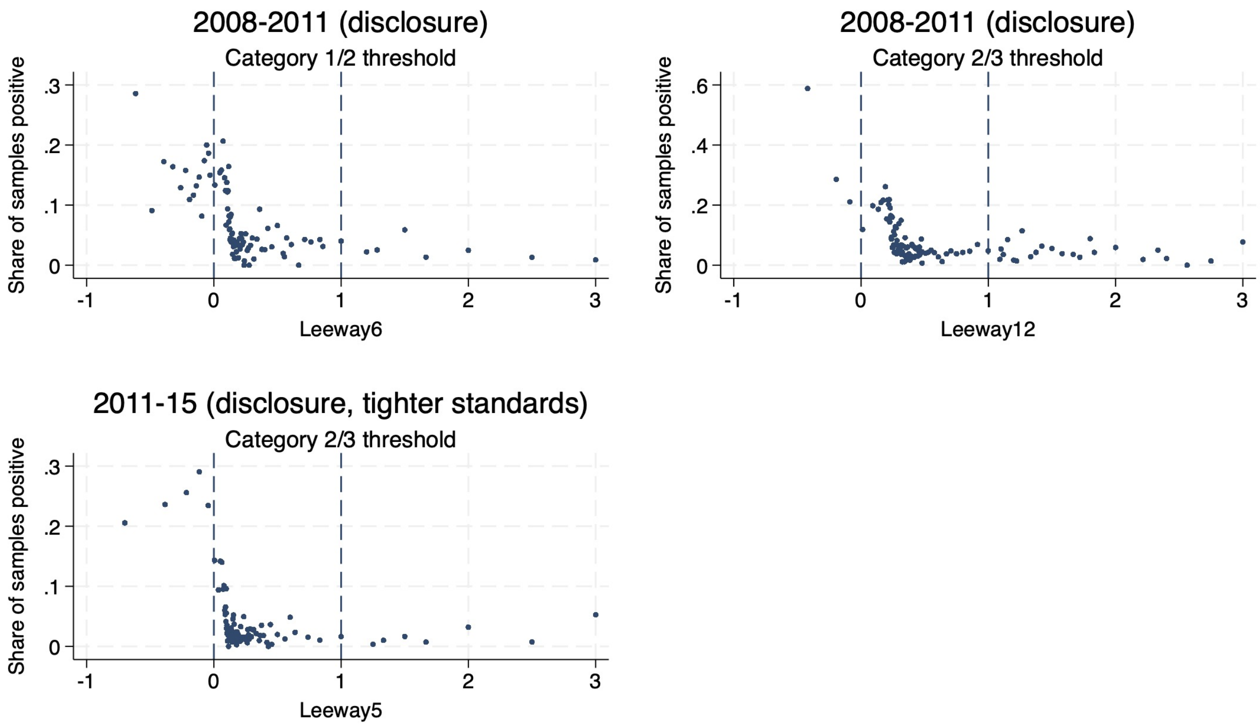
Notes: This figure illustrates how having more leeway with respect to the categorization threshold affects the incentives for establishments to control *Salmonella*. (In this paper, *leeway* is the share of remaining samples within a sample set that may test positive if an establishment is to achieve a certain categorization; categorization is determined by the total number of samples testing positive within the sample set.) 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.

Figure 5: RD plots: effects of known categorization (cutoffs not associated with disclosure) on *Salmonella* outcomes



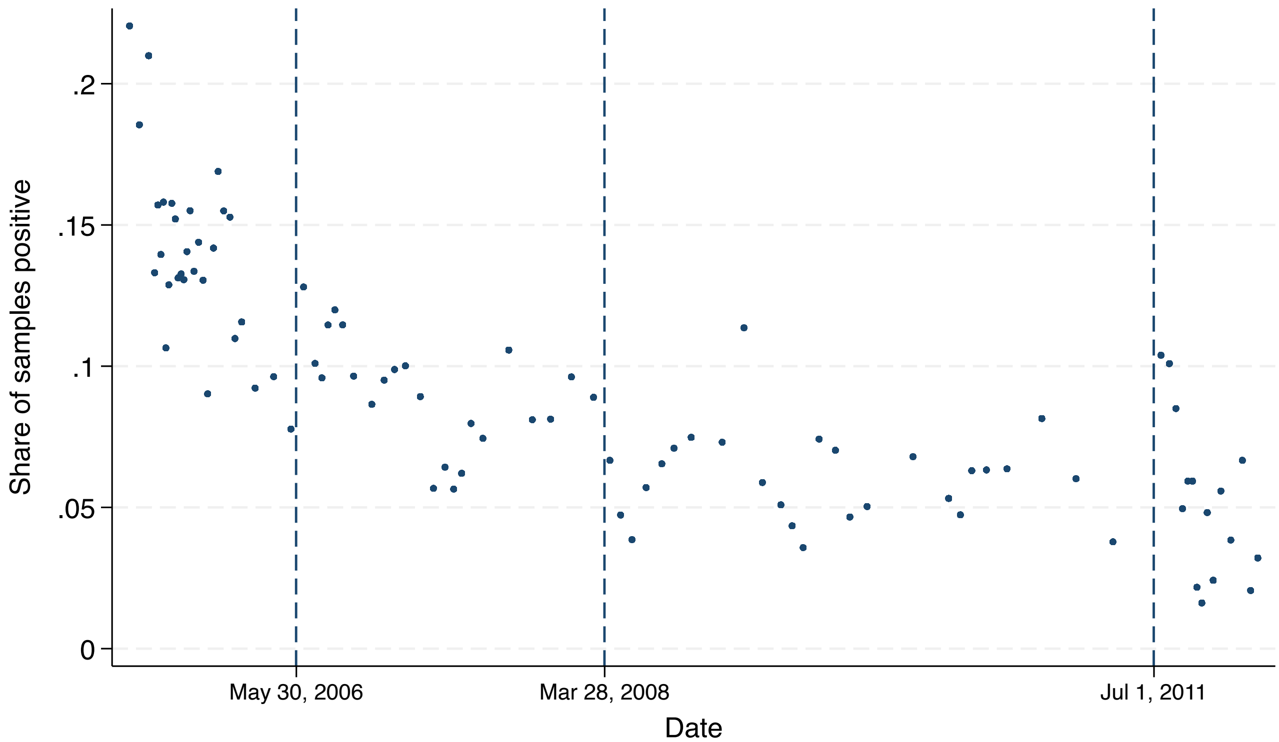
Notes: These RD plots provide graphical evidence corresponding with panel A in table 2, within the ranges of the running variables that correspond with the MSE-optimal bandwidths used in table 2. Quantile-spaced bins are generated using integrated MSE-optimal spacings estimators (Calonico, Cattaneo, and Titiunik, 2014a, 2015). Fit lines are not included because they tend to increase the type I error rate of visual inference (Korting et al., 2023). The $leeway2 = 1$ threshold is not shown in the lower-right plot because $leeway2$ takes on only two different values at and above the cutoff so the RD model cannot be estimated around this threshold.

Figure 6: RD plots: effects of known categorization (cutoffs associated with disclosure) on *Salmonella* outcomes



Notes: These RD plots provide graphical evidence corresponding with panel B in table 2. See additional notes to figure 5.

Figure 7: RD in time plot



Notes: This RD plot provides graphical evidence corresponding with panel B in table 4, within the temporal range that corresponds with the MSE-optimal bandwidths used in table 4. As in figures 5 and 6, quantile-spaced bins are generated using integrated MSE-optimal spacings estimators (Calonico, Cattaneo, and Titiunik, 2014a, 2015), and fit lines are not included because they tend to increase the type I error rate of visual inference (Korting et al., 2023).

Table 1: Number of sample sets by category and average share of samples positive for *Salmonella*, by period, 1999–2015

Policy regime	No categorization		Categorization (private)		Public disclosure		Public disclosure w/ tighter standards	
Years	1999 to 2006		2006 to 2008		2008 to 2011		2011 to 2015	
Number of Category 1 sets (share)	726	.603	266	.749	304	.851	428	.821
Number of Category 2 sets (share)	325	.270	71	.200	43	.120	61	.117
Number of Category 3 sets (share)	152	.126	18	.051	10	.028	32	.061
Share of samples positive (Number of obs.)	.129	71398	.100	20406	.068	21478	.042	35083

Notes: Number of sample sets in each category reflects sets ending in the period indicated. For 12 sets that had fewer than 51 observations, categorization could not be assigned. Since categorization was not in place until 2006, the number of sets in each category for 1999 to 2006 are listed based on the criteria in place from 2006 to 2011.

Table 2: Effects of known categorization on *Salmonella* test 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
Threshold	$D_0 = 1$	$D_1 = 1$	$D_0 = 1$	$D_1 = 1$	$D_0 = 1$	$D_1 = 1$	$D_0 = 1$
Implication	Fails std.	Meets std.	Cat. 2 or 3	Cat. 1	Cat. 3	Cat. 1 or 2	Cat. 2 or 3
Max. # pos. samples (κ)	12	12	6	6	12	12	2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Known categorization ($D_0 = 1$ or $D_1 = 1$)	0.034	0.029	0.084	0.049	0.149	0.102	0.078
Robust p-value	0.054	0.300	0.086	0.608	0.024	0.007	0.008
95% CI (lower limit)	−0.00	−0.03	−0.01	−0.07	0.03	0.03	0.01
(upper limit)	0.13	0.11	0.12	0.04	0.45	0.21	0.10
Observations	6623	8464	16153	15426	13231	2733	24546
Left bandwidth	1.12	0.19	1.66	1.00	1.02	0.18	2.02
Right bandwidth	0.21	3.15	0.59	2.92	0.64	2.34	0.26

Panel B: Cutoffs associated with disclosure

Policy regime	Public disclosure				Public disclosure w/ tighter standards	
Years	2008 to 2011				2011 to 2015	
Threshold	$D_0 = 1$	$D_1 = 1$	$D_0 = 1$	$D_1 = 1$	$D_0 = 1$	$D_1 = 1$
Implication	Cat. 2 or 3	Cat. 1	Cat. 3	Cat. 1 or 2	Cat. 3	Cat. 1 or 2
Max. # pos. samples (κ)	6	6	12	12	5	5
	(1)	(2)	(3)	(4)	(5)	(6)
Known categorization ($D_0 = 1$ or $D_1 = 1$)	0.015	0.021	−0.085	−0.018	0.158	0.001
Robust p-value	0.710	0.179	0.824	0.633	0.045	0.139
95% CI (lower limit)	−0.08	−0.08	−0.21	−0.09	0.00	−0.05
(upper limit)	0.06	0.02	0.17	0.05	0.21	0.01
Observations	14512	16051	5339	3380	18969	25531
Left bandwidth	0.70	0.98	0.61	0.23	0.95	1.00
Right bandwidth	0.42	2.56	0.29	2.15	0.26	2.26

Notes: Each pair or quartet of columns represents regressions using carcass-level 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 in the later period if the samples were taken after the Federal Register announcement of the policy change. All regressions are local linear RD regressions with triangular kernels, using $leeway\kappa$ as the running variable, as described in the text. $D_0 = 1$ if $leeway\kappa < 0$ and $D_1 = 1$ if $leeway\kappa \geq 1$; each of these conditions are equivalent to the known categorization outcomes reflected by the Implication rows in the table. Bandwidths, robust p-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo, Titiunik, and Vazquez-Bare, 2020), 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. The RD models cannot be estimated for $D_1 = 1$ and $\kappa = 2$ for 2011–15 because the running variable takes on too few different values at and above the cutoff $leeway\kappa = 1$ (namely, the only possible values are 1 and 2).

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

Policy regime Years	No categorization 1999 to 2006		Categorization (private) 2006 to 2008		Public disclosure 2008 to 2011		Public disclosure w/ tighter standards 2011 to 2015	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A								
Distance from Cat. 1 threshold	0.124	0.105	0.141	0.130	0.156	0.156	0.0404	0.0349
	(0.015)	(0.015)	(0.023)	(0.025)	(0.021)	(0.022)	(0.011)	(0.012)
Test number, current sample set	-0.000224	-0.000139	-0.000327	-0.000305	-0.00144	-0.00157	-0.0000336	-0.0000137
	(0.00015)	(0.00017)	(0.00025)	(0.00030)	(0.00026)	(0.00030)	(0.00011)	(0.00012)
Share of samples positive, current sample set		-0.263		-0.354		-0.284		-0.273
		(0.024)		(0.063)		(0.058)		(0.041)
Observations	49064	47860	15351	15021	15427	15086	23917	23393
Elasticity	0.20	0.18	0.31	0.28	0.59	0.60	0.19	0.17
Panel B								
Distance from Cat. 2 threshold	0.211	0.197	0.227	0.224	0.273	0.279	0.0847	0.0776
	(0.017)	(0.018)	(0.036)	(0.042)	(0.034)	(0.037)	(0.014)	(0.014)
Test number, current sample set	-0.00239	-0.00232	-0.00259	-0.00270	-0.00424	-0.00453	-0.000872	-0.000807
	(0.00024)	(0.00027)	(0.00053)	(0.00064)	(0.00056)	(0.00063)	(0.00017)	(0.00018)
Share of samples positive, current sample set		-0.197		-0.265		-0.197		-0.168
		(0.024)		(0.055)		(0.059)		(0.033)
Observations	50771	49567	14633	14303	14400	14059	24009	23485
Elasticity	0.60	0.57	0.90	0.90	1.63	1.70	0.74	0.71

Notes: Panel A demonstrates the effects of distance from the Category 1 thresholds (i.e., the value of *leeway*6 from 1999 to 2011 and *leeway*2 from 2011 to 2015) on *Salmonella* test outcomes. Panel B demonstrates the effects of distance from the Category 2 thresholds (i.e., the value of *leeway*12 from 1999 to 2011 and *leeway*5 from 2011 to 2015). Horizontally, each pair of columns represents regressions using carcass-level 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 in the later period if the samples were taken after the Federal Register announcement of the policy change. All regressions use establishment–sample-set 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 $leeway\kappa$, calculated using the mean share of samples positive and the mean value of $leeway\kappa$. Observations are included only if $leeway\kappa \in [0, 1)$.

Table 4: 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	0.501	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 use carcass-level observations; the dependent variable is binary with a value of 1 if the *Salmonella* test result is positive. The 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, Titiunik, and Vazquez-Bare, 2020). Bandwidths are chosen to minimize mean squared error on either side of each cutoff.

Table 5: 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 use carcass-level observations; the dependent variable is binary with a value of 1 if the *Salmonella* test result is positive. The 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, Titiunik, and Vazquez-Bare, 2020). 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 5 and 6, 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: Placebo tests

Table B1 presents the results of RD models that use placebo cutoffs near the $c = 0$ and $c = 1$ cutoffs that yield significant estimates in table 2. As discussed in the main body of the paper, only three of the 36 cutoffs in table B1 have statistically significant coefficients with $p < 0.1$ and with the correct, positive sign; the distribution of p-values is approximately uniform. In conclusion, the placebo cutoffs do not raise concerns about the validity of the main results.

Effects of policy changes: robustness tests

This appendix subsection presents the results of robustness tests relevant to the RDiT design discussed in section 7. For RDiT approaches to analysis of policy changes, Hausman and Rapson (2018) recommend a few additional robustness tests.[32] First, as recommended by Cattaneo, Idrobo, and Titiunik (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 B2, essentially conform with the results in panel B of table 4: 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 4, 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 4. 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 table B2 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

[32] All results described in the rest of this appendix use the same data set as panel B of table 4, dropping all establishments that ever exited.

observations from sample sets that span policy periods, the estimated RDiT effects change somewhat: the introduction of disclosure in 2008 resulted in a 2.9 percentage point decrease in the share of samples positive (though not statistically significant), while the 2011 tightening of standards led to an 11.1 percentage point increase ($p = 0.002$). 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 table B3, 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 table B3, 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 table B3, 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 4 are valid.

Table B1: Placebo effects of known categorization on *Salmonella* outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Placebo RD cutoff (c)	−0.15	−0.1	−0.05	0.05	0.1	0.15
Panel A: 1999 to 2006, $\kappa = 12$ positive samples						
$D_0^p = 1$	−0.021	0.016	0.076	−0.063	−0.053	0.001
Robust p-value	0.554	0.427	0.107	0.009	0.010	0.844
Panel B: 2006 to 2008, $\kappa = 6$ positive samples						
$D_0^p = 1$	0.024	−0.030	0.001	−0.016	0.004	−0.009
Robust p-value	0.437	0.195	0.597	0.317	0.866	0.235
Panel C: 2006 to 2008, $\kappa = 12$ positive samples						
$D_0^p = 1$	0.070	0.070	0.018	−0.096	−0.158	−0.071
Robust p-value	0.221	0.194	0.924	0.096	0.041	0.262
Panel D: 2011 to 2015, $\kappa = 2$ positive samples						
$D_0^p = 1$	0.059	0.011	−0.018	0.004	−0.013	−0.023
Robust p-value	0.086	0.807	0.110	0.345	0.002	0.000
Panel E: 2011 to 2015, $\kappa = 5$ positive samples						
$D_0^p = 1$	−0.056	−0.022	0.102	0.001	0.035	−0.011
Robust p-value	0.181	0.331	0.430	0.796	0.000	0.019
Placebo RD cutoff (c)	0.8	0.85	0.9	1.05	1.1	1.15
Panel F: 2006 to 2008, $\kappa = 12$ positive samples						
$D_1^p = 1$	−0.007	0.014	0.008	0.056	0.010	−0.011
Robust p-value	0.827	0.517	0.761	0.042	0.832	0.528

Notes: This table presents results of regressions paralleling those in table 2 with statistically significant results but for placebo cutoffs not associated with any change in disclosure status. D_0^p and D_1^p are analogous to D_0 and D_1 in table 2 but use the placebo cutoffs indicated at the top of the columns. Panels A through E report results for three placebo cutoffs on either side of the actual cutoff ($c = 0$) according to $c \pm 0.05n$, where $n = \{1, 2, 3\}$. Panel F uses the nearest placebo cutoffs to the actual cutoff ($c = 1$) that are multiples of 0.05, for which there are enough observations on either side of the placebo cutoffs to estimate the optimal bandwidths. 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, observations are included in the later period if the samples were taken after the Federal Register announcement that preceded the policy change. All regressions are local linear RD regressions with triangular kernels, using *leeway*κ as the running variable. Bandwidths, robust p-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo, Titiunik, and Vazquez-Bare, 2020), although bandwidths and confidence intervals are suppressed in this table. Bandwidths are chosen to minimize mean squared error on either side of each cutoff.

Table B2: 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			
$t \geq c$	0.034	−0.029	0.111
Robust p-value	0.248	0.385	0.002
95% CI (lower limit)	−0.03	−0.11	0.05
(upper limit)	0.13	0.04	0.21
Observations	11540	3625	3692
Left bandwidth	345	148	177
Right bandwidth	212	199	258

Notes: See notes to table 4.

Table B3: 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			
$t \geq c$	−0.038	−0.031	0.015
Robust p-value	0.640	0.883	0.269
95% CI (lower limit)	−0.15	−0.10	−0.01
(upper limit)	0.09	0.08	0.05
Observations	9621	3289	5747
Left bandwidth	356	172	165
Right bandwidth	89	60	139
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 4.

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, 2015. 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.[33] Moreover, the rolling-window system seemed it would 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.[34] After this change, which was implemented on July 1, 2016, the share of positive test results rapidly rose, and on November 20, 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.

Tables C1, C3, and C4 in this appendix present the results of regression models equivalent to those in tables 2, 3 and 4, covering the periods 2015–16 (rolling windows) and 2016–17 (disclosure hiatus). Figures C1 and C2 present RD plots that correspond to tables C1 and C4. Tables C2, C5, and C6 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

RD plots for the 2015–16 and 2016–17 periods are shown in figure C1, and regression results equivalent to those shown in table 2 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 the FSIS website. Similar to the 2008–

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

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

11 period, establishment operators apparently exerted effort to meet the Category 1 standard but then reduced effort once exceeding the threshold. Establishments were 4.1 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). In addition, during this period, establishments appear to have reduced effort after good performance ensured they would meet the Category 2 standard and therefore avoid information disclosure. Establishments were 5.2 percentage points more likely to have positive samples after meeting the Category 2 standard for the soonest-ending window during 2015–16 (table C1, Panel B, column 2).

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

Table C2 presents results for regressions parallel to those in table C1 using placebo cutoff values for the running variables ($leeway\kappa$). Similar to table B1, 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 C2, three of the 12 RD coefficients are statistically significant with $p < 0.1$, but none of these have the correct (i.e., positive) sign. Moreover, when considering the results in table C2 together with those in table B1, only three of the 48 coefficients have the correct sign and $p < 0.1$. In conclusion, the placebo tests do not raise significant concerns about the conclusions drawn from table C1.

Distance from thresholds and Salmonella test outcomes, 2015–17

Table C3 presents results of regressions that demonstrate the positive correlations between $leeway2$ ($leeway5$) for the soonest-ending window and the likelihood of positive *Salmonella* test results. The regressions are similar to those in table 3, except that FSIS did not use sample sets during these periods. So, 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. Also, they use establishment–month–year fixed effects instead of establishment–sample-set fixed effects. In 2015–16, when the $leeway2$ value was 10 percentage points higher, the probability of a positive *Salmonella* test result was 2.00 percentage points higher ($p = 0.012$; elasticity = 0.92; panel A, column 2). Under the disclosure hiatus, there was no statistically significant relationship between $leeway2$ and *Salmonella* test results in the preferred specification, which controls for the share of samples positive in the soonest-ending window. When the $leeway5$ value was 10 percentage points higher, the probability of a positive *Salmonella* test result was 3.03 percentage points higher in 2015–16 ($p < 0.001$; elasticity = 3.36; panel B, column 2) and 4.16 percentage points higher in 2016–17 ($p = 0.018$; elasticity = 0.55; panel B, column 4).

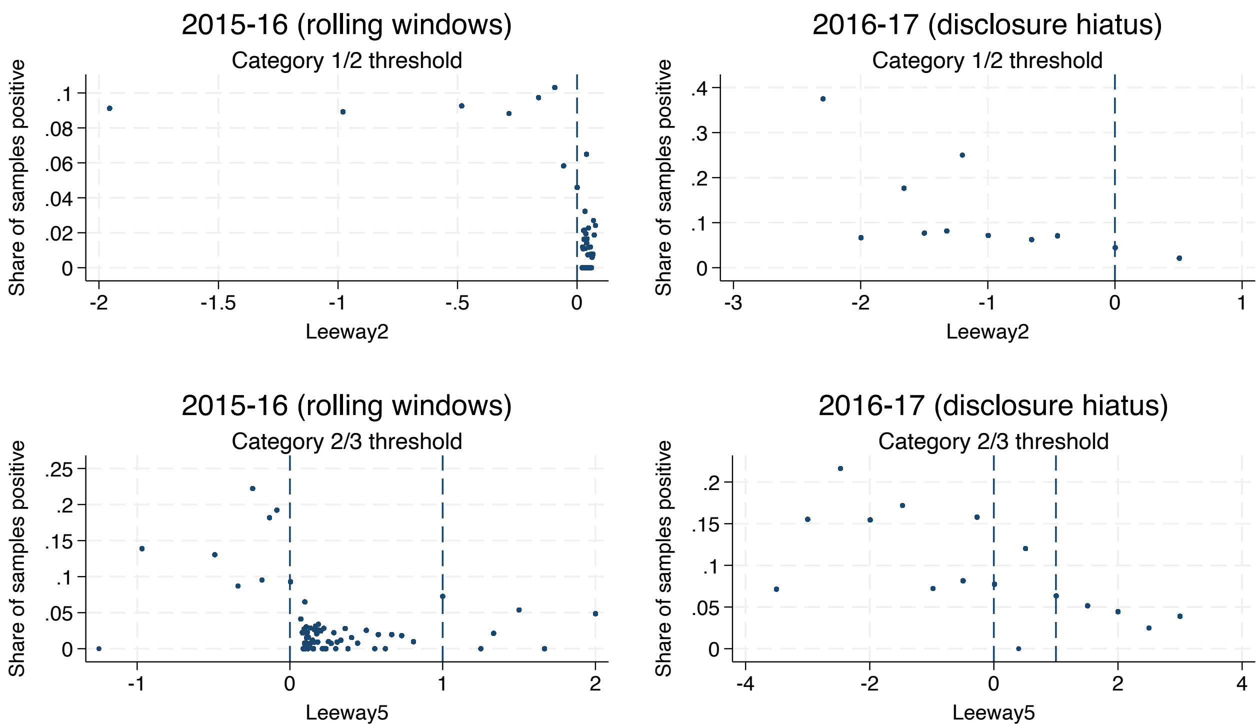
Although the correlation between $leeway2$ and positive *Salmonella* test results lessened under the disclosure hiatus from 2016 to 2017, the correlation between $leeway5$ and test results increased in this period, relative to 2011–15 and 2015–16. In other words, establishment operators appear to have 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 C4 presents results of RDiT regressions for the policy changes in 2015 and 2016. Figure C2 presents a corresponding RD plot. These policy changes had insignificant effects on average *Salmonella* test outcomes under the main specifications.[35] 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 C5, panel A, column 1). The additional robustness tests and placebo tests presented in tables C5 and C6 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 4, which shows that the introduction of public disclosure in 2008 reduced the share of samples positive by about 55 percent.

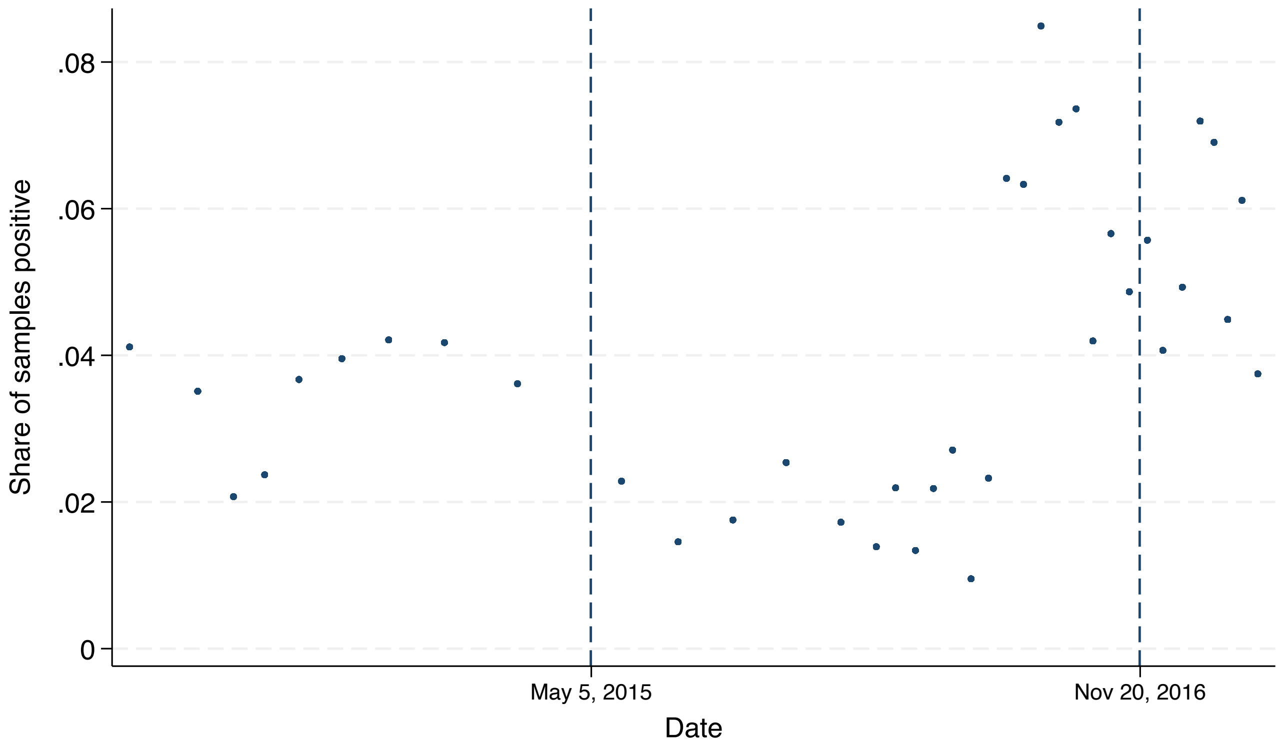
[35] In figure C2, a large rise in the share of samples positive starting July 1, 2016, can be seen. As mentioned above, this can be attributed to a change in the FSIS test procedure. The bandwidths used in estimating the effects of the 2015 and 2016 policy changes do not include July 1, 2016, when the test procedure changed.

Figure C1: RD plots: effects of known categorization on *Salmonella* outcomes



Notes: These RD plots provide graphical evidence corresponding with table C1, within the ranges of the running variables that correspond with the MSE-optimal bandwidths used in table C1. As in figures 5 and 6, quantile-spaced bins are generated using integrated MSE-optimal spacings estimators (Calonico, Cattaneo, and Titiunik, 2014a, 2015), and fit lines are not included because they tend to increase the type I error rate of visual inference (Korting et al., 2023). The $leeway2 = 1$ threshold is not shown in the upper two plots because $leeway2$ takes on only two different values at and above the cutoff so the RD models cannot be estimated around this threshold.

Figure C2: RD in time plot



Notes: This RD plot provides graphical evidence corresponding with panel B of table C4, within the temporal range that corresponds with the MSE-optimal bandwidths used in table C4. As in figure 7, quantile-spaced bins are generated using integrated MSE-optimal spacings estimators (Calonico, Cattaneo, and Titiunik, 2014a, 2015), and fit lines are not included because they tend to increase the type I error rate of visual inference (Korting et al., 2023).

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
Threshold	$D_0 = 1$	$D_0 = 1$
Implication	Cat. 2 or 3	Cat. 2 or 3
Max. # pos. samples (κ)	2	2
	(1)	(2)
Known categorization ($D_0 = 1$)	0.041	0.029
Robust p-value	0.047	0.125
95% CI (lower limit)	0.00	−0.01
(upper limit)	0.08	0.07
Observations	7467	4390
Left bandwidth	2.13	2.49
Right bandwidth	0.08	0.67

Panel B: Cutoffs associated with disclosure				
Policy regime	Rolling windows		Disclosure hiatus	
Years	2015 to 2016		2016 to 2017	
Threshold	$D_0 = 1$	$D_1 = 1$	$D_0 = 1$	$D_1 = 1$
Implication	Cat. 3	Cat. 1 or 2	Cat. 3	Cat. 1 or 2
Max. # pos. samples (κ)	5	5	5	5
	(1)	(2)	(3)	(4)
Known categorization ($D_0 = 1$ or $D_1 = 1$)	0.125	0.052	−0.004	−0.107
Robust p-value	0.217	0.008	0.860	0.587
95% CI (lower limit)	−0.03	0.01	−0.06	−0.60
(upper limit)	0.15	0.09	0.05	1.05
Observations	7435	8494	2365	4145
Left bandwidth	1.27	1.00	3.72	1.00
Right bandwidth	0.62	1.48	0.67	2.02

Notes: Each regression uses observations from the policy regimes beginning and ending in the indicated years. All regressions are local linear RD regressions with triangular kernels, using *leeway*κ as the running variable, as described in the text. $D_0 = 1$ if *leeway*$\kappa < 0$ and $D_1 = 1$ if *leeway*$\kappa \geq 1$; each of these conditions are equivalent to the known categorization outcomes reflected by the Implications rows in the table. Bandwidths, robust p-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo, Titiunik, and Vazquez-Bare, 2020), 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. The RD models cannot be estimated with $D_1 = 1$ and $\kappa = 2$ because the running variable takes on too few different values at and above the cutoff $c = 1$ (namely, the only possible values are 1 and 2).

Table C2: Placebo effects of known categorization on *Salmonella* outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Placebo RD cutoff (c)	−0.15	−0.1	−0.05	0.1	0.15	0.2
Panel A: 2015 to 2016, $\kappa = 2$ positive samples						
$D_0^p = 1$	−0.015	0.011	0.032	−0.002	−0.003	−0.010
Robust p-value	0.295	0.715	0.589	0.756	0.799	0.039
Placebo RD cutoff (c)	0.7	0.75	0.8	1.25	1.3	1.5
Panel B: 2015 to 2016, $\kappa = 5$ positive samples						
$D_1^p = 1$	0.035	0.032	0.039	−0.048	−0.036	−0.015
Robust p-value	0.102	0.268	0.130	0.001	0.020	0.878

Notes: This table presents results of regressions paralleling those in table C1 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 panel A; $c = 1$ in panel B) that are multiples of 0.05, for which there are enough observations on either side of the placebo cutoffs to estimate the optimal bandwidths around c. Each panel represents regressions using observations from the policy regimes beginning and ending in the indicated years. As in table B1, D_0^p and D_1^p are analogous to D_0 and D_1 in table 2 but use the placebo cutoffs indicated at the top of the columns. All regressions are local linear RD regressions with triangular kernels, using *leeway*κ as the running variable. Bandwidths, robust p-values, and confidence intervals are calculated using the `rdms` command in Stata (Cattaneo, Titiunik, and Vazquez-Bare, 2020), although bandwidths and confidence intervals are suppressed in this table. Bandwidths are chosen to minimize mean squared error on either side of each cutoff.

Table C3: Effects of distance from category thresholds on *Salmonella* test outcomes, 2015–17

	Rolling windows, 2015 to 2016		Disclosure hiatus, 2016 to 2017	
	(1)	(2)	(3)	(4)
Panel A				
Distance from Category 1 threshold (soonest-ending window)	0.258	0.200	0.261	−0.121
	(0.08)	(0.08)	(0.11)	(0.15)
Share of samples positive, soonest-ending window		−2.942		−14.065
		(1.02)		(6.50)
Observations	7787	7604	1863	1863
Elasticity	1.12	0.92	0.72	−0.34
Panel B				
Distance from Category 2 threshold (soonest-ending window)	0.368	0.303	0.660	0.416
	(0.07)	(0.06)	(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	3.36	0.88	0.55

Notes: This table represents the results of similar regressions to those shown in table 3, 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 distance from the Category 1 threshold ($leeway2$) on *Salmonella* test outcomes; Panel B the effects of distance from the Category 2 threshold ($leeway5$). The main variables of interest are $leeway2$ and $leeway5$ 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 $leeway\kappa$, calculated using the mean share of samples positive and the mean value of $leeway\kappa$. Observations are included only if $leeway\kappa \in [0, 1)$.

Table C4: 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). See additional notes to table 4.

Table C5: 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		
$t \geq c$	−0.012	0.005
Robust p-value	0.475	0.785
95% CI (lower limit)	−0.04	−0.03
(upper limit)	0.02	0.04
Observations	13245	5790
Left bandwidth	496	98
Right bandwidth	170	130

Notes: For additional details on the regression specifications, see notes to table 4.

Table C6: 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		
$t \geq c$	−0.015	
Robust p-value	0.813	
95% CI (lower limit)	−0.05	
(upper limit)	0.04	
Observations	12262	
Left bandwidth	502	
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 4.