

Feasible Policy Evaluation by Design: A Randomized Synthetic Stepped-Wedge Trial of Mandated Disclosure in King County

Evaluation Review

1-48

© The Author(s) 2020


Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0193841X20930852

journals.sagepub.com/home/erx



Cassandra Handan-Nader^{1,2} ,
Daniel E. Ho^{2,3,4,5}, and Becky Elias⁶

Abstract

Evidence-based policy is limited by the perception that randomized controlled trials (RCTs) are expensive and infeasible. We argue that carefully tailored research design can overcome these challenges and enable more widespread randomized evaluations of policy implementation. We demonstrate how a stepped-wedge (randomized rollout) design that adapts

¹ Department of Political Science, Stanford University, CA, USA

² Regulation, Evaluation and Governance Lab, Stanford University, CA, USA

³ Department of Political Science, Stanford Law School, Stanford University, CA, USA

⁴ Stanford Institute for Economic Policy Research, Stanford University, CA, USA

⁵ Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, CA, USA

⁶ Food Protection Program, Environmental Health Services, Public Health—Seattle & King County, Seattle, WA, USA

Corresponding Author:

Cassandra Handan-Nader, Department of Political Science, Stanford University, Encina Hall West, Suite 100, Stanford, CA 94305, USA.

Email: slnader@stanford.edu

synthetic control methods overcame substantial practical, administrative, political, and statistical constraints to evaluating King County's new food safety rating system. The core RCT component of the evaluation came at little financial cost to the government, allowed the entire county to be treated, and resulted in no functional implementation delay. The case of restaurant sanitation grading has played a critical role in the scholarship on information disclosure, and our study provides the first evidence from a randomized trial of the causal effects of grading on health outcomes. We find that the grading system had no appreciable effects on foodborne illness, hospitalization, or food handling practices but that the system may have marginally increased public engagement by encouraging higher reporting.

Keywords

design and evaluation of programs and policies, outcome evaluation (other than economic evaluation), design and evaluation of programs and policies, real-world dissemination

While evidence-based policy has made substantial strides over the generation (see, e.g., Gueron & Rolston, 2013; Haskins & Margolis, 2014), vast parts of policy-making remain uninformed by a rigorous evidence base. Haskins and Margolis (2014, p. 4) whimsically suggest that research constitutes only 1% of factors driving the adoption of social legislation. Bridgeland and Orszag (2013) estimate that only US\$1 of every US\$100 of federal spending is evidence-based. Rather than "evidence-based policy making," the default may be "policy-based evidence making" (House of Commons Science and Technology Committee, 2006).

A core challenge is that randomized controlled trials (RCTs)—the gold standard of policy evaluation (Imbens, 2010; Rubin, 2008)—are perceived of as limited in applicability when it comes to substantial questions of law and policy (Abramowicz et al., 2011; Ho, 2015; Levitt & List, 2009; Nathan, 2008). Some perceive RCTs as *infeasible* (Easterly, 2009; Ravalion, 2012; Rodrik, 2008), when policy makers seek to implement large legal and policy changes wholesale. Many perceive RCTs as slow and *expensive* (Bothwell et al., 2016).¹ And others charge RCTs as *unethical*, as, for instance, the (perceived) benefit of treatment should not be withheld (Desposato, 2015; Greiner & Matthews, 2016). These concerns can be particularly acute from the perspective of public management and administration (Margetts, 2011; Perry, 2012).

We argue that these critiques stem in part from an unduly restrictive notion of RCTs. Practical research design can overcome many of these challenges to public administration while retaining the benefits of a randomized design. Implementation typically cannot be carried out instantaneously across a large jurisdiction. Vaccinations cannot be administered to all children simultaneously (Hemming et al., 2015). Cash transfers cannot be granted in one fell swoop (Fernald et al., 2008). Universal health care cannot be rolled out instantaneously across a country (King et al., 2007). As a result, practical constraints often mandate that implementation be *sequenced*, and sequencing facilitates rigorous evaluation of the policy during implementation.² In keeping with recent practical guidelines for RCTs in real-life settings (Todak et al., 2018), we argue that rigorous, high-quality experimental designs can be flexible enough to accommodate logistical constraints and other practical challenges.

We illustrate with an RCT we designed to evaluate the food safety rating system adopted by the public health department in Seattle & King County (Public Health—Seattle & King County). In that context, political constraints required that the grading system for restaurants be implemented across the entire county in 2017. Treatment could not be withheld. Randomization at the restaurant level was infeasible, primarily because cases of foodborne illness cannot be attributed to specific establishments at scale. Operational concerns about training inspectors and onboarding restaurants, as well as equity concerns about the impact of grading, loomed large (Ho, 2017b). The initial perception was hence that an RCT would be too expensive, operationally unmanageable, and potentially unfair. These perceptions explain why not a single randomized evaluation of grading exists to date despite the fact that many jurisdictions have adopted grading policies.

The research design we developed in collaboration with Public Health aimed to solve these problems. To address the fact that the entire county needed to be treated, we adapt a stepped-wedge (randomized rollout) design that randomized the order in which grading was rolled out in four subregions of King County. While prominent in medicine, the design remains underutilized in evaluations of law and policy (Mdege et al., 2011, pp. 938–939). The design also allowed the department to improve operational logistics based on the early implementing regions. Second, we develop an algorithm to construct synthetic regions—which are contiguous and respect municipal boundaries—that are statistically indistinguishable in preexisting foodborne illness trends. We show that adapting insights from synthetic control methods (Abadie et al., 2010) can improve statistical power, addressing a key limitation of cluster-based randomization. Third,

to control the false positive rate, which is particularly problematic in this setting due to regional outbreaks (Ho et al., 2019), we use randomization inference that directly accounts for the randomization scheme (Ho & Imai, 2006; Imbens & Rubin, 2015; Rosenbaum, 2002a, 2002b). We show that in our setting, this approach outperforms the cluster-robust standard errors conventionally used for controlling test size. Last, we leverage administrative data to minimize the costs of outcome data collection (Buck & McGee, 2015; Coalition for Evidence-Based Policy, 2015, p. 1).

Our substantive results address a major question in the scholarship on mandated information disclosure for which restaurant grading has proven a pivotal case (Ben-Shahar & Schneider, 2014; Bubb, 2015; Fung et al., 2007; Ho, 2012; Loewenstein et al., 2014). We find that the grading system had no statistically or substantively significant effects on foodborne illnesses or hospitalizations. We corroborate these results by examining the effects of the grading system on risk factors as measured by an independent inspection team created specifically for this evaluation. The team was trained by the Food and Drug Administration (FDA) and for 18 months observed restaurants (without grading them or issuing any inspection report) including periods before and after grading implementation in three subregions. While this inspection component increased the budget of the evaluation, the evaluation could have been conducted without that line item. We find that the grading system may have reduced the propensity by inspectors to *cite* critical violations for unchanged sanitation practices but find suggestive evidence that the grading system sparked more public submissions of food poisoning complaints. The grading system might hence facilitate greater public engagement with food safety and investigation of outbreaks, although the durability of this effect is much less certain.

This article proceeds with the following sections. First, we provide background on food safety and our collaboration with King County. Second, we spell out some of the key practical constraints on the evaluation. Third, we show how research design overcame these practical barriers. We then present results and conclude with a brief discussion.

Background

Foodborne illness accounts for some 46 million illnesses, 128,000 hospitalizations, and 3,000 deaths annually in the United States (Centers for Disease Control and Prevention [CDC], 2011). In the past 20 years, an increasingly popular reform has been the *disclosure* of restaurant inspection results to consumers (Ho, 2012; Kovacs et al., 2018; Seiver & Hatfield,

2000). By simplifying the results of an inspection visit and posting placards in entryways of restaurants, “restaurant grading” promises to inform consumers about food risk and provide market incentives for restaurants to adopt safer sanitation practices (Fung et al., 2007; Weil et al., 2006). Beginning in 2014, the public health department of Seattle & King County (Public Health—Seattle & King County) explored the possibility of a public grading system with the express goals of reducing foodborne illness and improving risk communication with the public. To provide context for the evaluation and because the process consumed considerable departmental resources, we detail background on the grading system and planning process here.

A comprehensive program review in 2014 recommended that the Public Health “implement a restaurant reporting system that is transparent, credible and intuitive” by “[c]onducting further research into window placard system” (Public Health—Seattle & King County, 2014, p. 16). At the same time, the program review identified the challenge of maintaining the accuracy and consistency of frontline inspections. The report recommended “an ongoing quality control program to increase operational consistency and quality in inspections” (Public Health—Seattle & King County, 2014, p. 2).

In response to that program review, Public Health began a process of extensive stakeholder engagement. It constituted a “Food Program Stakeholder Advisory Committee” comprised of health officials, restaurant representatives, academics, and members of the public. A subcommittee reviewed eight existing restaurant reporting systems. The committee articulated the view that the rating system, in contrast to existing ones, should consider using multiple inspections, high-risk violations, relative performance of establishments, and account for effects on the diverse community in the county.³ The committee also considered the question of the quality and consistency of restaurant inspections, as inconsistencies among frontline inspectors would undermine the reliability of disclosure (Boehnke & Graham, 2000; Seiver & Hatfield, 2000). Senior food program managers expressed the concern that grading might reduce the propensity by inspectors to cite code violations, as evidence of “grade inflation” exists in many other jurisdictions (Ho, 2012; Kovacs et al., 2018).

The department also convened a series of all-staff meetings around restaurant grading and the consistency of the inspection process. One meeting focused on the value of consistency/reliability of the inspection process expressly connecting the issue to the credibility of the grading system. To address the problem of the accuracy and consistency of inspections, Public

Table 1. Description of the County’s Restaurant Grading Categories.

Rating	Description
Excellent	No or few red critical violations over the last four inspections
Good	Some red critical violations over the last four inspections
Okay	Many red critical violations over the last four inspections
Needs to improve	The restaurant was either closed by Public Health—Seattle & King County within the last year or the restaurant needed multiple return inspections to fix food safety practices

Health developed a “peer review” initiative, whereby inspectors were randomly paired up to conduct joint inspection visits, independently scored health code violations, and developed training materials based on these results. The peer review initiative was evaluated with an RCT in 2016, showing gains in accuracy and consistency (Elias & Ho, 2016; Ho, 2017a).

The design of the grading system involved two parts. The first part focused on usability. Roughly 35% of King County’s population is minority (King County Executive Office, 2015), 20% is foreign-born, and over 10% has limited English proficiency (Felt, 2017). The county used a community-based participatory approach to design the placard, convening a series of (multilingual) focus groups with community partners. A graphic designer then developed placards that would communicate the notion of food risk effectively to a diverse population. Appendix A presents the placard that emerged out of this usability process. The second part focused the reliability of grading. One of the basic criticisms of extant grading systems is that a single visit presents merely a “snapshot in time” (Boehnke & Graham, 2000; Seiver & Hatfield, 2000; Wiant, 1999). In contrast to other jurisdictions, the county decided to base grades on (a) multiple routine (unannounced) inspections, (b) “critical” (red) code citations (as peer review showed that noncritical code items led to greater inconsistency), and (c) an adjustment for differences in stringency across inspection areas (Ashwood et al., 2016). Table 1 provides the grading categories and their respective descriptions.

In January 2017, the King County Board of Health adopted the food safety rating system. As part of the County Executive’s campaign for creating the “best-run government” that prioritizes evidence to improve government and employee performance, the county was open to an evaluation. Such an evaluation would be of tremendous public value, given the popularity of grading and resource costs of implementing a grading system. New

York City, for instance, allocated some US\$3.2 million for the implementation of its grading system, constituting a 19% increase in its program budget (Collins, 2010). The only systematic evidence to date about the effect of grading on health outcomes stems from an observational (nonrandomized) study in Los Angeles (LA) County, which found a 20% reduction in foodborne illness hospitalizations manifesting itself in the first quarter after adoption.⁴ A rigorous understanding of the effect of a grading system on health outcomes has implications beyond public health, as restaurant grading has played in a pivotal role in the broader scholarship on mandated information disclosure (Ben-Shahar & Schneider, 2011, 2014; Bubb, 2015; Fung et al., 2007; Ho, 2012; Loewenstein et al., 2014; Weil et al., 2006; Winston, 2008).

Notwithstanding the county's openness to an evaluation, there were substantial reservations about the feasibility, expense, and ethics of a randomized evaluation, which we detail below. The county had originally planned a series of surveys and community-based focus groups, but those components would not assess the effect on health outcomes. Beginning in 2014, we worked closely with Public Health to design an evaluation that could overcome the considerable practical challenges. Our goal was to facilitate an agreement on a rigorous design by defraying any perceived costs to an evaluation. None of the academic work was compensated, providing a unique model for academic–agency collaboration to promote evidence-based policy.

Practical Barriers to Unit-Level Randomization

A priori, one appealing way to identify the effect of restaurant grading on foodborne illness might be at the restaurant level. For restaurant $i \in \{1, \dots, N\}$ randomly assigned to one of the two possible conditions, $T_i = 1$ for restaurant grading and $T_i = 0$ otherwise, we would estimate the average treatment effect as $\hat{\tau} = \frac{1}{N_1} \sum_{i=1}^N Y_i \cdot T_i - \frac{1}{N_0} \sum_{i=1}^N Y_i \cdot (1 - T_i)$, where Y_i represents the foodborne illness cases arising from restaurant i and N_1 and N_0 are the number of restaurants assigned to the treatment and control conditions, respectively (Neyman, 1923; Rubin, 1974). But numerous practical constraints prohibit this ideal experiment.

Illness Attribution

The most important limitation is that foodborne illnesses cannot be attributed to specific establishments at scale. Public Health receives complaints

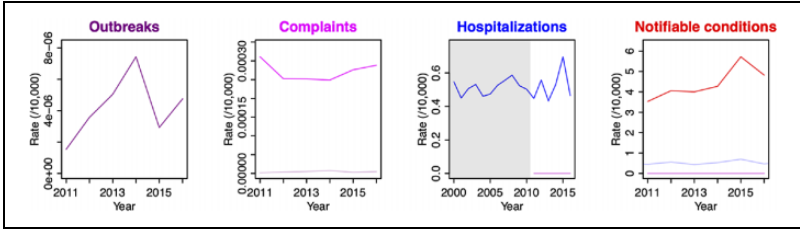


Figure 1. Four sources of data on foodborne illness for King County: Probable or lab-confirmed foodborne illness outbreaks (first panel, purple), public complaints (second panel, pink), hospitalizations (third panel, blue), and illnesses via state-mandated report (right panel, red). Panels are organized by increasing prevalence, and each sequential panel includes the previous panel's data (in faint lines) for comparison. Illnesses provide roughly 8.5 times the volume of hospitalizations and 25 times the volume of complaints. The gray box indicates years available for hospitalization data but unavailable for other outcomes.

about establishments (second panel of Figure 1), but given the incubation period of most common foodborne pathogens, attribution requires an extensive and costly epidemiological investigation. This investigation includes completing a 2-week food history for potentially affected individuals, a case-control analysis to identify likely sources, laboratory testing where available, and an on-site inspection of suspected establishments. From 2011 to 2016, only 21 of 3,280 consumer complaints to Public Health were attributable to King County restaurants by lab confirmation, and only 69 complaints resulted in an epidemiological classification of “probable” attribution. Only 53 outbreaks (defined as two or more individuals affected) were probable or lab-confirmed from complaints (left panel of Figure 1). This constitutes a very small fraction of foodborne illnesses in the county, making restaurant-level randomization infeasible.⁵

Outbreaks

An alternative outcome, which is not attributable to specific establishments, lies in hospitalization discharge data. Indeed, because the LA effect was based on hospitalizations, our evaluation started with this outcome in mind. Given some uncertainty about diagnostic codes (Ho et al., 2019), we consulted with epidemiologists specializing in foodborne disease in King County and the CDC to compile a comprehensive set of discharge codes for foodborne illnesses likely to be affected by restaurant sanitation practices (see Appendix B). Yet, we found that a small number of large

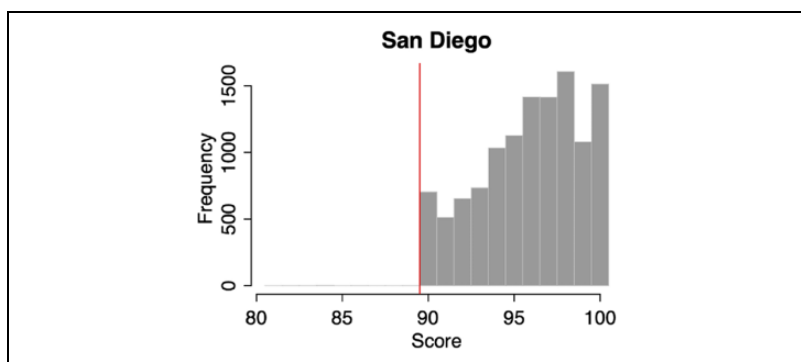


Figure 2. Score distribution in San Diego, where 90 points is the cutoff for an “A” grade, in 2011.

outbreaks meant hospitalization trends were highly stochastic as can be seen in the third panel of Figure 1. Between 2013 and 2015, with outbreaks stemming from Chipotle and a large meat producer, the rate increased by 60%. Even assuming a large 20% grading effect from LA, a design based on hospitalization data alone would be significantly underpowered.

The best available outcome data consist of “notifiable conditions” that laboratories, physicians, and hospitals are required to report to state authorities after a diagnosis is made. Seven illnesses that are primarily foodborne (and not travel-related) are available: campylobacteriosis, cyclosporiasis, listeriosis, salmonellosis, shiga toxin-producing *E. coli*, vibriosis, and yersiniosis. The right panel of Figure 1 shows that these data provide roughly 8.5 times the volume of foodborne hospitalizations, providing a more stable time series. Like hospitalizations, these illnesses cannot be attributed to specific establishments but are aggregated at the zip code level based on the residential address of the patient.

Nonindependence

Even if outcomes were attributable to specific restaurants, the premise of restaurant grading is that it creates competitive pressure for sanitary practices. The rating that one restaurant receives can hence affect potential outcomes for other restaurants, violating conventional independence assumptions (Rubin, 1980). Grading must hence be implemented at a large enough scale to minimize interference (e.g., diner crossover between treated and control establishments) but small enough to allow for actual comparisons.

Political Pressure for Adoption

Due to political pressure, the grading system needed to be implemented no later than 2017. No parts of the county could be left untreated. The pressure intensified due to high-profile, large outbreaks in 2015 (e.g., Chipotle), posing an inferential threat to any simple comparison over time: foodborne illness may decrease in subsequent years by “regression-to-the-mean” alone. Reforms may *appear* effective solely because large outbreaks have died down. This form of endogenous policy adoption threatens causal inference for many legal and policy settings (see, e.g., Grambsch, 2008).

Equity Implications and Staff Division

A 2014 King County Executive Order mandated that departments adopt initiatives with equity and social justice in mind. Public Health was, as a result, particularly conscientious of potential inequities of the grading system along income, ethnicity, and cuisine. This posed another challenge to randomization, as any single randomization (particularly at the cluster level) might be seen to unfairly burden some set of establishments. Grading also proved controversial among inspection staff, and Public Health wanted to equitably share the implementation burden across two field offices.

Budgetary Constraints and Inspector Behavior

The county began with essentially no budget for evaluation. This is common for adoptions of grading systems with none to date building in a rigorous evaluation. As a result, observational studies focus on whether the citation of violations decreases after the enactment of grading. In addition to regression-to-the-mean effects, such designs are limited as they conflate two distinct effects: (a) the effect of grading on whether *inspectors* cite violations for otherwise identical practices and (b) the effect of grading on *restaurant* sanitation practices. To see why reliance on inspection scores alone can be problematic, Figure 2 displays the score distribution of establishments in San Diego. Higher scores represent better inspection performance, and the vertical line at 90 represents the cutoff to receive an “A” grade. Given considerable discretion and ambiguity in the health code, it is implausible that establishments are cleaning up precisely to the threshold. The much more likely explanation is that inspectors use the discretion to avoid a confrontation, a dynamic widely noted in the public health literature (Boehnke & Graham, 2000) and of concern to senior management in King

County. Kovacs et al. (2018), for instance, show that “fudging” inspection scores to meet the grading threshold increases when there are social ties between the inspector and establishment. As one San Diego inspector noted, “Some inspectors will give out a B for an 89. I usually warn somebody at that point. It’s a judgment call” (Sylvester, 1980).

Addressing Practical Barriers via Research Design

We now articulate how advanced research design can be deployed to address the significant practical barriers to evaluation.

Stepped-Wedge (Randomized Rollout) Design

Because of close communication around the grading initiative, our team was aware of considerable operational challenges in launching the grading system.⁶ These operational challenges helped persuade the county to roll out the grading system in stages. The chief benefit from the county’s perspective would be that a smaller group of inspectors could be onboarded and then learn from the pilot jurisdiction to train the rest of the staff. One point of negotiation was the duration of the rollout and the number of stages. There was no ability to extend the rollout beyond 2017, and operationally, the county felt that sequencing over four quarters would be ideal.

After this commitment, we adopted a cross-sectional stepped-wedge randomized trial (The Gambia Hepatitis Study Group, 1987; Hussey & Hughes, 2007). Conventionally, stepped-wedge designs randomize the order in which an intervention is implemented across preexisting groups (e.g., municipalities). All groups end up being treated, but the treatment effect is identified based on (randomized) group–time variation in the intervention. In a classic setting, stepped-wedge trials model individual-level data with a random effect for the group and a fixed effect for the time period and use a random effects error correlation structure to account for correlation across individuals within the same group (Hooper et al., 2016; Hussey & Hughes, 2007; Kasza et al., 2017). We make two significant alterations to the conventional analysis. First, due to the stochastic nature of foodborne illness trends discussed in the Practical Barriers to Unit-Level Randomization section, we aggregate our data to the cluster-period level. In place of a random effects model, we estimate a least squares two-way fixed effects model as shown in Equation 1:

$$Y_{ij} = \alpha_i + \beta_j + \theta T_{ij} + \varepsilon_{ij}, \quad (1)$$

where Y_{ij} is the foodborne illness rate for group i at time j , α_i is a fixed effect for group i , β_j is a fixed effect for time j , θ is the treatment effect, and T_{ij} equals 1 for groups in time periods when they are assigned the treatment and 0 otherwise. Once a group has received the treatment, it will continue to receive the treatment for all subsequent time periods. The two-way fixed effects model assumes a common secular time trend across units in the absence of the treatment and additive and constant treatment effects. These assumptions seemed appropriate given the nearly immediate 20% effect reported for LA that we used to inform our power analysis. Second, due to the small number of clusters involved, we use a randomization inference procedure discussed in Randomization Inference subsection rather than using the conventional cluster-robust standard errors which typically accompany two-way fixed effects models (Cameron & Miller, 2014). This procedure does not impose any parametric assumptions on the null distribution of the test statistic.

With this strategy, if grading is implemented over the course of four quarters in four regions, we should be able to observe reductions in foodborne illnesses corresponding to when grading is rolled out in each region. The design helps to account for a quarter-specific shocks (e.g., an outbreak) across King County, as well as region-specific (but time-invariant) differences, and randomization allows for a “reasoned basis for inference” as we spell out below (Fisher, 1925, 1935; Ho & Imai, 2006; Imbens & Rubin, 2015).

Improving Power With Synthetic Regions

While the application of a stepped-wedge design is more novel for law and public policy, it can also be underpowered. Power limitations are exacerbated by the fact that researchers commonly roll out the intervention across existing administrative units (e.g., municipalities). There is, hence, no guarantee that these units are comparable in baseline outcomes. To illustrate this, the left panel of Figure 3 plots five existing administrative regions in King County (top) and the associated time series for foodborne illnesses (bottom).⁷ While the time series of the different regions exhibit similar seasonality, there is substantial evidence of quarter- and region-specific shocks. This poor balance on pretreatment trends results in noisy estimates of treatment effects. To illustrate this, we simulate treatment effects of 20% magnitude and 0% magnitude on 2014 or 2015 and fit Equation 1 above across these treatments. The left column of Table 2 displays statistical power (i.e., the proportion of tests that reject the null hypothesis given a

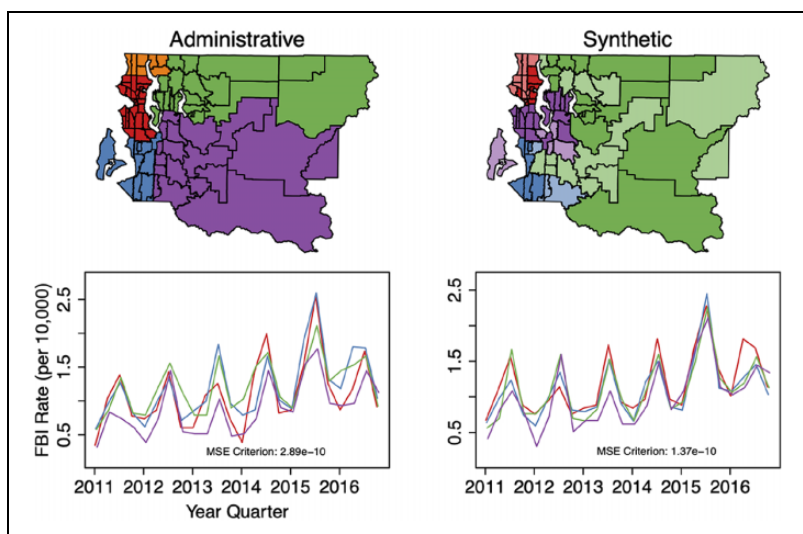


Figure 3. Pretreatment foodborne illness trends when using existing administrative regions to divide up the county (left) and algorithmically derived synthetic regions (right). The synthetic regions show improved pretreatment balance between regions in part by excluding zip codes from the evaluation regions if quarterly balanced is improved (excluded zip codes are shaded more lightly in the right panel). Quantification of quarterly pretreatment balance (see Equation 4) is shown at the bottom of each plot.

20% effect) and size (i.e., the false rejection rate or Type I error given no treatment effect). A well-designed trial should have high power (conventionally $>80\%$) and size at the α level (5%). Using administrative units, however, results in low power of 70% coupled with a high false rejection rate of 25%.

We sought to improve the power of the design by leveraging insights from synthetic control methods (Abadie et al., 2010, 2015). The basic insight is that by creating synthetic rollout regions, we may be able to construct more plausible comparison groups that exhibit parallel foodborne illness trends. Because inspectors are principally assigned based on zip code, we begin with zip codes as the primitive units. Our goal is to aggregate 85 zip codes in King County to four regions that are comparable on foodborne illness trends so as to maximize power while controlling size. An additional degree of freedom comes from the fact that while all zip codes would be included in the *implementation* of grading, individual zip codes

Table 2. Statistical power at a 20% effect size and size (false positive rate or Type I error) of the stepped-wedge quarterly rollout test using administrative regions (Admin.) or synthetic regions (Synth.), using $\alpha = .05$ in a one-tailed test for reduction in foodborne illness.

Unit	Parametric		Bootstrap		Randomization Inference	
	Admin.	Synth.	Admin.	Synth.	Admin.	Synth.
Power	.70	.95	.30	.55	.33	.89
Size	.25	.25	.16	.15	.04	.04

Note. Three inferential methods are presented: a standard parametric method using cluster-robust standard errors (Wooldridge, 2001), the wild bootstrap approach designed to account for a small number of clusters (Cameron et al., 2008; Esarey & Menger, 2017), and the randomization inference approach described in Randomization Inference subsection (Imbens & Rubin, 2015). There are five administrative regions, which can be rolled out in 120 different ways, and 2 years, so power and size are calculated over 240 tests. Similarly, because there are five possible synthetic region allocations, each of which can be rolled out in 24 different ways, and 2 years, power and size are calculated over 240 tests.

could be excluded from the *evaluation* of grading. This trades off higher internal validity for lower external validity and brings our total target regions to five (one for each quarter, plus a leave-out group).

To improve performance, we imposed several other constraints on region identification. First, we require regions to be geographically contiguous. This minimizes the chance that diners cross over treated and nontreated areas, as it is well established that most restaurant competition is local (Parsa et al., 2011). However, as we discuss in the Discussion section, this design does not completely preclude spillover effects across regions sharing border zip codes. We supplement this potential weakness in the primary design with a supporting cross-county evaluation presented in the Foodborne Disease Incidence subsection, where spillovers are much less of a concern due to the extensive geographic scale of counties. Second, for similar reasons, for 31 municipalities in King County, we required zip codes in the same municipality (other than Seattle) to remain in the same cluster.⁸ Third, because Seattle is a large municipality that spans 37 zip codes, with lower likelihood of contamination across the city, we subdivided Seattle into municipal districts based on boundaries of Seattle’s Office of Planning and Community Development (Assefa, 2010).⁹ Fourth, because foodborne illness has a relatively low baseline incidence of roughly 42 cases per 100,000 people

(annually from 2011 to 2016), we imposed a population threshold of 50,000 for each region.¹⁰

The sample space of all possible ways to group 85 zip codes in King County into five mutually exclusive groups is large:

$$\binom{85}{5} = \frac{1}{5!} \left[5^{85} - \left[\binom{5}{1} 4^{85} - \binom{5}{2} 3^{85} + \binom{5}{3} 2^{85} - \binom{5}{4} 1^{85} \right] \right] \sim 2.15 \times 10^{57}.$$

Searching this sample space is computationally intensive, and simple random sampling is inefficient. We hence developed an algorithm to efficiently identify regions subject to our constraints.

The algorithm begins by randomly seeding four zip codes, and randomly adding contiguous zip codes until population threshold (50,000 people) is met in each region. Then, we randomly propose adding neighboring zip codes based on whether the addition improves the mean absolute value (MAV) of pairwise differences in yearly foodborne illness rates in the pretreatment period (2011–2016):

$$\text{MAV}_a = \frac{1}{IJ(I-1)} \sum_{i=0}^I \sum_{l \neq i}^I \sum_{j=1}^J |Y_{ij}^a - Y_{lj}^a|, \quad (2)$$

where Y_{ij}^a is the aggregated rate of foodborne illness in year j ($j = 1, \dots, J$) in region i ($i = 1 \dots I$) of allocation a .¹¹ Figure 4 shows an example of one run of the algorithm, which iterated through 49 possible region allocations. We ran this algorithm 2.3 million times, producing an average of 38 possible region allocations per run. We discard duplicate allocations. Because allocations within a run can be quite similar, we also thin potential region allocations by storing every fifth step. This results in 17.5 million possible allocations that we subset to the most balanced ones as we explain below in the Selecting Potential Region Allocations subsection.

The second column of Table 2 shows how the construction of such synthetic regions can improve statistical power. Instead of 70% power with administrative units, statistical power with synthetic regions increases to 95%. That said, size remains poor at 25% with conventional parametric inference using cluster-robust standard errors.

Randomization Inference

The reason for poor size with conventional methods lies in the small number of clusters (Cameron et al., 2008; Cameron & Miller, 2014). The middle panel of Table 2 applies the wild cluster bootstrap of Cameron et al. (2008), which seeks

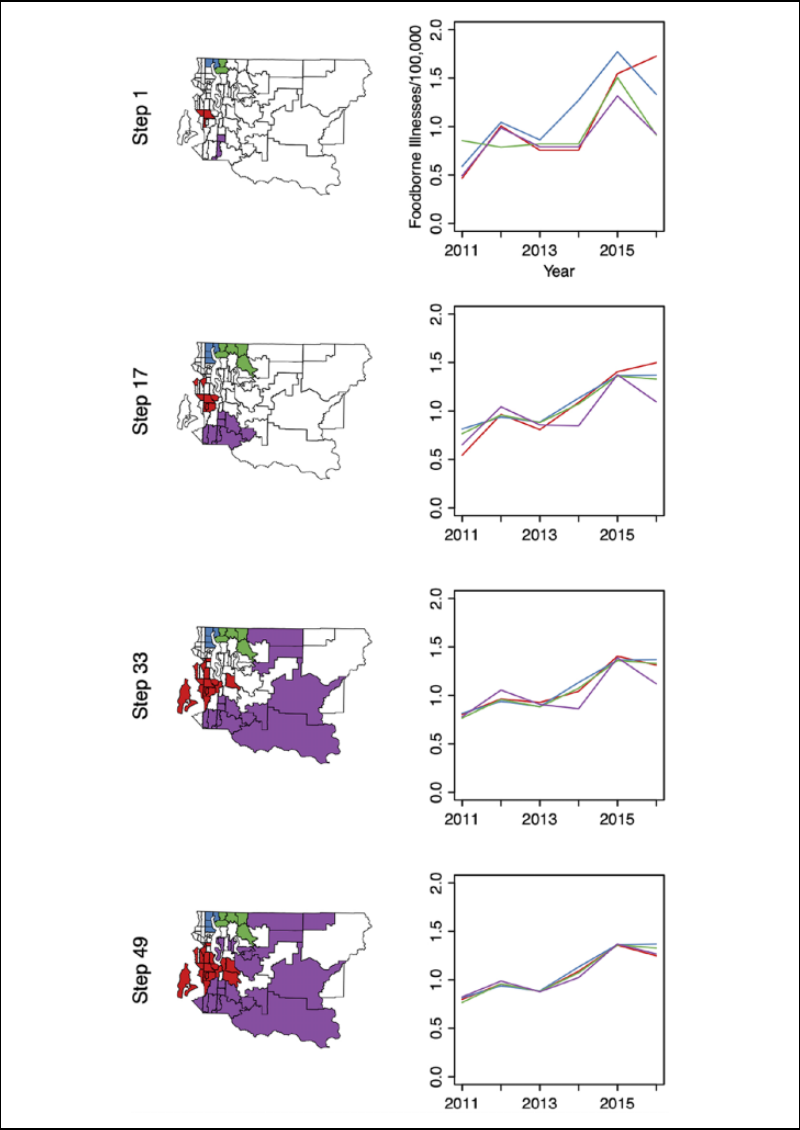


Figure 4. Region growth algorithm progression. After starting regions are formed with random seeds in Step 1, zip codes in contiguous municipalities are added to the regions if they minimize the criterion in Equation 2. The algorithm stops when no available zip codes can further minimize the criterion.

to correct for this deficiency. Even this technique, however, results in false positive rates 3 times larger than α of .05 and reduces power substantially. We hence study the use of randomization inference, which can account precisely for the treatment assignment scheme of randomizing from (a) potential regional allocations and (b) potential rollout orders with four regions within each allocation ($24 = 4!$ in our scheme). Under the sharp null hypothesis, outcomes are fixed: $(Y_{ij}|T_{ij} = 1) = (Y_{ij}|T_{ij} = 0)$ for the observed treatment vector \mathbf{T} and foodborne illness rate Y_{ij} in region i and quarter j . The only stochastic component comes from random treatment assignment.

Let Ω be the set of all possible treatment vectors. The size of Ω is $K = M \times 4!$, where M is the number of possible region allocations and $4!$ is the possible number rollout orders given a region allocation. Let $W(\mathbf{t})$ represent the test statistic, namely θ from the stepped-wedge regression in Equation 1, given a treatment vector \mathbf{t} . Under the sharp null, this test statistic can be calculated across all possible K randomizations, and the set $W(\mathbf{t})$ forms the randomization distribution. The one-tailed p value can hence be calculated as the following:

$$\Pr(W(\mathbf{T}) \leq W(\mathbf{t})) = \frac{\sum_{\mathbf{t} \in \Omega} 1(W(\mathbf{T}) \leq W(\mathbf{t}))}{K}. \quad (3)$$

The stepped-wedge test statistic allows us to adjust for region- and quarter-specific effects separately, while randomization inference leverages the rollout design over regions and time to test the null hypothesis that foodborne illness rates were invariant to the adoption of grading. The intuition is that we examine how extreme the observed test statistic is relative to all the other ways that grading could have been implemented under our randomization scheme. In that sense, randomization provides the “reasoned basis for inference” and imposes no parametric assumptions on the distribution of the test statistic under the sharp null hypothesis (Fisher, 1925, 1935; Ho & Imai, 2006; Imbens & Rubin, 2015; Rosenbaum, 2002a, 2002b).

The right panel of Table 2 shows that this method of inference has desirable properties in our setting. Randomization inference controls the false positive rate at roughly $\alpha = .05$, which stands in sharp contrast to parametric and bootstrap methods.

Selecting Potential Region Allocations

Power. Having established that randomization inference with synthetic regions has desirable properties, we now spell out the selection of M

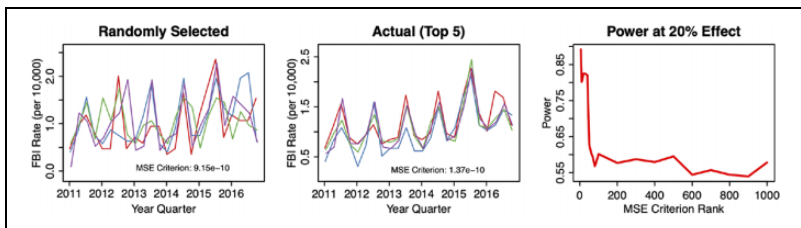


Figure 5. Power at a 20% effect size when ranking region allocations by the mean squared error criterion described in Equation 4 (left). Pretreatment balance on foodborne illness trends is shown for a randomly selected region allocation and the chosen region allocation, which was ranked in the top 5 (right). We achieve the highest power (89%) when selecting the top 5 allocations for the randomization distribution.

potential rollout regions from the 17.5 million possible region allocations. As noted before, M region allocations \times 24 orders (per allocation) form the “randomization distribution” (i.e., the reference distribution). The left panel of Figure 5 plots a random allocation of the 17.5 million, showing that there is still substantial imbalance in parts of the full distribution.

To select our region allocations, we hence ranked allocations by yearly MAV (originally used to optimize the region creation algorithm) and mean squared error (MSE, quarterly prediction):¹²

$$\text{MSE}_a = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J \left(Y_{ij}^a - \hat{Y}_{ij}^a \right)^2, \quad (4)$$

$$\hat{Y}_{ij}^a = \alpha_i^a + \tau_j, \quad (5)$$

for region i in year-quarter j in allocation a . We then used simulation-based power calculations to examine power and size as we include region allocations going down the ranked list (Baio et al., 2015).¹³ The right panel of Figure 5 plots power against the number of potential allocations (ranked by MSE, which we found to outperform MAV). We observe that power is substantially higher for low M . With $M = 5$, we observe the highest statistical power at 89%, while the false positive rate is controlled at .04.

Equity constraints. With a well-powered design, we explored addressing equity constraints articulated by the county: that regions should not contain lopsided numbers of inspectors to train; that the operational burden of implementation should be evenly shared across the two offices; that no

single region should have an overwhelming majority of gradable restaurants; and that grading resources should be “fairly” allocated to areas of particular income levels or urbanization. We hence trimmed region allocations by rejecting allocations with severe demographic imbalances across regions as displayed in Table D1 of Appendix D. For external validity, we also excluded any allocation that included fewer than 50% of King County zip codes in the evaluation region. To equitably distribute the operational burden of implementing the grading system, we allocated zip codes not in the evaluation regions in a way that balanced the number of restaurants subject to grading in each implementation region (see Appendix C for more details).

After trimming, randomization inference remained adequately powered at 85% with Type I error controlled at 4%. Figure 6 shows the final randomization distribution with zip codes sized relative to their populations for visibility. Not all rollout orders are shown, but there are a total of 120 ways in which the grading system could have been rolled out. The top right box outlines the allocation and rollout order drawn on December 13, 2016.

Baseline Risk Factor Study

As the evaluation plan unfolded, enthusiasm for the study grew. In particular, to study the effects on *citation* behavior, we explored partnering with the FDA around their risk factor analysis. Under the Government Performance and Results Act, FDA trains an independent corps of federal inspectors to observe risk factors across the country, independent of the local health code and without citations to operators. Because of the concern that inspectors would be loathe to cite violations with the additional pressure of grading, such a risk factor analysis could be invaluable to measuring outcomes independent of how grading affects citation behavior. Unfortunately, FDA’s study plan permitted only a very limited number of such inspections in King County. In late 2016, the county identified additional funding for contractor inspectors, who would be hired from outside the county, trained in part by FDA, and carry out risk factor inspections independent of the King County inspection corps. Based on a power analysis, we created a custom risk factor checklist that would enable these inspectors to carry out observations quickly (on average in about 20 min, see Appendix E), in lieu of the typical 45-min–1-hr visit, and randomized establishments for this team to visit. One downside to this risk factor data collection is that hiring constraints prevented commencing it before 2017. At the same time,



Figure 6. Final randomization distribution ($K = 120$). Zip codes are displayed as rectangles proportional to their population. The top of the figure shows how this rectangle transformation looks for all King County zip codes. The top 5 ranked region allocations are shown along the columns, and the 24 possible order permutations are shown along the rows (darker colored zip codes get grading earlier than lighter colored zip codes). Zip codes not included in the evaluation are shown in gray. The allocation randomly selected for the actual rollout is outlined in red in the top right.

political pressure meant grading had to begin in January 2017. This meant that only posttreatment data were available for region 1, but we stratified the sampling scheme to conduct balanced inspections before and after grading implementation for the remaining three regions.

It is important to note that all of these extensive design efforts transpired before grading began. The analysis plan follows readily from the design, as there is no way to change the regions (and underlying randomization distribution) after implementation. We view this precommitment as a virtue of prioritizing the experimental design (Rubin, 2008).

Results

Although our design was based around the analysis of foodborne illnesses (notifiable conditions), for completeness, we present analyses of the effect of the grading system on all available outcomes. The Foodborne Disease Incidence subsection examines the effect on measures of foodborne disease incidence. The Risk Factors subsection assesses the effect on more direct measures of food handling practices from the risk factor study. The Restaurant Complaint Data subsection discusses effects on foodborne illness complaints and their resolution.

Foodborne Disease Incidence

Hospitalizations. We begin by fitting Equation 1 to quarterly hospitalization rates per region using the patient zip code from Washington's Comprehensive Hospital Abstract Reporting System available up through 2017. The left panel of Figure 7 provides the time series of each of the regions with darker lines indicating regions randomized to implement grading earlier. The dashed line represents the start of the quarterly restaurant grading rollout. If grading had as immediate and large an effect as suggested by earlier work, the darker lines should exhibit earlier decreases in disease incidence. The time series, however, confirm that hospitalizations are quite stochastic. As shown in Table 3, we find a nonstatistically significant *increase* in foodborne hospitalizations of .01 per 10,000 (one-tailed $p = .57$) with a 95% confidence interval (CI) of $[-0.09, 0.11]$ hospitalizations per 10,000.¹⁴

Illnesses. The middle panel of Figure 7 displays the time series in the four regions of reported foodborne illness rates (per 10,000) from 2011 to the first quarter of 2018. Trends are much more stable than for hospitalizations.

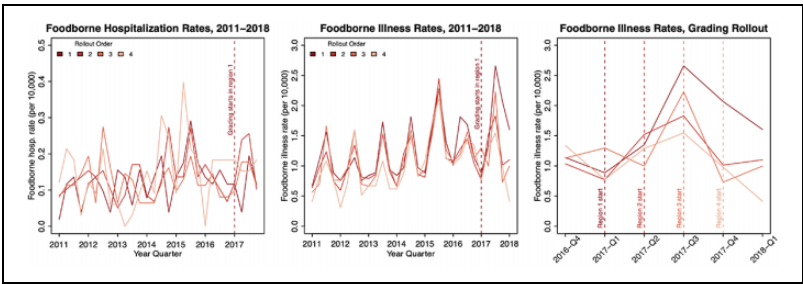


Figure 7. Foodborne hospitalization rates from 2011 to the first quarter of Q1 in King County (left), foodborne illness rates from 2011 to the first quarter of Q1 in King County (middle), and foodborne illness rates specifically during the grading rollout period from Q1 2017 to Q1 2018 (right). We find no significant change in foodborne illness due to restaurant grading.

Table 3. Results Leveraging the Synthetic Region Design Described in Improving Power With Synthetic Regions Subsection.

Outcome	Effect	<i>p</i> value	95% CI
Foodborne illness (rate)			
Hospitalizations	0.01	.57	[−0.09, 0.11]
Reported illnesses	0.15	.90	[−0.07, 0.44]
Risk factors (IRR)			
Grading in region	1.03	.65	[0.88, 1.16]
Grading in restaurant	0.95	.45	[0.81, 1.07]
Illness complaints (rate)			
All complaints	0.13	.18	[−0.03, 0.30]
Probable/confirmed outbreaks	0.09	.03	[0.01, 0.14]

Note. For reported illnesses, the zip samples are the evaluation zips. For other estimates, the sample comprises all zip codes. Rates are per 10,000 residents; 95% CI is the confidence interval derived from inverting the randomization inference test for a one-sided alternative. IRR = incidence rate ratio.

Yet contrary to claims about the benefits of grading, the earlier implementing regions, if anything, experienced higher surges in foodborne illness in 2017. King County experienced a rise in foodborne illness in the third quarter of 2017,¹⁵ and the first implementing region, if anything, exhibited a higher uptick in foodborne illness than the other regions. This finding is inconsistent with claims that restaurant grading can mitigate the impact of an outbreak by improving food sanitation practices. Using the model in

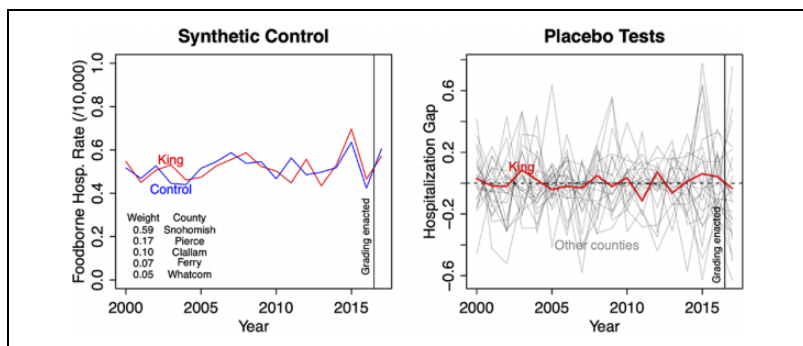


Figure 8. Synthetic control results using Comprehensive Hospital Abstract Reporting System hospitalization data for all counties in Washington. A synthetic control county was developed for King County using the weights displayed in the lower left corner of the left panel. The difference between King County's mean squared prediction error pre- and posttreatment is not statistically different from the placebo distribution of 24 comparison counties (one-tailed, $p = .38$).

Equation 1, we find a nonstatistically significant *increase* of .15 cases per 10,000 associated with the rollout of restaurant grading or an increase of 14% over the pretreatment quarterly average baseline (one-tailed $p = .90$).¹⁶ Inverting the randomization test provides us with a 95% CI of $[-0.07, 0.44]$ cases per 10,000.

Synthetic controls from outside King County. As an additional check, we also take a synthetic controls approach to compare King County's foodborne hospitalization rate (per 10,000) to that of other counties in Washington state (Abadie et al., 2010).¹⁷ We create a "synthetic control" county that is comparable to King County in pretreatment foodborne hospitalization trends from 2000 to 2016. If restaurant grading caused a reduction in hospitalizations of 20%, we should observe a drop in King County relative to its synthetic control.¹⁸

Figure 8 presents results. The left panel shows that the synthetic control tracks King County's pretreatment foodborne hospitalization trends very closely ($MSPE_{pre} = 0.003$). The group is comprised primarily of King County's immediate neighbors, Snohomish and Pierce counties, which make substantive sense. Contrary to a large grading effect, we observe no appreciable difference between King County and comparison counties in 2017. The right panel shows the hospitalization gap between King County and its synthetic control in red, along with the hospitalization gaps for all

placebo treated counties.¹⁹ The hospitalization gap is comparable to that of placebo-treated Washington counties, and we fail to reject the null hypothesis using permutation inference with a difference-in-difference test statistic (one-tailed $p = 0.38$).

Taking these findings together, we hence detect no statistically significant effects of restaurant grading on foodborne disease incidence. Given the null results, we now attempt to quantify the policy relevance of nonzero effect sizes that would suggest public health benefits and still be consistent with our data. The lower bound of the 95% CIs provides us with a *probable* best-case reduction scenario for the policy. The 95% CI in a randomization inference framework provides the lowest and highest null effect sizes that would be retained in a one-tailed test at $\alpha = .05$ given the data. Due to the discreteness of the p values in the randomization distribution, the CIs tend to be conservative (i.e., coverage is at least 95%; Agresti, 2003). Therefore, the lower bounds should be interpreted as optimistic about the policy's potential public health benefits.

The Washington State Department of Health's (2013) food safety goals targeted an overall decrease of 34% averaged across the four most prominent foodborne illness types by 2020. The lower bound on the hospitalization effect, $-.09$, would represent a decrease of 11% from the baseline rate. In our data, we observed about one foodborne illness hospitalization for every five reported foodborne illness cases in the baseline period. Therefore, we estimate that an 11% decrease in hospitalizations would account for about 6.5% of Washington State's reduction targets. By the same logic, a lower bound of a 7% reduction in foodborne illness rates would account for 20.6% of Washington State's 2020 reduction targets. Since hospitalizations are presumably a subset of reported illnesses, 20.6% of the target reduction represents the most optimistic picture our data could support. Meeting one fifth of the target through restaurant grading alone would be a policy relevant effect. That said, our data suggest that such an effect could be observed through chance alone.

In addition, to assess the policy relevance of the effect, one also has to weigh the substantial costs of implementing restaurant grading for the county. These costs included roughly 50% of the food program director's time during the 2 years of preparation for restaurant grading, core staff time (e.g., 15%–20% of time of first two inspections to explain the system to operators), industry and community relations around the initiative, back-end data management, the design process, and a full audit of posting compliance, with ongoing maintenance costs. While these costs are difficult to

monetize, one estimate is that between 5% and 20% of program energy (with a budget of US\$11 million) was devoted to the initiative in the 2 years of development. In the most optimistic scenario, if 7% of cases of the two most prevalent illnesses (salmonellosis and campylobacteriosis) were avoided, the annual public health benefit, based on United States Department of Agriculture (USDA) estimates would be roughly US\$220k.²⁰ Ongoing costs would of course be significantly lower, but it is worth emphasizing that our evidence, if anything, is consistent with an *increase* in foodborne illnesses.

Risk Factors

Because foodborne illnesses are often underreported, we now turn to direct evidence of “risk factors” in restaurants. If the threat of a poor restaurant grade improves sanitation practices, we should be able to observe these directly in food handling practices (e.g., handwashing, cross-contamination). Several studies examine whether restaurant grading reduced the number of violation citations, but the grading system may directly affect whether an inspector *cites* a violation for the same behavior (Boehnke & Graham, 2000; Kovacs et al., 2018). Evidence of sharp disparities around grading thresholds, as displayed in Figure 2, suggests that grading may have a perverse public health effect: creating the *appearance* of all “A”-graded establishments, while actually decreasing operator incentives to correct behavior. Because this was a major concern for Public Health, we examine the effect of grading on risk factors as measured by the independent risk factor study. Recall that independent contractors randomly sampled restaurants in each region with a short-form inspection checklist that we designed for this process (see Appendix E).

To generate the test statistic for randomization inference, we fit the following set of quasi-likelihood overdispersed Poisson models for the count of violations cited $y_{ijk} = \sum_{l=1}^L y_{ijkl}$ over all violations $l = 1, \dots, L$:

$$y_{ijk} \sim \text{Poisson}(\mu_{ijk}), \quad (6)$$

$$\mu_{ijk} = n_{ijk} \exp(\alpha_i + \beta_j + \gamma_k + \theta T_{ij}), \quad (7)$$

$$\text{Var}(y_{ijk} | \mathbf{X}) = \phi \mu_{ijk}, \quad (8)$$

where n_{ijk} is an offset for the total number of *possible* violations observed for establishment k in region i and year-quarter j , α_i are region fixed effects, β_j are year-quarter fixed effects, γ_k are inspector fixed effects, T_{ij} is a dummy variable taking on the value of 1 if restaurant k in region i is subject

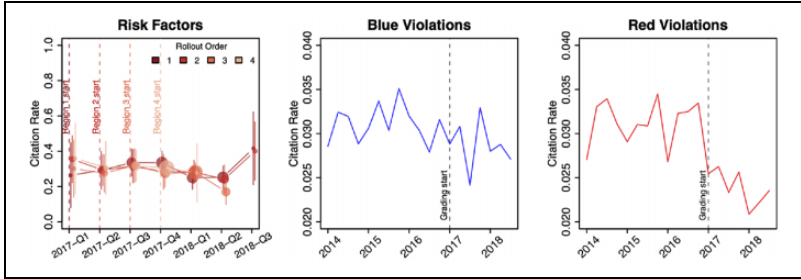


Figure 9. Risk factor citation rates by grading rollout region and quarter from the start of grading in the first quarter of 2017 to mid-third quarter of 2018 (left panel). Points are sized by the number of inspections in each region and quarter, which varied based on staffing and targets to achieve balance in the number of inspections in each region pre- and posttreatment. The middle and right panels plot the citation rate for blue and red violations for graded inspections.

to grading in year-quarter j and 0 otherwise, and ϕ is a dispersion parameter to be estimated from the data. We fit an analogous model that examines the effect of an establishment actually being graded:

$$\mu_{ijk} = n_{ijk} \exp(\alpha_i + \beta_j + \gamma_k + \theta G_{ijk}), \quad (9)$$

where G_{ijk} is an indicator variable for whether restaurant k in region i has received a graded inspection (and presumably, a grade placard) as of year-quarter j . For both models, θ is the test statistic we use for the randomization inference procedure.

The middle rows of Table 3 present results. We find that the average citation rate of restaurants in graded regions increased by 3% over that of restaurants in nongraded regions, but this difference is not statistically significant (one-tailed $p = .65$). Figure 9 corroborates this by showing relative stability in citation rates over the observation window. Restaurants that had a graded inspection exhibited a 5% decrease in their average citation rate, but this difference is again not statistically significant (one-tailed $p = .45$). In sum, we find no evidence that restaurant grading caused restaurants to improve sanitation practices. It is important to note that these results are independent of any direct citation effect of grading, as these inspectors were not administering the King County health code and issued no reports to operators after observation.

We illustrate the possibility of such citation effects by plotting the time series of low-risk (blue) and high-risk (red) violations in the right panels of

Figure 9. Recall that in contrast to other grading systems, King County relies only on high-risk (red) violations to grade establishments. While blue violations remain cited at comparable rates before and after grading, red violations decrease substantially.²¹ Given considerable ambiguity around what constitutes a high-risk classification (see Ho, 2017a; Ho et al., 2018) and the stability of risk factors and blue citations, these findings suggest that grading reduces the citation of existing high-risk violations.

This analysis also shows limitations to identifying the causal effect of grading by studying only health code violations (e.g., Wong et al., 2015). Such designs are akin to studying whether a grade cutoff for military enlistment has “effects” on student performance when cutoffs may have direct effects on instructor grading (Rojstaczer & Healy, 2012).

Using the type of logic presented in the Foodborne Disease Incidence subsection to interpret the 95% CI lower bounds as optimistic cases for the policy effect, we consider the policy relevance of a 12%–19% decrease in the average citation rate for restaurants. This would amount to a little over one fewer violation observed per inspection on average. The threshold for a return visit in King County is 35 high-risk violations. We therefore conclude that the most optimistic benefit of the policy for sanitation practices is minimal compared to the number of violations it takes for the county to classify a restaurant as out of compliance with its food safety standards. The much more policy-relevant quantity may be the evidence of citation effects—particularly in light of the risk factor study and stability in non-critical (ungraded) citations—whereby restaurant grading may cause inspectors to under-cite critical violations.

Restaurant complaint data

Finally, we investigate whether grading affected consumer foodborne illness complaints or the rate at which the county substantiates complaints to classify an outbreak. We refer to an outbreak as substantiated when the county deemed the outbreak “probable” or “confirmed” after investigation of complaints. We aggregate complaints and outbreaks investigated by the county by region and quarter. We fit the following least squares model for (1) all consumer complaints and (2) probable/confirmed outbreaks tied to consumer complaints:

$$Y_{ij} = \alpha_i + \beta_j + \theta T_{ij} + \varepsilon_{ij}, \quad (10)$$

where Y_{ij} is the complaint rate or outbreak rate of region i in year-quarter j , α_i are fixed effects for regions, β_j are fixed effects for year-quarters, and T_{ij}

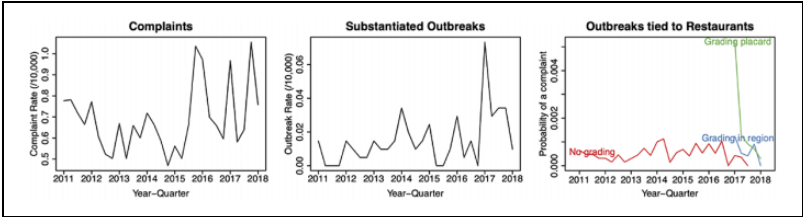


Figure 10. Foodborne illness complaints per 10,000 people (left), substantiated foodborne illness outbreaks per 10,000 people (middle), and the probability of a restaurant receiving a consumer complaint tied to a substantiated outbreak (right), from 2011 to 2018.

is a dummy variable taking on the value of 1 if region i is subject to grading in year-quarter j and 0 otherwise.

The bottom rows of Table 3 present results. We find that complaints increased by .13 per 10,000 people alongside the rollout of restaurant grading, but the association was not statistically significant (95% CI [−0.03, 0.30]). Restaurant grading was, however, statistically significantly associated with the detection of .09 more outbreaks per 10,000 people (two-sided $p = .03$, 95% CI [0.01, 0.14]). The left panel of Figure 10 shows the time series of complaints from 2011 to 2018, which depicts a secular increase in complaints. The middle panel shows that substantiated outbreaks increased starting in 2017, and the right panel shows that this was particularly driven by complaints about graded restaurants in region 1, which drives the statistically significant outbreak effect.

We consider several interpretations of these results. First, the interpretation most favorable to the grading system is that it increased the volume of complaints to Public Health. Grading placards contained quick response codes that led to a Public Health website linking to instructions on how to report foodborne illness. Because underreporting makes foodborne surveillance challenging, such reports may have facilitated the investigation of otherwise isolated complaints. There are, however, weak points to this interpretation. The increase in the overall volume of complaints was not statistically significant, and the same increases in substantiated complaints are not as evident in regions that implemented grading in subsequent quarters.

A second possibility is that media coverage of the launch of the grading system independently affected how consumers engaged with the department. This effect may have been short term and primarily manifested itself

in the first quarter in the region where placards became available. This interaction between media and popular attention on the launch and availability of placards would explain the patterns we see but suggests that the impact on public engagement may not be sustained.

A third possibility is that these effects are simply artifacts of shellfish norovirus outbreaks that disproportionately affected Region 1 in the first quarter. The number of outbreaks driving these effects, for instance, is low. Yet, we observe few prior outbreaks—particularly for low-grade illnesses such as norovirus—that exhibit such a sharp uptake in independent complaints. Moreover, the probability that an outbreak is substantiated is directly affected by public reporting of food poisoning cases. The fact that Region 1 had a higher substantiation rate may be a function of increased public engagement.

Last, it is also possible that King County's decision to post ongoing foodborne illness investigations in late 2015, as well as the hiring of a full-time foodborne illness/enteric disease investigator in late 2016, affected the probability that a complaint was substantiated. That said, these factors would seem to affect all grading regions equally, and we observe some evidence that the increase in substantiated complaints tracks the staggered rollout. But perhaps the investigative resources interacted with the greater public consciousness of food safety from the grading placards, providing a boost to investigations early on in the implementation. We have marginal evidence that consumers, on average, reported cases one day earlier than before grading.²² Earlier reports make for easier epidemiological investigation, which might explain the higher rate of substantiating complaints.²³

Our results provide suggestive evidence that the grading system increased engagement with public health surveillance, but we are not able to conclusively resolve which of these mechanisms may have driven the apparent increase in substantiated complaints and outbreaks.

Discussion

We hope to have demonstrated through this field experiment how research design can mitigate perceived limitations of RCTs. We believe this model of academic–agency collaboration—whereby academics can help offset the challenges to evaluation—is fertile ground for promoting evidence-based law and policy. While county resources for planning the grading system were substantial (see the Background section and the Foodborne Disease Incidence subsection), the county budget for the randomized *evaluation* was insubstantial.²⁴ The only substantial line item was the expense of contractor inspectors for the independent risk factor analysis, but the evaluation of

health outcomes could have proceeded without that element.²⁵ The chief reasons why this bargain could be struck were the following: (a) the county committed in principle to rigorously evaluating the policy initiative and agency leadership was committed to managing the process; (b) our academic team, in turn, was willing to invest considerable research efforts to develop a tailored research design that addressed practical, political, and ethical concerns; and (c) administrative data, which had never been analyzed for evaluating these kinds of policy initiatives, were available.

Our trial provides critical evidence about the value, implementation, and administration of grading systems. Most importantly, it grounds expectations citizens, policy makers, and academics should have of these kinds of disclosure initiatives. Contrary to often cited data about the health benefits, we find no evidence that foodborne illnesses decreased. Our trial also shows an unintended consequence of grading, namely the citation effect documented in the Risk Factors subsection, which is consistent with grade inflation in other jurisdictions (Ho, 2012; Kovacs et al., 2018). While these findings may inform the decision of whether to adopt a grading system, our results also inform prospective decisions within the confines of a given grading system. Management and training of staff should focus on avoiding the unintended citation effect. The results also strongly support King County's methodology to calibrate the grading thresholds each year to measure *relative* performance in a way that is more robust to citation effects (Ashwood et al., 2016). The evidence about increased public engagement suggests that grading systems may have the most salutary effects by public education about food safety principles but that sustaining such public engagement remains a challenge.

To be sure, our case study also illustrates limitations for how research design can facilitate RCTs of policy implementation. First, one challenge to customized research designs is scalability. It took considerable learning, mutual respect, understanding, and artful negotiation among key stakeholders to craft our solution. From that perspective, navigating political barriers is a nonpecuniary cost to conducting a randomized evaluation in this context. Nonetheless, we believe this is a fruitful avenue, as there remains significant under-provision of randomized trials in vast areas of law and policy (Abramowicz et al., 2011; Gerber et al., 2014; Green & Thorley, 2014; Margetts, 2011). Good observational design can also involve considerable research investment (Rubin, 2008), and natural experiments are often unavailable for critical policy questions of interest. One of the virtues of a large-scale field experiment—given the amount of planning invested in this research collaboration—is that these kinds of trials are less likely to be left to file drawers (Rosenthal, 1979).

Second, another potential limitation of our study is external validity. Our evaluation only provides valid estimates of the effect in King County. Unlike other grading systems, King County's system is based on multiple routine inspections, uses only high-risk violation, adjusts for interinspector differences, and does not use a letter grade. The treatment effect we recover may hence not generalize to other jurisdictions. While this may seem like a substantial limitation, it turns out that grading systems in fact vary dramatically across nearly all jurisdictions (Ho, 2012, p. 603, table 2). Some jurisdictions do not post grades until appeals are resolved, others disclose only whether a restaurant passed an inspection, some conduct only one routine inspection per restaurant annually, and yet others have no meaningful grade variation at all (Ho, 2012, p. 603, table 2). The concern about external validity is then more about the challenge of policy learning given the highly decentralized nature of retail food safety enforcement, which generates such dizzying variation, in the United States. Taken to the extreme, this concern undercuts the notion of states as laboratories of democracy. Nor is this problem one observational studies can address (LA's effect may be limited to LA's system and population), and there are in fact few alternative credible observational designs (Ho et al., 2019). The most promising path forward, in light of the rapid adoption of grading systems, would be to adapt our design to evaluate implementation across many jurisdictions.

Third, one limitation of the stepped-wedge design is that the treatment might be different in earlier periods than in later periods. For instance, if the agency is learning about how to implement the policy (e.g., informing operators), grading in the first period may not have the same effect as grading in a later period. We hence view a stepped-wedge design as complementary to implementation analyses better known in the public administration field. Such analyses should reveal operational challenges in rolling out a policy. Fortunately, the extent of planning in King County meant that operational issues were limited.²⁶ If such challenges exist, greater effects should be detectable in subsequent regions. Ideally, when a policy is anticipated to be difficult to implement, the design would include more stages of rollout, thereby facilitating both implementation analyses and increasing statistical power of the evaluation.

Fourth, although our design attempted to best account for potential nonindependence across restaurants, it is theoretically possible that treatment effect spillovers weakened power. Competition between neighboring restaurants for the same diners is one potential mechanism. Recent work examining occupational safety inspections, for example, has shown that disclosure of particularly poor inspection performance (accompanied by a press release) can have deterrence effects on geographically proximate firms, driven by their desire to

maintain their reputations with the public (Johnson, 2020). This explicit policy of “regulation by shaming,” however, was quite different from King County’s goal. By design, all restaurants within a given municipal area received placards at roughly same time, not just the worst performers, and the department did not issue press releases about poor performance. Without press releases, we might expect this type of spillover to be quite geographically concentrated, which would be captured by our use of subregions in a county that covers over 2,300 square miles. Combined with the well-established fact that restaurant competition is predominantly local (Parsa et al., 2011), this aspect of the design should give some assurance that spillovers do not negate the ability to detect a grading effect. Moreover, the synthetic control analysis that compares King County to other Washington counties also provides no evidence that King County’s foodborne illness rate dropped by 20% upon adoption.

Fifth, the fact that we largely find no statistically significant effects of the grading policy does not completely rule out policy-relevant effects in the most optimistic scenario. Although our study was adequately powered to detect a 20% decrease in foodborne illness rates, smaller decreases may have gone undetected and still have had policy relevance to the county. In the most optimistic scenario consistent with our findings, the county may have been brought one fifth of the way to its goal based on the state’s targeted decrease of 34% in foodborne illness by 2020. However, we cannot rule out that this “progress” would have occurred simply due to chance in the absence of restaurant grading. In the meantime, we can quantify concrete costs of the restaurant grading program and implementation.

Last, there are surely ways in which our algorithm could have been refined. For instance, rather than making a binary decision on whether to include a zip code in an area, continuous weights could be used, although the search space would become quite large. (In the current setup, the possible number of partitions already exceeds the estimated number of seconds since the big bang.) Practically, time was a considerable constraint. After institutional approvals and data transfer, our team had little time to provide the regions to the county to plan operations.

In short, while we acknowledge these potential limitations, we believe there are few better alternatives than a randomized evaluation to credibly assessing the benefits of such substantial policy initiatives. Research design—in this case a synthetic cluster randomized stepped-wedge design—holds great promise for overcoming practical barriers to evidence-based law and policy.

Appendix A

Sample Placard

Public Health

Seattle & King County



FOOD

SAFETY RATING

This business has received a food safety rating of:



EXCELLENT

Possible ratings:



NEEDS TO IMPROVE



OKAY



GOOD



EXCELLENT

For more information, text:
 4683 1111 4683 1111 4683 1111 4683 1111
 更多資訊，請發送簡訊：
 更多資訊，請發送簡訊：
 더 자세한 정보는, 연락해
 04 683 1111 4683 1111 4683 1111 4683 1111
 Untuk informasi lebih lanjut, hubungi
 Para obtener más información, envíe un mensaje de texto a



text king food to 468311

Business number: _____
 Permit #: _____

Date: _____



Petty Hayes, RNH, MPH,
 Director of Public Health —
 Seattle & King County

www.KingCounty.gov/FoodSafetyRating

Figure A1. Example of food safety placard.

Appendix B

Selection of Discharge Codes

MEMORANDUM

From:

To: Food Safety Group, Centers for Disease Control and Prevention (CDC)
 Date: September 20, 2016
 Subject: Foodborne Illness Selection for Prospective Evaluation

Based on earlier conversations, we understand that CDC does not currently retain a list of all foodborne illness diagnostic codes. As you know, Seattle & King County will be implementing a restaurant grading system in 2017 and in order to evaluate the effects of this policy change, we aimed to identify foodborne illnesses that restaurant sanitation practices are likely to affect.

We write to inform you of what our research has revealed and also seek your input on whether you agree with our selection of diseases. We hope that this research may be independently helpful to CDC's monitoring efforts of food safety.

Based on our review of the public health literature, we propose to focus on illnesses that are:

- (a) foodborne in at least 80% of cases (based on Mead¹ or Scallan²), and
- (b) travel-related in less than or equal to 20% of cases (based on Scallan²).

We include the travel restriction, as we do not expect restaurant grading to impact the incidence of conditions frequently acquired abroad. The travel restriction, however, may not be appropriate for monitoring non-restaurant related food safety conditions.

For reference, the Appendix summarizes data at the pathogen level in Table A and at the ICD code level in Table B. Our review has concluded that the ICD-9 and ICD-10 codes corresponding to these illnesses are:

	ICD-9	ICD-10
Salmonella (non-typhoidal)	0030 - 0039	A020 - A029
Campylobacter	00843	A045
E. Coli (shiga-toxin producing)	00804	A043
Vibriosis	0054, 00581	A053, A055
Listeriosis	0270	A320 - A329
Yersiniosis	00844	A046, A282
Trichinosis	124	B75
Other foodborne bacterial illnesses	0050-0053, 00589, 0059	A050-A052, A058, A059

In addition, in the ICD-10 system, B9622, B9623, B962 are secondary diagnostic codes that identify shiga-toxin producing strains of *E. coli* as the cause of illness and B967 is a secondary code that identifies *Clostridium perfringens* as the cause of illness.

We hope this list is helpful and would welcome any input you might be able to provide.

Appendix C

Implementation Regions

In order to minimize the operational burden of implementing the grading system, we optimized how zip codes that are not in the set of evaluation regions were added to implementation regions. We balance the number restaurants subject to grading:

$$R_a = \frac{1}{I(I-1)} \sum_{i=1}^I \sum_{l \neq i}^I |n_i^a - n_l^a|, \tag{11}$$

where n_i^a and n_l^a are the number restaurants in region i ($i = 1, \dots, I$) or region l given allocation a . The right panel of Figure 3 shows the resulting evaluation and implementation regions that were delivered to the county for the actual grading rollout.

Appendix D

Trimming

Table D1. Balance covariates used to ensure that the comparison regions were operationally and politically feasible for the county. Population and household income are the 5-year estimates from the 2014 American Community Survey. Restaurant covariates are calculated from the baseline period 2014–2015. Gradable restaurants are restaurants eligible to receive a grade according to King County’s internal classification. The Seattle municipal area includes 37 zip codes.

Unit	Covariate	Thresholds
Region	(a) Population	Within 2.5th and 97.5th pctl
	(b) Number of gradable restaurants	
	(c) Number of full-time employees assigned to gradable restaurants	
	(d) Mean household income	
Allocation	(e) Percentage of included zips that are in the Seattle municipal area	Within 2.5th and 97.5th pctl
	(f) Percentage of restaurants in included zips that are inspected by the Central Office	
	(g) Percentage of total zips in the county that are included in the allocation	At least 50%

Note. Pctl = percentile.

Appendix E

Risk Factor Study Inspection

Environmental Health Services Division

401 Fifth Avenue, Suite 1100
Seattle, WA 98104

206-263-9566 Fax 206-296-0189

TTY Relay: 711

www.kingcounty.gov/health

Public Health 
Seattle & King County

Foodborne Illness Risk Factor Study Data Collection Form Marking Instructions

1) Handwashing: This item is marked IN Compliance only when employees are observed using proper handwashing techniques at appropriate times. Note: in the rare occasion where there is no need to wash hands and therefore handwashing is not observed, this item may be marked IN.

a. This item is marked IN Compliance only when employees are observed using proper handwashing techniques.

b. This item is marked IN Compliance only when employees are observed washing their hands at appropriate times.

If either (a) or (b) are marked OUT, (1) must be marked OUT.

2) BHC: This item is marked IN Compliance only when employees are observed using suitable utensils or gloves to prevent bare hand (or arm) contact with RTE foods or are observed following an alternative procedure from an approved plan that allows bare hand contact with RTE foods. Note: there are very few such approved plans in KC.

3) Cross Contamination: This item is marked IN Compliance only when raw animal foods are separated from RTE foods during storage, preparation, and display. Cross Contamination due to soiled equipment or lack of hand washing is not included here.

4) Clean/Sanitize: This item is marked OUT of Compliance when food contact surfaces are not cleaned and sanitized before uses with different types of meat (e.g. placing beef on a cutting board contaminated with raw chicken) or when changing from raw animal foods to RTE food. This item is marked IN Compliance when food contact surfaces and utensils are clean to sight and touch and sanitized before use.

5) TAAC: If PHF is above 41°F (45°F for raw shell eggs) or below 135°F (130°F for roasts) and the food is not in cold or hot holding equipment, and the establishment is not using time as a public health control (TAAC), mark this item OUT of Compliance. This item should also be marked OUT if the establishment is using TAAC and:

- The food is unlabeled or improperly labeled, or
- The food is not served or discarded within 4 hours after removal from temperature control, or
- Approved written procedures are not available.

This item may be marked NA when the establishment does not use TAAC.

This item may be marked NO when the establishment uses TAAC, but not at the time of inspection.

6) Cold Holding: Except for PHF that are in the cooling process or have been recently out at room temperature during active preparation (less than 2 hours), this item should be marked OUT of Compliance if any PHF that is being stored under refrigeration (mechanical or using ice as a coolant) are found to exceed critical limits by 1 degree Fahrenheit or more.

- a. Enter the number of mechanical refrigeration units that are being used to cold hold PHF.
- b. Enter the number of mechanical refrigeration units that are holding PHF 1 – 5° F above critical limits.
- c. Enter the number of mechanical refrigeration units that are holding PHF 6° F or above critical limits.

7) Hot Holding: This item should be marked OUT of Compliance when PHF are stored in hot holding equipment and is found to be 134°F or below [130°F for whole beef roasts, corned beef roasts, pork roasts, and cured port roasts such as ham that are cooked according to the cooking chart found at WAC 246-215-03400 (2)].

This item may be marked NO when the establishment does hot hold PHF, but none are being held hot during the time of inspection.

This item may be marked NA when the establishment does not hot hold PHF.

- a. Enter the number of hot holding units that are being used to hot hold PHF.
- b. Enter the number of hot holding units that are holding PHF 1- 5° F below critical limits.
- c. Enter the number of hot holding units that are holding PHF 6° F or below critical limits.

8) Cooling: This item should be marked OUT of Compliance if any of the cooling methods (time/temperature, shallow pan, or pieces of meat) are found to be out of compliance. See WAC 246-215-03515 (1). For this study, we are not assessing items (2), (3), or (4). For establishments using time/temperature, the item should be marked OUT of Compliance if the food does not cool from 135°F to 70°F in two hours or less; or if the food does not cool from 135°F to 41°F in six hours or less. Discussions with the PIC and Food Workers along with observations should be used to determine compliance.

This item may be marked NO when the establishment does cool PHF, but proper cooling temperature and time parameters cannot be determined during the length of the inspection.

This item may be marked NA when the establishment does not cool cooked PHF.

- a. Enter (+) if the establishment uses the Time and Temperature method as specified in the Food Code, WAC 246-215-03515 (1) (a) and (b). Enter (-) if this method is not used.
- b. Enter (+) if the establishment uses the shallow pan method as specified in the Food Code, WAC 246-215-03515 (c) (i). Enter (-) if the shallow pan method is not used.
- c. Enter (+) if the establishment cools pieces of meat as specified in the Food Code, WAC 246-215-03515 (c) (ii).

Appendix F

Quarterly Versus Yearly Criterion

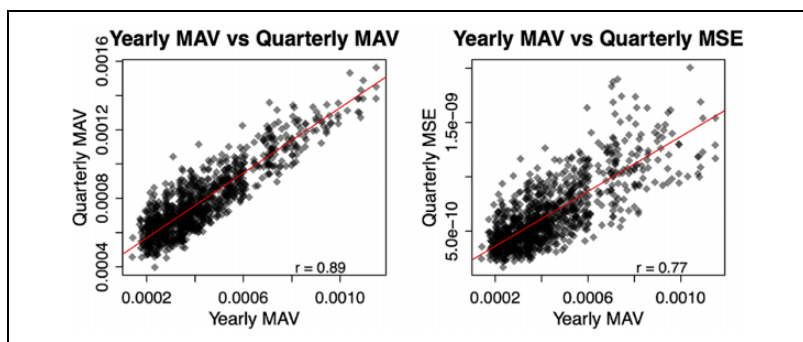


Figure F1. Criterion values in a random sample of results from the region algorithm optimized on yearly mean absolute value (see Equation 2). Balance calculated at the yearly level is highly correlated with balance calculated at the quarterly level.

Acknowledgments

The authors thank Zoe Ashwood for research assistance; Jenny Lloyd (King County) and Karen Wong (CDC) for help in validating hospitalization discharge codes; David Engelskirchen (FDA) and Katey Kennedy (FDA) for help in training our independent inspection team; Phil Wyman and Eyob Mazengia for designing the streamlined risk factor checklist; Ki Straughn, Jennifer Jessen, Hasina Wong, and Dalila Zelkanovic for carrying out the risk factor study; Meagan Kay and Sargis Pogojans for help in assembling notifiable conditions and complaints data; and Carrie Cihak, Eyob Mazengia, and Sargis Pogojans for helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This evaluation was supported exclusively by internal funding from Stanford University with no compensation from King County or any other entity.

ORCID iD

Cassandra Handan-Nader  <https://orcid.org/0000-0003-4868-2509>

Notes

1. RAND's landmark 15-year health care experiment, for instance, cost upward of US\$640 million in present dollars (Green-berg & Shroder, 1998).
2. This kind of implementation evaluation is distinct from the idea of "implementation analyses," whose principal goal is to assure faithful implementation of a policy initiative (Hollister, 2009). Instead, we argue that sequencing inherent in implementation facilitates rigorous randomized evaluation. We do not argue that our research design addresses the full panoply of criticisms of randomized controlled trials (see, e.g., Cartwright, 2007; Deaton & Cartwright, 2017). Our design addressed central constraints that otherwise would likely have led King County to compare only outcomes before and after the intervention.
3. The committee also noted that any grading system should not rely on an "appeals system" for poor grades and avoid introducing reinspections solely for regrading establishments.
4. Jin and Leslie (2003) compared municipal adoptions of grading in Los Angeles (LA) in 1998 to the rest of California before and after LA adopting grading.

Stemming from a power analysis conducted for King County's evaluation, Ho et al. (2019) found that the reduction in LA stemmed from the largest state recorded salmonella outbreak, which affected Southern California right before LA began grading.

5. In addition, diagnosing whether an illness is foodborne is challenging. So-called syndromic surveillance data for each hospital visit, for instance, provide only prediagnostic information, hence presenting too much risk of Type I error. By and large, foodborne illnesses present with nonspecific symptoms (e.g., diarrhea, vomiting).
6. Adapting the database and electronic tablets used for field inspections, in particular, proved challenging as did the training of inspectors around how to communicate the design of the grading system.
7. These regions are King County district court regions. It is worth noting that these regions do not map cleanly to zip codes, another downside of using existing boundaries when constrained to particular primitive units. We assigned border zip codes in a way that preserved contiguity between regions and was most faithful to the county's district map (see <https://www.kingcounty.gov//media/depts/elections/elections/maps/district-court-map/district-court.ashx?la=en>).
8. We do so by matching zip codes to municipalities using Coven (2012). These municipalities represent postal office classifications and hence differ slightly from conventional understandings of King County municipalities.
9. Because the neighborhood district boundaries are defined by census tract, we assigned each zip code to the district based on area.
10. King County has a population of roughly 2 million, and we tested population floors between 50,000 and 200,000. We also tested varying the minimum number of restaurants. We found that a population threshold of 50,000 provided the best pretreatment balance. Population was matched at the zip code tabulation area (ZCTA) level from the 2014 American Community Survey 5-year estimates.
11. We also tried a stochastic (as opposed to greedy) version of the algorithm, where a decrease in mean absolute value (MAV) due to adding a zip code would be stochastically accepted but found little improvement.
12. We also considered rerunning the region growth algorithm with this improved metric as the optimization criterion but found that it was unnecessary. First, our existing allocations ranked by the mean squared error (MSE) criterion provided over 80% power at $m \leq 40$ as shown in the left panel of Figure 5. Second, a random sample of allocations revealed that the yearly MAV criterion is highly correlated with quarterly MAV ($r = .89$) and the quarterly MSE ($r = 0.77$). See Figure 12 in Appendix F. Third, due to institutional approvals, we were under

considerable practical time constraints, as we received the foodborne illness data in November 2016 and were expected to deliver the final regions by early December 2016. We note that future work may well improve on the efficiency of the sampling algorithm, but for our purposes, the basic goal of improving power via synthetic regions was met.

13. When the randomization distribution was large, we used Monte Carlo simulation with 120 simulations for a given set of region allocations, 24 rollout orders for a single-region allocation, two effect sizes (0% and 20%), and 2 years.
14. To maximize hospitalization volume, we use all zip codes within the four regions.
15. This uptick appears to have been driven by campylobacter. Two small campylobacter outbreaks in this time period have been investigated by the county, one related to a restaurant and one from home cooking (see <https://www.king-county.gov/depts/health/communicable-diseases/disease-control/outbreak/private-event-august-2017.aspx> and <https://www.kingcounty.gov/depts/health/communicable-diseases/disease-control/outbreak/cafe-juanita.aspx>).
16. The units are limited to the evaluation zip codes in each of the four regions.
17. Reported illness data at the county level for 2017 was not yet available from the Washington State Department of Health at the time the analysis was conducted.
18. More formally, let $Y^1 = \{y^1_1, \dots, y^1_J\}$ represent the hospitalization rates in King County from year $j \in \{j, \dots, J\}$ and Y^0 represent an $I \times J$ matrix of hospitalization rates for I control counties over J years. The synthetic control calculates weights w_i for control county i which minimize the pretreatment mean squared prediction error:

$$\text{MSPE}_{\text{pre}} = \frac{1}{n_{\text{pre}}} \sum_{j < 2017} \left(y^1_j - y^0_j \right)^2, \quad (12)$$

where $y^0_j = \sum_{i=1}^I Y^0_{ij} w_i$ is weighted foodborne hospitalization rate for the synthetic control in year j , y^1_j is the foodborne hospitalization rate for King County in year j , and n_{pre} is the number of pretreatment years. We use permutation inference to test for a treatment effect with the following test statistic τ :

$$\tau = \frac{1}{n_{\text{post}}} \sum_{j \geq 2017} \left(y^1_j - y^0_j \right) - \frac{1}{n_{\text{pre}}} \sum_{j < 2017} \left(y^1_j - y^0_j \right), \quad (13)$$

where n_{post} is the number of posttreatment years.

19. To conduct permutation inference, we limit the reference distribution to counties where MSPE_{pre} is within 30 times that of King County. Thirty is the minimum threshold at which we can obtain a reference distribution large enough to reject the null hypothesis at $\alpha = .05$.

20. Here, we use the per case estimate of cost from the USDA, see <https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses.aspx>, and average annual counts of 1,000 cases of campylobacteriosis and 250 cases of salmonellosis in King County, see <https://kingcounty.gov/depts/health/communicable-diseases/disease-control.aspx>
21. Using the linear regression described in Equation 12, the coefficient on θ_2 , the grading rollout for red violations, is -0.83 ($SE = .32$).

$$Y_{ijkt} = \theta_1 \text{Grading}_{ij} + \theta_2 \text{Grading}_{ij} \times \text{Points}_t + \alpha_{it} + \beta_{jyear^t} + \gamma_{kt} + \varepsilon_{ijkt}, \quad (14)$$

where Y_{ijkt} is the number of violation points in for establishment i in year-quarter j from inspector k for violation type t (red or blue), α_{it} represents fixed effects for establishment i and violation type t , β_{jyear^t} represents fixed effects for the year of year-quarter j and violation type t , γ_{kt} represents fixed effects for inspector k and violation type t , θ_1 represents the grading rollout treatment for establishment i in year-quarter j for blue points, and θ_2 represents the grading rollout treatment for establishment i in year-quarter j for red points, as Points_t is dummy variable that takes on the value of 1 if violation type t is red and 0 otherwise.

This difference, however, is not statistically significant when using fixed effects for year-quarter j and type t ($\theta_2 = -.51$, $SE = .38$). This is because the largest drop in red points relative to blue points occurred simultaneously for all regions at the beginning of 2017, rather than in a quarterly fashion that tracked the rollout of grading over the course of 2017. The beginning of 2017 also coincided with an area rotation for inspectors, so these results are most plausible with inspectors citing fewer violations in anticipation of grading. The stability of the blue violations, given uncertainty about the red and blue distinction, suggests that the effect is driven by changes in inspector behavior not restaurant behavior.

22. Using randomization inference for time from meal to reporting, trimming outliers over 2 standard deviations from the mean reporting time, yields a p value of .07.
23. We thank Sargis Pogosjans for suggesting this analysis.
24. The main costs for the core RCT component were as follows: (a) US\$1,450 for hospital discharge data from Washington State and (b) additional staff time for training due to the staggered rollout.
25. The budget allowed for 900 hr for three contractors, comprising roughly US\$86k.

26. The main operational issue was about the validity of grade calculations, having to do with the database syncing issues, and compliance by operators to posting the grade. These were documented only in 2018, and hence do not invalidate the stepped-wedge design, and are limited in scope. If, however, the county changed the system, for instance, by introducing substantial penalties for noncompliance with placard posting, our evaluation may not speak to this effect.

References

- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505.
- Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2), 495–510.
- Abramowicz, M., Ayres, I., & Listokin, Y. (2011). Randomizing law. *University of Pennsylvania Law Review*, 159(4), 929–1005.
- Agresti, A. (2003). Dealing with discreteness: Making “exact” confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods in Medical Research*, 12, 3–21.
- Ashwood, Z. C., Elias, B., & Ho, D. E. (2016). Improving the reliability of food safety disclosure: A quantile adjusted restaurant grading system for Seattle-King County [Manuscript].
- Assefa, S. (2010). Census geography to neighborhood equivalency files. <https://www.seattle.gov/opcd/population-and-demographics/geographic-files-and-maps#2010census>.
- Baio, G., Copas, A., Ambler, G., Hargreaves, J., Beard, E., & Omar, R. Z. (2015). Sample size calculation for a stepped wedge trial. *Trials*, 16, 354.
- Ben-Shahar, O., & Schneider, C. E. (2011). The failure of mandated disclosure. *University of Pennsylvania Law Review*, 159, 647–749.
- Ben-Shahar, O., & Schneider, C. E. (2014). *More than you wanted to know: The failure of mandated disclosure*. Princeton University Press.
- Boehnke, R. H., & Graham, C. (2000). *International survey on public posting of restaurant inspection reports, and/or grade card posting schemes based upon health inspections*. Region of Ottawa-Carleton Health Department.
- Bothwell, L. E., Greene, J. A., Podolsky, S. H., & Jones, D. S. (2016). Assessing the gold standard—Lessons from the history of RCTs. *New England Journal of Medicine*, 374(22), 2175–2181.
- Bridgeland, J., & Orszag, P. (2013, July/August). Can government play moneyball? *The Atlantic*, 312, 63–66.

- Bubb, R. (2015). TMI? Why the optimal architecture of disclosure remains TBD. *Michigan Law Review*, 113(6), 1021–1042.
- Buck, S., & McGee, J. (2015). *Why government needs more randomized controlled trials: Refuting the myths*. Laura and John Arnold Foundation.
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90, 414–427.
- Cameron, A. C., & Miller, D. L. (2014). A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50, 317–371.
- Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, 2(1), 11–20.
- Centers for Disease Control and Prevention. (2011). *CDC estimates of foodborne illness in the United States* (Technical Report CS218786-A). Author.
- Coalition for Evidence-Based Policy. (2015). Demonstrating how low-cost randomized controlled trials can drive effective social spending: Project overview and request for proposals. *Project Overview and Request for Proposals*. <http://coalition4evidence.org/wp-content/uploads/2014/02/Low-cost-RCT-competition-December-2013.pdf>.
- Collins, G. (2010, July 28). For restaurants it's like going back to school. *New York Times*, A15; New York City, 2009 Executive Budget.
- Coven, D. S. (2012). Free Zipcode Database: Unique Zipcode [data file]. <http://federalgovernmentzipcodes.us>
- Deaton, A., & Cartwright, N. (2017). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.
- Desposato, S. (2015). *Ethics and experiments: Problems and solutions for social scientists and policy professionals*. Routledge.
- Easterly, W. (2009). Development experiments: Ethical? Feasible? Useful? *Aid Watch: Just Asking That Aid Benefit the Poor*. <http://www.nyudri.org/aidwatch/archive/2009/07/development-experiments-ethical-feasible-useful>
- Elias, B., & Ho, D. H. (2016). Government under review. *Boston Review*. <http://bostonreview.net/editors-picks-us-books-ideas/daniel-e-ho-becky-elias-government-under-review>
- Esarey, J., & Menger, A. (2017). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods, First View*, 1–35.
- Felt, C. (2017, February 23). King County's Changing Demographics. <https://www.kingcounty.gov//media/depts/executive/performance-strategy-budget/documents/pdf/RLSJC/2017/Feb23/KingCountyDemographics022317>.
- Fernald, L. C. H., Gertler, P. J., & Neufeld, L. M. (2008). Role of cash in conditional cash transfer programmes for child health, growth, and development: An analysis of Mexico's Oportunidades. *The Lancet*, 371(9615), 828–837.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

- Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.
- Fung, A., Graham, M., & Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge University Press.
- The Gambia Hepatitis Study Group. (1987). The Gambia hepatitis intervention study. The Gambia Hepatitis Study Group. *Cancer Research*, 47(21), 5782–5787.
- Gerber, A. S., Green, D. P., & Kaplan, E. H. (2014). The illusion of learning from observational research. In D. L. Teele (Ed.), *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* (pp. 9–32). Yale University Press.
- Grambsch, P. (2008). Regression to the mean, murder rates, and shall-issue laws. *The American Statistician*, 62(4), 289–295.
- Green, D. P., & Thorley, D. R. (2014). Field experimentation and the study of law and policy. *Annual Review of Law and Social Science*, 10(1), 53–72.
- Greenberg, D. H., & Shroder, M. (1998). *The digest of social experiments* (2nd ed.). The Urban Institute Press.
- Greiner, D. J., & Matthews, A. (2016). Randomized control trials in the United States legal profession. *Annual Review of Law and Social Science*, 12(1), 295–312.
- Gueron, J. M., & Rolston, H. (2013). *Fighting for reliable evidence*. Russell Sage Foundation.
- Haskins, R., & Margolis, G. (2014). *Show me the evidence: Obama's fight for rigor and results in social policy*. Brookings Institution Press.
- Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ*, 350, h391.
- Ho, D. E. (2012). Fudging the nudge: Information disclosure and restaurant grading. *Yale Law Journal*, 122(3), 574–688.
- Ho, D. E. (2015). Randomizing... What? A field experiment of child access voting laws. *Journal of Institutional and Theoretical Economics*, 171(1), 150–170.
- Ho, D. E. (2017a). Does peer review work: An experiment of experimentalism. *Stanford Law Review*, 69, 1–119.
- Ho, D. E. (2017b). Equity in the bureaucracy. *Irvine Law Review*, 7, 401–451.
- Ho, D. E., Ashwood, Z. C., & Handan-Nader, C. (2019). New evidence on information disclosure through restaurant hygiene grading. *American Economic Journal: Economic Policy*, 11(4), 404–428.
- Ho, D. E., & Imai, K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *Journal of the American Statistical Association*, 101(475), 888–900.

- Ho, D. E., Sherman, S., & Wyman, P. (2018). Do checklists make a difference? A natural experiment from food safety enforcement. *Journal of Empirical Legal Studies*, 15(2), 242–277.
- Hollister, R. (2009). Reply comments, the role of random assignment in social policy research. *Journal of Policy Analysis and Management*, 28(1), 178–180.
- Hooper, R., Teerenstra, S., de Hoop, E., & Eldridge, S. (2016). Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, 35, 4718–4728.
- House of Commons Science and Technology Committee. (2006). Scientific advice, risk and evidence based policy making. <https://publications.parliament.uk/pa/cm200506/cmselect/cmsctech/900/900-i.pdf>
- Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2), 182–191.
- Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2), 399–423.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Jin, G., & Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics*, 118(2), 409–451.
- Johnson, M. S. (2020). Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws. *American Economic Review*, 110(6), 1866–1904.
- Kasza, J., Hemming, K., Hooper, R., Matthews, J. N. S., & Forbes, A. B. (2017). Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Statistical Methods in Medical Research*, 28(3), 703–716.
- King County Board of Health. (2017, January 19). Rule & Regulation 17-01. https://www.kingcounty.gov/depts/health/board-of-health/~/_media/depts/health/board-of-health/documents/regulations/BOH-regulation-17-01.ashx
- King County Executive Office. (2015). Statistical Profile of King County. <https://www.kingcounty.gov//media/depts/executive/performance-strategy-budget/regional-planning/Demographics/KC-profile2016.ashx?la=en>
- King County Executive Order. (2014). Advancing equity and social justice through development of a strategic innovation priority plan and executive department actions. Document code no. ACO 9-2 (AEO). <https://www.kingcounty.gov/about/policies/executive/administrationaeco/aco92aeco.aspx>
- King, G., Gakidou, E., Ravishankar, N., Moore, R. T., Lakin, J., Vargas, M., Téllez-Rojo, M. M., Hernández Ávila, J. E., Hernández Ávila, M., &

- Hernández Llamas, H. (2007). A “Politically Robust” experimental design for public policy evaluation, with application to the Mexican Universal health insurance program. *Journal of Policy Analysis and Management*, 26(3), 479–506.
- Kovacs, B., Lehman, D., & Carroll, G. R. (2018). Boundary kinking in public grading schemes: The effects of tie strength in social relationships. *Academy of Management Proceedings*, 2018(1), 10704.
- Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1), 1–18.
- Loewenstein, G., Sunstein, C. R., & Golman, R. (2014). Disclosure: Psychology changes everything. *Annual Review of Economics*, 6(1), 391–419.
- Margetts, H. Z. (2011). Experiments for public management research. *Public Management Review*, 13(2), 189–208.
- Mdege, N. D., Man, M.-S., Taylor, C. A. (nee Brown), & Torgerson, D. J. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, 64(9), 936–948.
- Nathan, R. P. (2008). The role of random assignment in social policy research. *Journal of Policy Analysis and Management*, 27(2), 401–415.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–480.
- Parsa, H. G., Self, J., Sydnor-Busso, S., & Yoon, H. J. (2011). Why restaurants fail? Part II—The impact of affiliation, location, and size on restaurant failures: Results from a survival analysis. *Journal of Foodservice Business Research*, 14(4), 360–379.
- Perry, J. L. (2012). How can we improve our science to generate more usable knowledge for public professionals? *Public Administration Review*, 72(4), 479–482.
- Public Health—Seattle & King County. (2014). *King county food protection program review: Final report*. Author.
- Ravallion, M. (2012). Fighting poverty one experiment at a time: Poor economics: A radical rethinking of the way to fight global poverty: Review essay. *Journal of Economic Literature*, 50(1), 103–114.
- Rodrik, D. (2008). The new development economics: We shall experiment, but how shall we learn? Harvard Kennedy School Faculty Research Working Papers Series RWP08-055. <https://j.mp/2n75Zrp>.
- Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, 114(7), 1–23.

- Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3), 286–327.
- Rosenbaum, P. R. (2002b). *Observational studies*. Springer.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.
- Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591–593.
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840.
- Seiver, O. H., & Hatfield, T. H. (2000). Grading systems for retail food facilities: A risk-based analysis. *Journal of Environmental Health*, 63(3), 22.
- Sylvester, E. (1980, May 25). Making sure your eating places are A-OK: Inspectors rate S. D. Restaurants, *L.A. Times*, A1.
- Todak, N., White, M. D., Dario, L. M., & Borrego, A. R. (2018). Overcoming the challenges of experimental research: Lessons from a criminal justice case study involving TASER exposure. *Evaluation Review*, 42(3), 358–385.
- Washington State Department of Public Health. (2013). *Foodborne Illnesses* (Technical report). Washington, USA.
- Weil, D., Fung, A., Graham, M., & Fagotto, E. (2006). The effectiveness of regulatory disclosure policies. *Journal of Policy Analysis and Management*, 25(1), 155–181.
- Wiant, C. J. (1999). Scores, grades, and communicating about food safety. *Journal of Environmental Health*, 61(9), 37–39.
- Winston, C. (2008). The efficacy of information policy: A review of Archon Fung, Mary Graham, and David Weil's full disclosure: The perils and promise of transparency. *Journal of Economic Literature*, 46(3), 704–717.
- Wong, M. R., McKelvey, W., Ito, K., Schiff, C., Bryan Jacobson, J., & Kass, D. (2015). Impact of a letter-grade program on restaurant sanitary conditions and diner behavior in New York City. *American Journal of Public Health*, 105(3), e81–e87.
- Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data*. The MIT Press.

Author Biographies

Cassandra Handan-Nader is a PhD student in the Department of Political Science at Stanford University and a graduate student fellow in the Regulation, Evaluation and Governance Lab at Stanford University.

Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law at Stanford Law School; a professor of political science at Stanford University; a senior fellow at Stanford Institute for Economic Policy Research; Director of the Regulation, Evaluation, and Governance Lab at Stanford University; and an associate director at Stanford Institute for Human-Centered Artificial Intelligence at Stanford University.

Becky Elias is a former manager in the Food Protection Program, Environmental Health Services at Public Health—Seattle & King County. The work was completed during her tenure at Public Health—Seattle & King County. She currently serves as Director of Global Retail Food Safety and Quality Assurance at Starbucks, which had no role in this study.