JOURNAL OF
EMPIRICAL LEGAL STUDIES

Journal of Empirical Legal Studies
Volume 15, Issue 2, 00–00, June 2018

Do Checklists Make a Difference? A Natural Experiment from Food Safety Enforcement

*Daniel E. Ho, Sam Sherman, and Phil Wyman**

Inspired by Atul Gawande's bestselling *Checklist Manifesto*, many commentators have called for checklists to solve complex problems in law and public policy. We study a unique natural experiment to provide the first systematic evidence of checklists in law. In 2005, the Public Health Department of Seattle and King County revised its health code, subjecting half of inspection items to a checklist, with others remaining on a free-form recall basis. Through in-depth qualitative analysis, we identify the subset of code items that remained substantively identical across revisions, and then apply difference-in-differences to isolate the checklist effect in more than 95,000 inspections from 2001–2009. Contrary to scholarly and popular claims that checklists can improve the administration of law, the checklist has no detectable effect on inspector behavior. Making a violation more salient by elevating it from "noncritical" to "critical" status, however, has a pronounced effect. The benefits of checklists alone are considerably overstated.

I. Introduction

Atul Gawande's bestselling *Checklist Manifesto* argued that dramatic improvements in complex decisions can be achieved by cheap and easy-to-use checklists (Gawande 2009). Generalizing from the evident success in surgery, Gawande claimed that checklists—simple, enumerated lists of information or steps—could improve complex decisions in "virtually any endeavor" (2009:158). While Gawande acknowledged potential limitations, others quickly latched on: "Checklists are hot [and have] captured the imagination of the media and have inspired publication of a manifesto on their

*Address correspondence to Daniel E. Ho, William Benjamin Scott and Luna M. Scott Professor of Law, Stanford Law School; Professor (by courtesy) of Political Science; Senior Fellow, Stanford Institute for Economic and Policy Research; 559 Nathan Abbott Way, Stanford, CA 94305; email: dho@law.stanford.edu. Sherman is J.D. Candidate, Stanford Law School; Wyman was Health and Environmental Investigator, Public Health—Seattle and King County and is Environmental Health Specialist, Snohomish Health District.

We thank Mark Bossart for help securing data, Janet Anderberg for insights into the health code revision, Becky Elias, Ki Straughn, and Dalila Zelkanovic for help in developing and carrying out a simplified inspection model, Claire Greenberg and Aubrey Jones for research assistance, and Becky Elias, Atul Gawande, Sandy Handan-Nader, Tim Lytton, Oluchi Mbonu, Bill Simon, David Studdert, and two anonymous referees for comments.

power in managing complexity" (Davidoff 2010:206).[1] Or as the *Times* put it, checklists are "a classic magic bullet" (Aaronovitch 2010).

Fueled in part by Gawande's manifesto, scholars have, in turn, advocated for greater adoption of checklists in law and public policy. Richard Thaler and Cass Sunstein consider the checklist as part of choice architecture (Thaler et al. 2012:432). Bill Simon (2012:394) urged "[p]rotocols of the sort that Gawande developed," suggesting provocatively that checklists might have averted major financial and accounting scandals, such as options backdating and Enron's collapse. In criminal law, scholars have proposed checklists to help police and prosecutors comply with *Brady* obligations (Findley 2010; Griffin 2011), to improve defense lawyering (Bibas 2011), and to help judges weigh testimony (Levine 2016). A Federal Courts Study Committee recommended a legislative drafting checklist (Maggs 1992) and checklists have been proposed to improve election administration (Douglas 2015), safety regulation (Hale & Borys 2013), and education (Institute of Education Sciences & National Center for Education Statistics, Department of Education 2014). In the regulatory context, Gawande himself argued that checklists would enable building inspectors and industry alike to make "the reliable management of complexity a routine" (Gawande 2009:79). If so, checklists might address longstanding challenges with decentralized law enforcement and administration (Ho 2017:5–12; Mashaw 1985), where few proven techniques exist to address dramatic disparities in complex legal decisions (Ho & Sherman 2017).

Yet the legal reception to checklists has not been uniform. Checklists, rather than solving the problem of bureaucracy, may create it (Ho 2017:57, 82–83; Simon 2012:395). Ford (2005) associates checklists with the pathologies of rigid command-and-control regulation, and McAllister (2012) warns that checklists may lead to mechanistic application failing to detect real risk. Warned Stephen Cutler, Director of the Division of Enforcement at the Securities and Exchange Commission: "don't fall victim to a checklist mentality" (Cutler 2004). Opening a congressional hearing on emergency preparedness, entitled "Beyond the Checklist," Representative Langevin admonished: "Our nation's leaders are not seeing the big picture. Instead, they are driving our departments and agencies to focus so much effort on checking boxes that there is barely time left to actually combat a potential pandemic" (Langevin 2007). Across areas, checklist skepticism abounds. Koski (2009:831) critiqued the checklist mentality in school reform, noting that parties in structural litigation "develop[ed] more paperwork, checklists, and bureaucratic oversight, essentially 'teacher-proofing' the reform process." In family law, Ross (2003:206) concluded that "[t]here is . . . no easy checklist that agencies and courts can rely upon to predict whether a child can be safe with his or her parents." In securities regulation, Ford (2008:29) opined that "[c]reating ever-longer lists of prohibited behavior or checklists of compliance-related best practices will not be effective if the basic culture of the firm does not foster law-abiding behavior." In food safety, Stuart

[1] Gawande's book treatment rightly acknowledges conditions under which checklists may not work, such as a command-and-control ethos (2009:75–76), the role of institutional culture (2009:160), and difficulties in implementation (2009:170), yet many commentators simply focus on the claims that checklists are "simple, cheap, effective" (2009:97) and "quick and simple" (2009:128).

and Worosz (2012:294) documented one "veteran food safety auditor" who claimed that "about 70% of the items on food safety checklists are irrelevant to food safety."

A core empirical premise of proponents is that checklists, by *reminding* individuals, should improve the detection of errors, violations, or mistakes for complex tasks (Gawande 2007, 2009; Hales et al. 2008; Thaler et al. 2012). Despite the litany of reform proposals and sharp disagreement over their effectiveness, as we articulate below, this empirical premise has never been subject to rigorous empirical scrutiny in law and public policy.

We address this lacuna by providing the first evidence in a legal context of the effectiveness of checklists. We study a unique and inadvertent natural experiment in food safety enforcement conducted by the Public Health Department in Seattle and King County. In 2005, the department revised its health code, subjecting roughly half of violations (classified as critical) to a checklist to be employed by inspectors, but leaving other (noncritical) violations to a free-form recall basis, as was the case before for all violations. Most importantly, while the revision added and revised a range of items, through in-depth qualitative research, including engagement with staff responsible for implementing the code revision, we classify the subset of violations that remained *identical* before and after the revision.

This approach offers several advantages to understanding the causal effect of checklists on code citation. First, a common critique of checklist studies is that the introduction of the checklist is accompanied by greater training, teamwork, and a reorientation of tasks, hence confounding the intervention (Bosk et al. 2009). By focusing on identical code items—for which there was no additional training and change in instructions—we are able to isolate the effect of a checklist independent of other factors. Second, because the checklist applied only to half the violations, a difference-in-differences design accounts for temporal changes in sanitation and enforcement practices, as well as selection bias in who adopts checklists. Our design therefore addresses two common limitations to extant checklist studies: before-after design studies cannot rule out temporal changes (e.g., increased managerial commitment to quality improvement) and cross-sectional comparisons cannot rule out preexisting quality differences between adopting and nonadopting institutions (see De Jager et al. 2016).

Third, our design uses internal administrative data to account for assignments of establishments to inspectors. Our treatment effect is hence identified by changes in how the same establishments are inspected by the same inspectors before and after the code revision, for the subset of code items that remained identical except for the checklist format. This has a considerable advantage in the inspection context, where there are well-documented differences in inspection stringency by inspector (Ho 2012, 2017). Last, because the checklist format was merely incidental to the substantive changes in the health code, our design rules out Hawthorne effects, whereby research subjects may improve performance due to the knowledge of being observed (Haynes et al. 2009:497).

We study data from 95,087 inspections, and 15 identical violations (i.e., 1,426,305 potential violations) scored in each inspection by 37 inspectors serving before and after the intervention for 2001–2009. Applying logistic difference-in-differences regressions with inspector and establishment fixed effects, we find that the checklist has no appreciable effect on inspection behavior. Due to the large sample size, our estimates are relatively precise, allowing us to rule out moderately sized effects. On the other hand, elevating the salience of a violation has a pronounced effect, even when the violation remains the same.

These findings have considerable implications for regulatory enforcement and how we conceive of "choice architecture" for regulatory behavior. As we spell out below, checklists are no panacea and cannot resolve core issues of administrative law and the design of regulatory institutions. The benefits of checklists are likely to emanate from the process of focusing and simplifying core responsibilities on areas of highest risk, not checklists alone. Our findings hence suggest that proper prioritization of risk factors and code item simplification are critical to improving regulatory enforcement.

The article proceeds as follows. Section II situates this study in the existing literature on checklists. Section III provides institutional background on food safety enforcement. Section IV describes the research design and data. Section V presents results. Section VI describes limitations and Section VII concludes with implications.

II. Prior Evidence on Checklists

The value of a natural experiment such as King County's is best seen in the context of the existing evidence base on checklists.

The core evidence stems from healthcare. Recent meta-analyses and systematic reviews suggest a weakly positive effect of surgical checklists on postoperative outcomes (Bergs et al. 2014; De Jager et al. 2016; Gillespie et al. 2014; Lau & Chamberlain 2016; Lyons & Popejoy 2014). The rigor of observational designs and estimated effects, however, varies considerably, with only a limited number of randomized controlled trials. De Jager et al. conduct a systematic review of 25 studies of the Surgical Safety Checklist promoted by the World Health Organization, and conclude that "poor study designs" mean that "many of the positive changes associated with the use of the checklist were due to temporal changes, confounding factors and publication bias" (2016:1842).

To understand the case for exporting checklists to law and public policy, we searched ProQuest and Google Scholar for all quantitative studies purporting to assess the effect of a checklist on outcomes in any domain.[2] We focus on outcomes external to the checklist (e.g., complication rates in surgery), so we do not include studies of interrater reliability, construct validity, checklist adoption rates, and compliance. Our search yielded 79 prior studies, which we classified by research design and substantive area. We classified research design into three types: (1) randomized controlled trials, where the checklist was randomly assigned; (2) observational studies where the principal comparison is between outcomes before and after the adoption a checklist; and (3) observational studies where the principal comparison is cross-sectional (e.g., hospitals adopting vs. not adopting a checklist). For each study we collected information about the main outcome(s) studied, and recorded whether the finding was statistically significantly

[2] As another check, we also searched Web of Science to verify that results were comparable.

Table 1: Summary of Checklist Literature by Substantive Area, Research Design, and Findings

		Findings				
Area	*Design*	*Neg.*	*Insig.*	*Pos.*	*Not Rep.*	*Prop.*
Medicine	Randomized	1	9	11	0	0.19
	Observational: cross-section	2	4	11	0	0.15
	Observational: before-after	1	27	28	4	0.55
Software	Randomized	2	5	0	4	0.10
	Observational: cross-section	0	1	0	0	0.01
	Observational: before-after	0	0	0	0	0.00

Notes: "Neg." indicates statistically significant negative findings, "Insig." indicates statistically insignificant findings, "Pos." indicates statistically significant positive findings, and "Not Rep." indicates that insufficient information was reported. Each cell count represents the number of main findings of 110 outcomes in 78 studies. The table excludes one observational study on the effect of checklists on scuba diving that had statistically insignificant results.

positive, negative, inconclusive, or not reported. Table 1 presents an overview of the studies along these dimensions.

We make three observations on the state of the literature. First, the overwhelming majority of work on checklists (68 of 79 studies) is limited to healthcare. The only other active research area is in software (10 studies), with one study on scuba diving (excluded from Table 1). Surprisingly, while proponents commonly invoke the use of checklists in aviation and aeronautics and product manufacturing, we were not able to identify any published empirical findings about the effect of checklists in those sectors. For instance, Hales and Pronovost (2006) point to an aviation finding that electronic checklists reduced errors by 46 percent compared to paper-based checklists, but this aviation finding stems from an unpublished Boeing simulation study about errors in checklist completion (Boorman 2001). A commonly cited reference for airline aviation checklists provides only anecdotal evidence of checklist usage (Degani & Wiener 1990). Hersch (2009:8) notes that checklists were widely adopted in cockpits "despite little formal study of its effectiveness in flight testing" and Gordon et al. (2012) argue that safety improvements in aviation stemmed not from checklists, but from a cultural shift away from pilot-driven to team-based management after major airline crashes. In product manufacturing, Hales and Pronovost discuss checklists employed by the Food and Drug Administration for drug manufacturing and by the Canadian Food Inspection Agency for food safety, but cite no evidence of the impact of such checklists on outcomes. Of course, usage is not effectiveness. Pronovost credits James Reason's *Managing the Risks of Organization Accidents* as inspiring the idea for a patient safety checklist, but Reason mentions checklists as only one of 20 quality assurance tools (Reason 1997:130). Notwithstanding calls to expand checklists to law and public policy and the ambitions for checklists to reduce failures "from medicine to finance, business to government" (Gawande 2009:13), there is virtually no evidence base pertaining directly to law and public policy.

Second, roughly 70 percent of studies are observational designs, either before-after or cross-sectional comparisons. Each of these designs has limitations. Consider the seminal eight-hospital study on checklists, which used a before-after design. The eight participating hospitals were selected from dozens of applicants, and the complication rate of surgical procedures decreased from 11 percent to 7 percent after checklists were introduced (Haynes et al. 2009). But the decision to apply for a quality improvement program may itself reflect increased managerial commitment to improving safety. Gains might therefore stem from managerial commitment, not the checklist; adoption may be endogenous.[3] In addition, as the study recognized, Hawthorne effects were "difficult to disentangle" (2009:497). Cross-sectional comparisons of (1) hospitals that adopt checklists and those that do not (e.g., Jammer et al. 2015) or (2) cases with high checklist compliance versus low checklist compliance (e.g., Mayer et al. 2016) may be confounded by quality differences across hospitals and physicians. The positive correlation between checklist compliance and fewer complications could reflect differences in the care of surgical teams, not the causal effect of a checklist (Mayer et al. 2016: 63 ["Compliance with interventions like a checklist can thus be a surrogate of an underlying positive team culture"]). Strikingly few studies have attempted to deploy more sophisticated designs. For instance, the eight-hospital study could have randomly selected hospitals from applicants and compared outcomes between treatment and control groups before and after the intervention, thereby potentially adjusting for temporal changes that could be the result of improved managerial commitment.

Third, the checklist intervention is typically accompanied by a host of other changes. Peter Pronovost, for instance, expanded the checklist intervention to a far more substantial "comprehensive unit-based safety program" to focus on teamwork and cultural change (Pronovost & Vohr 2010:78–112). Or consider again the eight-hospital study. As described by Gawande, who made site visits to participating hospitals as part of the intervention:

> The hospital leaders committed to introducing the concept systematically. They made presentations not only to their surgeons but also to their anesthetists, nurses, and other surgical personnel. We supplied the hospitals with their failure data so the staff could see what they were trying to address. We gave them some PowerPoint slides and a couple of YouTube videos ... For some hospitals, the checklist would also compel systemic changes—for example, stocking more antibiotic supplies in the operating rooms .... Using the checklist involved a major cultural change, as well—a shift in authority, responsibility, and expectations about care—and the hospitals needed to recognize that.[4]

As an initial matter, it is unclear how to conceive of the treatment. On the one hand, the introduction of a checklist is *confounded* by many other changes (e.g., training modules, site visits, observation of failure data, communication, teamwork, leadership commitment to change). On the other hand, the checklist intervention may be considered a *compound* treatment, with the checklist only as a component of the treatment.

[3] If complication rates are high in one year, for instance, these rates may drive managers to institute a new quality improvement program, but complication rates may decrease by regression-to-the-mean alone.

[4] Gawande (2009:145–46).

Perhaps from the perspective of hospital care, the distinction is of less practical import: if the treatment works in totality and is cost justified, it should be adopted.

But from the perspective of exporting the intervention to other domains, isolating the effect and mechanism of the checklist is critical. Policy proposals may differ dramatically if the gains stem from aspects distinct from the checklist. For instance, if gains are explained by providing hospitals with failure data, the optimal intervention may be less about a checklist than about providing data to learn from prior mistakes. Similarly, policy interventions may hinge on the mechanism of the checklist effect. A principal mechanism is that checklists serve as a *reminder*, "especially with mundane matters that are easily overlooked" (Gawande 2007). Checklists as reminders may be much more easily exported to other domains and should be verifiable in our setting where individuals, not teams, tally violations. An alternative mechanism is that checklists *empower* lower-level staff to monitor for errors, such as by shifting decisional authority from physicians to nurses (Gawande 2007; Simon 2012; Pronovost & Vohr 2010). If a cultural shift toward team-based learning is critical, the checklist alone may have limited success in law and regulation, where serious peer learning is limited (Ho 2017; Simon 2015; Noonan et al. 2009).[5]

Our study contributes to the understanding of checklists by isolating the checklist effect in a new domain that lends itself to a credible observational design.

III. Institutional Background

A. *Food and Drug Administration*

While retail food safety enforcement is principally conducted at the local level, the Food and Drug Administration (FDA) publishes a Model Food Code, revised every few years, aimed to help states, counties, and municipalities in promulgating codes for food safety. Food inspections are widely recognized to constitute complex risk assessments, given the myriad of conditions at retail establishments. The Model Food Code from 2001 comprises nearly 600 pages to account for such complexity (U.S. Food and Drug Administration 2001).

In 2002, a committee of the Conference for Food Protection (CFP), a nonprofit organization comprised of industry, government, academic, and consumer representatives that provides guidance on food safety, recommended adopting a new version of the model inspection score sheet, drawing in part on Washington State's form. While prior FDA model forms did not include checklists, the new form placed 27 "critical" violations on a checklist basis and was adopted by the FDA in 2004.[6] FDA instructions evince a desire for this checklist system to remind inspectors of potential critical

[5] In the medical context, Duclos et al. find no evidence that a team training intervention reduced adverse events, and conclude: "Checklist use and [team training] are not, in isolation, magic bullets" (2016:1811).

[6] As we describe below, the CFP also recommended relabeling categories of violations, but for convenience, we continue to refer to these as critical and noncritical categories as shorthand.

violations, but not noncritical violations.[7] CFP discussion focused on the fact that the checklist facilitated consumer understanding of inspection results.[8] Industry members, however, expressed concern that the checklist might reduce the quality of comments written to help facilities come into compliance.[9]

The CFP also questioned the distinction between critical and noncritical violations. The chair of the Inspection Form Committee noted that "designating items as 'critical' in the Food Code and on many inspection reports may be misunderstood in relation to the severity and importance of violations" (Conference for Food Protection 2004:I-11). "Some critical violations are seldom identified as contributing to foodborne illness" (2004: II-28). In an effort to align code items with their evidence base,[10] the committee created new categories to relabel violations. Critical violations became either "risk factors" or "public health interventions." Noncritical violations became "good retail practices." FDA adopted this relabeling, but the model form continued to place the former two categories on a checklist basis and the latter on a free-form basis.

These categories are far from clear. "Risk factors" are "food preparation practices and employee behaviors most commonly reported to the Centers for Disease Control and Prevention (CDC) as contributing factors in foodborne illness outbreaks"[11] and include improper holding temperatures, inadequate cooking, and contaminated equipment. "Public health interventions" are interventions "to protect consumer health,"[12] including hand-washing, time and temperature controls for pathogens, and consumer advisories (e.g., raw fish warning). "Good retail practices" are "[s]ystems to control basic operational and sanitation conditions within a food establishment," which "are the foundation of a successful food safety management system,"[13] including pasteurization of eggs, prevention of cross-contamination, and proper sewage disposal. Ho (2017:54–58) documents substantial confusion among inspectors about the relationship between violations, with inspectors using not only different violations, but also different categories, for the same observed conduct. The CFP itself anticipated this confusion, with the

[7] U.S. Food and Drug Administration (2005, Annex 7, Guide 3-C, pp. 9–10) ("Since the major emphasis of an inspection should be on the Risk Factors that cause foodborne illness and the Public Health interventions that have the greatest impact on preventing foodborne illness, the GRPs have been given less importance on the inspection form and a differentiation between IN, OUT, N.A. and N.O. is not made in this area.").

[8] Conference for Food Protection (2002:7).

[9] Conference for Food Protection (2004:II-18).

[10] Id., at I-11 (attempting to reformulate the inspection form to focus only on violations linked to epidemiological outbreak data).

[11] U.S. Food and Drug Administration (2005:2).

[12] Id., at ii.

[13] Id., at 534.

committee chair emphasizing the need to define the categories to "help regulatory agencies, industry, and the consumer understand which items are a focus or point of emphasis."[14]

More importantly, the FDA's relabeling failed to address the criticism that the distinctions may not track risk (see Jones et al. 2004). None of the references cited by the FDA, for instance, support the idea that raw or undercooked food advisories on menus have any public health impact.[15] The FDA recognizes that "sewage backing up in the kitchen" is a violation of good retail practices.[16] A sewage backup would warrant shutting down an establishment as an imminent public health hazard, and it is unclear why it would not be considered a critical risk factor.

This FDA background turns out to be advantageous for our design in two respects. First, the fluidity across critical versus noncritical categories substantively justifies the choice of noncritical violations as a control group. An exogenous shock in sanitation practices affecting exclusively critical items is highly implausible. Second, the lack of clarity between these items explains why we observe code items that are elevated from noncritical to critical status. As the FDA's relabeling is opaque, we will continue to use the critical/noncritical distinction for clarity of exposition and consistency with common usage.

B. *King County*

The food program in the Public Health Department of Seattle and King County is responsible for food safety enforcement at retail establishments, principally restaurants. Prior to the code revision in 2005, King County employed its own county health code. The food program then employed nearly 40 food safety inspectors for roughly 10,000 establishments. The typical full-service restaurant requires three unannounced inspections per year. During routine inspections, inspectors observe premises and mark violations of roughly 50 health code items on a score sheet. Prior to 2005, roughly half the items comprised "red items," requiring immediate correction, while the other half comprised "blue items," requiring correction by the next inspection.

Starting in 2002, the state department of health began a planning process to amend the state health code and apply it uniformly to localities.[17] After a multiyear planning process, the state adopted the FDA's 2001 Model Food Code, effective May 2005. The principal goals were to update food safety enforcement based on changes in

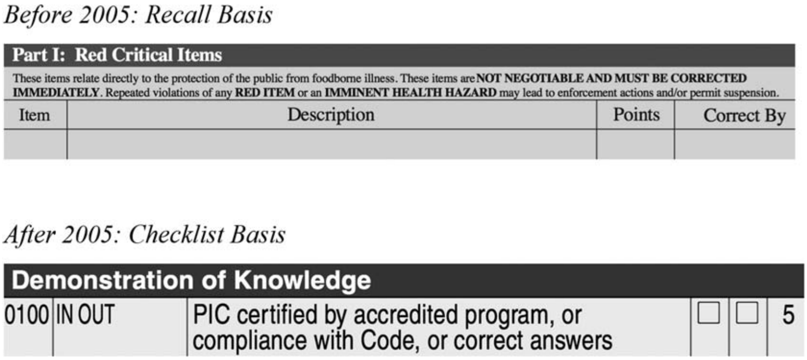
[14] Conference for Food Protection (2004:I-11).

[15] U.S. Food and Drug Administration (2005:283–84). Indeed, none of the references even attempts to evaluate the impact of consumer advisories.

[16] Id., at 534.

[17] In its 2003 session, the state legislature explicitly required the state department to consider the FDA Model Food Code. See RCW 43.20.145(1) ("The state board shall consider the most recent version of the United States food and drug administration's food code for the purpose of adopting rules for food service.").

Figure 1: Inspection score sheet for sample red violation before and after the 2005 code revision.



Before 2005: Recall Basis

Part I: Red Critical Items

These items relate directly to the protection of the public from foodborne illness. These items are **NOT NEGOTIABLE AND MUST BE CORRECTED IMMEDIATELY**. Repeated violations of any **RED ITEM** or an **IMMINENT HEALTH HAZARD** may lead to enforcement actions and/or permit suspension.

Item	Description	Points	Correct By

After 2005: Checklist Basis

Demonstration of Knowledge

0100	IN OUT	PIC certified by accredited program, or compliance with Code, or correct answers	□	□	5

Notes: The 2005 revision required all red items to be marked as "IN" or "OUT" of compliance.

food science and to align state with national standards. Key changes included provisions for cooling, room temperature storage, bare hand contact, and consumer advisories.

In addition to these substantive code changes, the format of the score sheet was revised for consistency with FDA's evolving score sheet. Previously, inspectors wrote down all violations observed in a free-form field of the score sheet, shown in the top of Figure 1. On the new form, depicted in the bottom of Figure 1, inspectors were required to mark all red items as "IN" or "OUT" of compliance (i.e., to check all red violations), but blue items remained on a recall basis as before. In addition, three code items were reclassified from blue to red, which we refer to as "elevated" code items. Instead of the distinction between items that required immediate correction versus items that needed to be corrected by the next visit, the red/blue distinction was reformulated to track the FDA's: red items were "practices or procedures identified as the most prevalent contributing factors of foodborne illness" (critical items), while blue items were "preventative measures to control the addition of pathogens, chemicals, and physical objects into foods" (noncritical items).

From May 1 to October 19, 2005, King County paused regular field activity to train inspectors on the substantive code revisions. No training was provided for previously existing code items or on how to use the checklist.

IV. Identifying Equivalent Violations Across Code Revisions

A. Matching Identical Violations

Our analysis hinges on identifying the subset of code items that remained identical across code revisions. Excluding violations that substantively changed allows us to attribute scoring differences to the checklist format. To do so, we engage in in-depth

Table 2: Example of Code Items Classified as Equivalent Across 2005 Code Revision

	Before 2005	*After 2005*
Blue item	Health Code	Health Code
	“Non-food contact surfaces of equipment are cleaned at such intervals to keep them clean and in a sanitary condition”	“Nonfood-contact surfaces of equipment shall be cleaned at a frequency necessary to preclude accumulation of soil residues”
	Score Sheet	Score Sheet
	“Non-food contact surfaces maintained and clean”	“Non-food–contact surfaces maintained and clean”
Red item	Health Code	Health Code
	Food shall be “[s]afe for human consumption” and “[c]lean, wholesome, and free from spoilage and adulteration”	“[F]ood shall be safe, unadulterated, and ... honestly presented”
	Score Sheet	Score Sheet
	“Foods wholesome, free from spoilage, not adulterated”	“Food in good condition, safe and unadulterated; approved additives”

Notes: The blue item appeared as Item 241 in the pre-2005 score sheet and King County Food Code § 5.22.060(C) and as Item 4300 in the post-2005 score sheet and Washington State Retail Food Code § 4–602.13 (2005), and the red item appeared as Item 102a in the pre-2005 score sheet and King County Food Code § 5.06.010(C) & (E) and as Item 1000 in the post-2005 score sheet and Washington State Retail Food Code § 3–101.11 (2005).

analysis of four sources: (1) the food code before and after the revision, (2) the inspection score sheet, (3) nonpublic marking instructions provided to each inspector, and (4) validation with county and state officials.

We classify code items along two dimensions. First, we classify whether code items before and after 2005 are equivalent, similar, or approximate matches. In many instances, the similarity of code language made this clear. The principal criterion, however, was whether inspection staff treated the violations as identical, which is why we consulted extensively with staff. Each classification decision was validated by a senior staff member responsible for administering the code revision in 2005. Second, we classify whether the number of score sheet items remained the same for the violation, or whether the violation was aggregated or disaggregated. The pre-2005 score sheet, for instance, had two different violations for garbage storage and containers, which were aggregated into a single garbage maintenance violation post-2005.

Table 2 provides examples of red and blue items coded as equivalent across code revisions. The blue item is a violation for whether non-food contact surfaces are clean. Code and score sheet language remains largely identical before and after 2005. The red item is a broad violation for whether food is safe and unadulterated. Again, the code and score sheet language is comparable.[18] Appendix B provides further details on classification of individual code items.

[18] We note that the post-2005 score sheet adds language about “approved additives.” Further qualitative research reveals that this addition does not change the violation in any material sense. Pre-2005, the marking instructions specifically cite King County Food Code § 5.08.030A, which prohibited “sulfiting agents in the food service establishment.” The post-2005 marking instructions provide as the principal example of “unapproved additives” the case of “sulfites being applied to fresh fruits and vegetables.”

Table 3: Code Items by Substantive Similarity of Match and Aggregation Level

	Aggregation				
Substance	*Unchanged*	*Aggregated*	*Disaggregated*	*Total*	*Prop.*
Equivalent	11	4	0	15	0.23
Similar	7	5	1	13	0.20
Approximate	4	3	2	9	0.14
Dropped				17	0.26
Added				12	0.18

NOTES: Because some violations are aggregated and others are disaggregated, the units here are consolidated violations (e.g., a violation represented as two items pre-2005 and one item post-2005 is one unit). Violations that exist only in the pre-2005 score sheet are indicated as "dropped" and violations that are newly added after 2005 are indicated as "added."

Table 3 provides a summary of our classification. Counts represent consolidated violations: for example, if a violation was aggregated in 2005, we treat the constituent code items as a single unit.[19] Fewer than one in four violations are equivalent across code revision, with 11 violations remaining the same both in substance and aggregation. A substantial number of code items bore no similarity across code revisions, but these are not the focus of our investigation.

B. Data

We obtained inspection data from King County, reporting violations cited for each establishment during a routine inspection of full-service establishments.[20] Appendix A describes details on how we cleaned the data. Because there was a substantial code revision implemented late in 2001, we begin the observation period in November 2001. We use May 2005 as the treatment date,[21] and to maintain balanced observation windows before and after the treatment, we end our observation window in April 2009. For simplicity, and because the results are comparable, we focus exclusively on the 15 violations classified as equivalent before and after 2005. All equivalent violations are listed in Appendix B.

To exclude inspector effects, we limit our analysis to 37 inspectors who conducted at least 200 field inspections before and after the code revision.[22] To assess robustness to heterogeneity in establishments, we also collect information on area assignments for each inspector. Inspectors are "rotated" to different area assignments once every few

[19] For example, if violations A and B were aggregated to violation C in 2005, we consolidate A and B pre-2005 to represent violation C. A violation of either A or B would be counted as a violation of C.

[20] These are referred to as "risk III" establishments. Comparable food inspection data are publicly available from 2006 to the present at https://data.kingcounty.gov/Health/Food-Establishment-Inspection-Data/f29f-zza5, but because we are interested specifically in the 2005 code revision, we obtained data going back earlier.

[21] No regular field inspections were conducted from May to mid-October due to training for the code revision.

[22] The annual target number of inspections is 870, so this threshold merely ensures sufficiently long service as a field inspector in the pre and post periods.

Table 4: Summary Statistics of Inspection Data for Matched Violations

		Before	*After*
Citation rate	Red violations	1.2	1.1
	Elevated violation	7.5	11.0
	All blue violations	4.4	4.2
	Aggregated blue violations	3.9	4.1
	Unchanged blue violations	4.8	4.4
	No. of inspectors	37	37
	No. of businesses	10,724	11,621
	No. of inspections	53,803	56,294

NOTES: Citation rate is the mean percentage of violations cited (averaged across inspections and citations). Rates for red and blue violations exclude the violation that is elevated from blue to red in 2005.

years. Areas are based on ZIP codes, and the rotation does not necessarily affect all establishments assigned to an inspector. During the period in question, inspectors retained the same area assignments from January 2004 to December 2006, allowing us to identify effects based solely on changes in red items compared to blue items by the same inspectors within the same areas.[23]

Table 4 provides summary statistics of our data before and after the code revision. The citation rate represents the percentage of inspections during which an item is cited, averaged across the class (red, blue, or elevated). The average blue item, for instance, is cited 4.4 percent of the time before the revision and 4.2 percent after. The average red item is cited 1.2 percent of the time, which decreases to 1.1 percent of the time after the code revision. The one equivalent elevated violation is for employee-wide food safety training, which increases in citation from 7.5 percent to 11 percent of the time, in spite of the fact that the training requirement remained the same. We note here that the item was reclassified from blue to red and moved from the tail end of the score sheet to near the top as the second code item. Its salience in that sense was heightened considerably, and we isolate this effect below.[24]

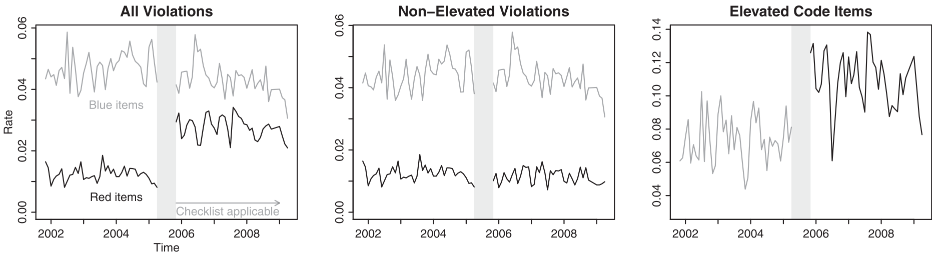
V. Results

We begin by visualizing the citation rates over time. The left panel of Figure 2 plots average citation rates by month for blue and red items that were equivalent over the code revision. Consistent with a difference-in-differences design, the citation rates are roughly parallel in the pretreatment period. The gray vertical bar indicates the training period for the code revision.

[23] Using this alternative end date also allows us to assess robustness to a series of tablet crashes that occurred in November and December 2008. These crashes resulted in large numbers of missing inspections in these months, so we exclude these months from our data. Conditioning on no area rotation also ends the observation period before any of these tablet crashes occurred.

[24] We do not have examples of code items that were only reclassified from blue to red (or vice versa) or only moved up the score sheet. The reclassification from blue to red necessarily moves a code item up higher in the score sheet.

Figure 2: Time series depicting the overall citation rate.



NOTES: Rates are calculated on a monthly basis, with the gray bar depicting the training period for new code items from May–October 2005. Critical code items are shown in black, and noncritical code items are in gray. The left panel illustrates the respective citation rates pooling across all critical, noncritical, and elevated code items. The middle panel pools across only nonelevated code items, while the right panel pools across only elevated items. Only the citation rates of equivalently matched code items are depicted.

If checklists serve as reminders, the citation rate for red items should increase after 2005 relative to blue items. The raw data are suggestive of a checklist effect: the citation rate for red items roughly doubles after 2005. The middle panel, however, shows that the effect disappears when the elevated code item is excluded. Contrary to a checklist effect, there is no breakpoint in red items after 2005. The right panel confirms that the citation increase is driven principally by the elevated code item.

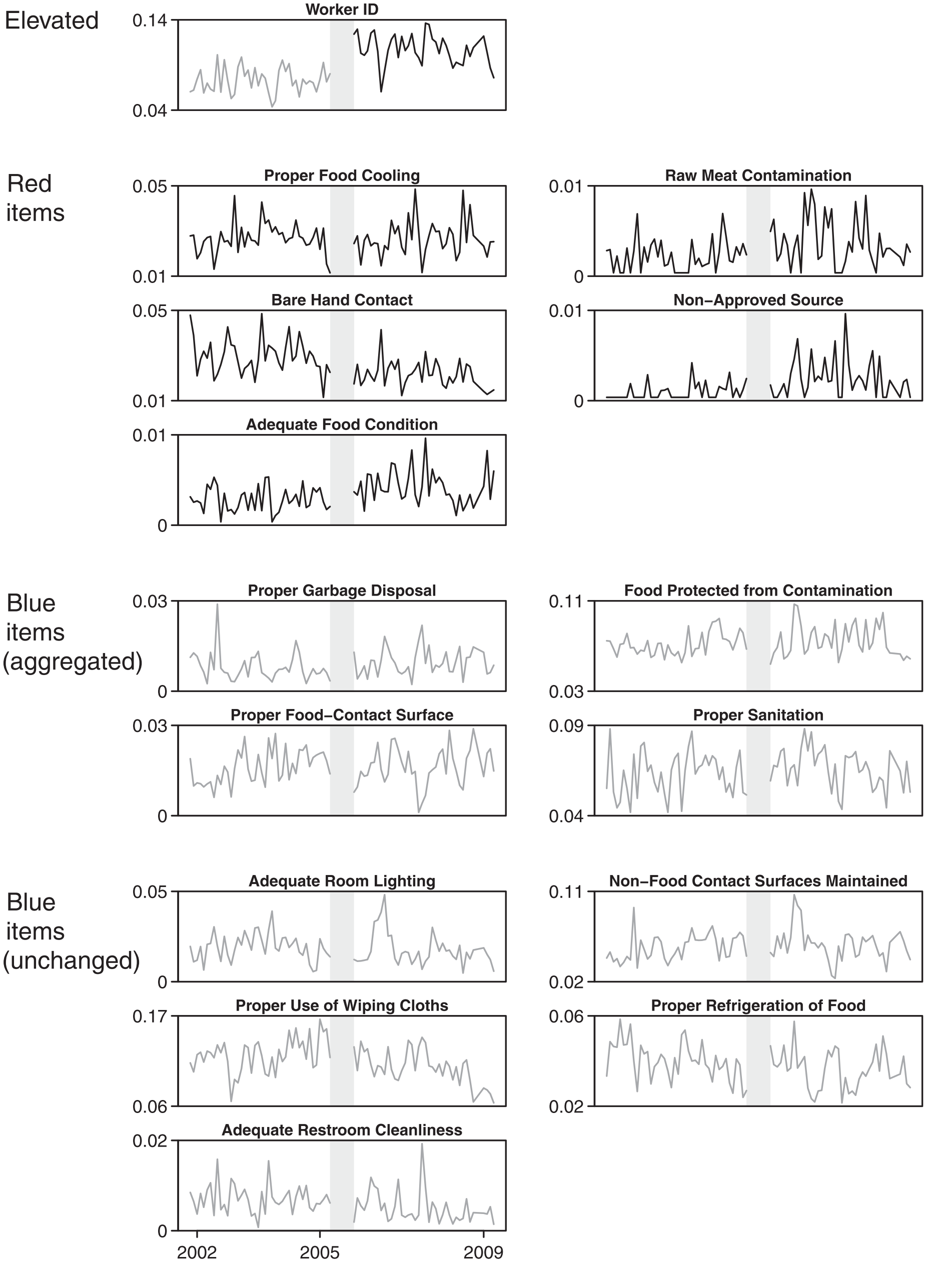
To ensure that there are no secular trends for specific code items, Figure 3 plots time series for each of the 15 equivalent code items. There are no obvious deviations from parallel pretreatment trends. The slope for the trend line in the pretreatment period for both red and blue items is effectively 0. This also appears to be the case for four violations that were aggregated in 2005.

To construct a formal test of the checklist effect, we estimate a fixed effects logistic model. The fully saturated model is specified as follows:

$$\log \frac{\mathrm{P}\left(y_{ijkl}=1\right)}{1-\mathrm{P}\left(y_{ijkl}=1\right)} = \tau_1\left(\mathrm{Post}_j \times \mathrm{Red}_i\right) + \tau_2\left(\mathrm{Post}_j \times \mathrm{Elevated}_i\right) + \tau_3\left(\mathrm{Post}_j \times \mathrm{Aggregated}_i\right) + \beta\,\mathrm{Post}_j + \delta_i + \vartheta_k + \alpha_l,$$

where y_{ijkl} is a binary indicator for whether code item i in inspection j of establishment k is scored by inspector l, Post$_j$ is an indicator for whether the code item occurred in the postintervention period, Red$_i$ is an indicator for a red (critical) code item, Elevated$_i$ is an indicator for whether the code item was elevated, and Aggregated$_i$ is an indicator for a code item that was aggregated in 2005. β accounts for item-invariant temporal differences before and after 2005. There are three sets of fixed effects: δ_i are code item fixed effects that account for differences in baseline citation rates for each of the 15 code items; ϑ_k are establishment fixed effects, which account for time-invariant differences in sanitation practices across establishments; and α_l are inspector fixed effects to account for differences in stringency across inspectors. Inspector and establishment fixed effects

Figure 3: Citation rates by code item, 2001–2009.



Notes: These time series depict the changes in the citation rate for each equivalently matched code item over time. Gray bars represent the training period for new code items from May–October 2005. Code items are displayed by item severity, and within severity, by aggregation level.

are separately identified even holding the area rotation fixed because inspectors may inspect establishments outside of their home areas due to vacation, illness, and other exigent circumstances. The treatment effects of interest are τ_1 and τ_2, the effects of the checklist and violation elevation, respectively, which are principally identified by temporal changes before and after 2005 in red citation rates relative to blue citation rates.

Table 5 presents results. For each labeled pair of columns, models are fit to the full observation window (2001–2009) or limiting the observation window to the same area assignments (2004–2006). Models in Column (A) present the simple model with no establishment and inspector fixed effects. Models in Columns (B), (C), and (D) sequentially add in establishment and inspector fixed effects, as we require an establishment to be inspected both before and after 2005. Results are comparable across models: elevating the worker training violation causes a (statistically significant) 2.1 percent increase in citation, but there is no statistically detectable effect of the checklist on whether red items are cited. Given the amount of data, the confidence interval on the checklist effect is relatively narrow: for the median code item, the interval is [–1.3%, 0.7%], allowing us to rule out substantial positive effects.[25]

In principle, it is possible that behavioral biases may make the citation of a single violation different from the citation of two violations, even if the predicate conditions are identical. Appendix C presents results for violation-specific models, again with comparable results.

The absence of detectable average effects, of course, does not preclude variance effects. An alternative theory of checklists may be that they should reduce interinspector differences in citations (with heterogeneous treatment effects potentially averaging to zero). Inspectors who undercite an item, for instance, may increase citations when the checklist reminds them of the item. Inspectors who overcite an item, on the other hand, might actually *decrease* citations with a checklist. One reason is that code item distinctions are by no means crystalline: residual sanitizing solution on a food contact surface, for instance, could arguably be scored either as a safe and unadulterated food violation or a raw meat cross-contamination violation (Ho 2017). The checklist might hence remind some inspectors that observed behavior previously scored one way falls more appropriately into a different violation category. If so, the checklist might produce no average citation effects, but decrease variance across inspectors.

Appendix D investigates whether there is any evidence of variance effects, using a hierarchical logistic regression model, with the variance of inspector random effects allowed to vary as a function of the checklist. We find no evidence of variance effects. While elevating a code item increases the citation rate, it may slightly reduce interinspector variability, but this result reaches only borderline significance.[26]

[25] This marginal effects calculation is based on the first model in Column (A), but coefficients and standard errors are comparable for all the models, with coefficients very close to 0.

[26] It is worth noting here that the reduction in interinspector variability is consistent even with increases in observed dispersion across inspectors in citation rates. The reason is that any average increase from a low baseline rate will necessarily increase the binomial variance. But our hierarchical models suggest that the binomial variance increases at a rate lower than from a pure mean shift.

Table 5: Logistic Regression of Code Item Citation

	(A)		*(B)*		*(C)*		*(D)*	
Post × Elevated	0.50**	0.53**	0.52**	0.55**	0.53**	0.58**	0.54**	0.57**
	(0.24)	(0.25)	(0.25)	(0.25)	(0.25)	(0.26)	(0.25)	(0.26)
Post × Red	−0.01	−0.02	−0.01	−0.01	−0.01	−0.02	−0.00	−0.01
	(0.07)	(0.10)	(0.08)	(0.10)	(0.08)	(0.10)	(0.08)	(0.10)
Post × Aggregated	0.13	0.03	0.13	0.03	0.13	0.03	0.14	0.03
	(0.09)	(0.10)	(0.09)	(0.10)	(0.35)	(0.22)	(0.09)	(0.10)
Post	−0.06	0.00	−0.08	−0.05	−0.07	0.00	−0.07	−0.07
	(0.07)	(0.09)	(0.06)	(0.09)	(0.07)	(0.08)	(0.06)	(0.09)
Estab. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inspector FE	No	No	No	No	Yes	Yes	Yes	Yes
Violation FE	No	No	Yes	Yes	No	No	Yes	Yes
N	1,426,305	544,050	1,426,305	544,050	1,283,835	411,420	1,283,835	411,420
Years	2001–2009	2004–2006	2001–2009	2004–2006	2001–2009	2004–2006	2001–2009	2004–2006

Notes: Each cell presents the point estimate with standard error in parentheses. For each model, the 2001–2009 period is presented in the left column, and the 2004–2006 period is presented on the right. Models (A) and (B) are fit using an unconditional logistic regression, but due to a large number of establishments in the data, Models (C) and (D) are fit with Chamberlain's (1980) conditional logistic regression. Standard errors are clustered at the inspector level. Data only include establishments inspected in both the pre- and postintervention periods. Sample sizes decrease when establishment fixed effects are added because establishments with no outcome variation are dropped. Coefficients reported on the logit scale. */**/*** denote statistical significance at α levels of 0.1, 0.05, and 0.01, respectively.

VI. Limitations

While our design provides us rich information about the effects of a checklist, we now discuss several potential limitations in interpreting the results.

First, red code items may already have been cited at the appropriate level, so that the null checklist effect simply reflects accurate preexisting inspection behavior. There are several reasons to doubt this account, however. The food code leaves considerable discretion to food inspectors, who are under time pressure to complete inspections (De Sitter & Van de Haar 1998:131), and a common lament is that inspectors fail to observe important food safety violations (e.g., Powell et al. 2013:688–89). The FDA itself intended for the checklists to remind inspectors of critical code items. The CFP debate also shows that the distinction between critical and noncritical code items is quite fluid—especially given the ambiguous distinctions between "risk factors," "public health interventions," and "good retail practices"—and the distinction certainly is not one turning on whether violations are "obvious" when observed (De Sitter & Van de Haar 1998:131). Moreover, there is considerable evidence that numerous inspectors cite violations less frequently than they occur (Ho 2017:73–74). In an internal study, the food program showed that average citation rates for many red items were substantially below the rates observed by FDA inspectors. For instance, while hand-washing violations were cited in 3 percent of inspections, FDA inspectors observed hand-washing violations in 76 percent of visits (Food and Drug Administration 2010:122). Given substantial differences in baseline citation rates across inspectors—which appear unaffected by the introduction of the checklist (see Appendix D)—it is doubtful that red code items were already uniformly at optimal levels.

Second, while our one positive result is that elevating a code item increases its citation rate, checklist and salience effects may interact. We distinguish here between the role of a checklist to remind inspectors of a code item and between increasing the ostensible importance of a violation by elevating a code item. On the one hand, we lack evidence on a code item that exclusively elevated (i.e., without being placed on a checklist basis). It is hence possible that the increase in citation of food worker card violations was the result of the joint effect of the checklist and increased salience. On the other hand, our evidence strongly suggests that the checklist alone has no appreciable effect.[27]

Last, our evidence may not speak to the effect of a checklist with more comprehensive reform. The median time on the job of King County food program employees was 15 years in 2017 (Ho 2017:47), so the effect of checklists with the 2005 code revision may have been small because inspectors were settled in old habits. Of course, one of the virtues touted by many checklist proponents is its simplicity and power to promote

[27] It is possible that the checklist might have an effect on low-salience violations, for which we also lack evidence. Two reasons to be skeptical of such an effect are that (1) the CFP debate shows that there is no clean distinction between critical and noncritical code items, and (2) the FDA's own rationale for adopting the checklist for critical items was that it was most important for those violations.

cultural change, even in the face of longstanding institutions. At minimum, our evidence challenges that account.[28]

VII. Implications

We conclude with several implications. First, checklists are no panacea. Structural problems in the administration of law cannot be papered over by a checklist. Promises of silver bullets may ring hollow for administrative agencies when underlying structural problems may stem from limited budgets, high turnover, constrained supervision and removal authority, and/or labyrinthine code provisions. Nor is this insight limited to law. As we spelled out above, checklist interventions are rarely about checklists alone. As proponents of checklists in healthcare put it: "When we begin to believe and act on the notion that safety is simple and inexpensive, that all it requires is a checklist, we abandon any serious attempt to achieve safer, higher quality care" (Bosk et al. 2009:445).

Second, our findings suggest that checklist benefits may emanate from dimensions other than the checklist itself. As mentioned earlier, such interventions often are accompanied by a concurrent increase in training, teamwork, evidence on failures, and reorientation on core tasks. Our study rejects the notion that checklists serve their main benefit as reminders alone. Focusing on other components, however, also underscores that quality improvement may be less easy and less simple than promised by the most fervent checklist proponents. If a cultural change toward teamwork is what accounts for earlier checklist findings (Bosk et al. 2009 [medicine]; Daughtrey & Carroll 2007 [software]; Gordon et al. 2012 [aviation]), more serious reform toward peer-based learning may be warranted in contexts where bureaucrats, lawyers, and judges typically go it alone.

Third, checklist design may nonetheless matter. Our affirmative finding is that elevating the salience of a violation increases its citation considerably.[29] It is well-known that compliance with checklists can be low (Pickering et al. 2013; Van Klei et al. 2012). In the nuclear safety context, for instance, Nichols and Wildavsky (1987) report that the complexity of checklists leads some code items to be simply forgotten. The dynamic is similar to that reported in aviation: "in some checklists, items that are not very critical are made so, consuming valuable time, adding workload and shifting attention from

[28] Similarly, one might distinguish between the treatment effect of checklists under full compliance, where there is better evidence, and the intention-to-treat effect of checklist programs that are weakened by noncompliance. A surgery simulation study, for instance, finds considerable effects (Arriaga et al. 2013), while efforts to adopt the checklists across South Carolina were plagued by substantial (nonrandom) noncompliance (Haynes et al. 2017). When considering a policy decision to implement checklists, noncompliance weakens the claim that checklists are simple and effective.

[29] We note that we do not have evidence of the effect of elevating a code item without a checklist in place. It is hence possible that the elevation effect is a combined effect of reclassifying the violation, moving it to the top of the inspection sheet, and subjecting it to a checklist. However, because (1) the violation remained identical and was widely known among inspection staff and (2) we find no evidence of checklist effects for nonelevated items, our best inference is that the effect stems from the increased salience rather than the checklist.

very critical checklist items to non-critical ones" (Degani & Wiener 1990:56). Our findings underscore the importance of determining which items are deemed critical and salient, which can vary considerably across jurisdictions.

Fourth, our findings demonstrate the inadequacy of the FDA's prioritization of violations. Consider the elevated code item for employee-wide food safety training. We researched the state of training requirements across the top 20 metropolitan areas (see Appendix F). Next to King County, only six require employee-wide training, with the rest requiring only manager training. In the six jurisdictions requiring employee-wide training, no jurisdiction classifies it as a critical item.[30] For the 13 jurisdictions requiring only manager training, five do not consider it a critical item, notwithstanding the FDA's classification as such. Even worse, the evidence base for employee-wide food worker training on sanitation practice remains quite mixed (Egan et al. 2007; York et al. 2009).[31] While elevating the violation increased its detection, the net effect may hence have been to worsen the allocation of inspection resources. This evident confusion about risk classification illustrates why the CFP saw a need to solidify the evidence base for code item prioritization. It also shows the futility of FDA's relabeling. The 2009 FDA Food Code heightened the confusion by relabeling violations again, this time to "priority," "priority foundation," and "core" items. And again, the CFP rightly found fault with the labels: "The new terms and levels of priority . . . are difficult for regulators to articulate and difficult for regulated industry to understand" (Conference for Food Protection 2012). But the problem is not just semantic; it is the lack of evidence for what to prioritize.

Fifth, our evidence corroborates the core insight by checklist proponents to manage complexity by simplifying core tasks (e.g., Gordon et al. 2012:129). In food safety, the FDA Model Food Code has ballooned from 21 pages in 1934 to 698 pages in 2009. Much of that growth, of course, reflects advancements in the scientific understanding of foodborne pathogens. But the sheer complexity of the volume means that line-level inspectors may rarely read the Model Food Code. And just as the code can become too complex, so can checklists (Bardach & Kagan 1982; Braithwaite & Braithwaite 1995; Ho 2012). The motivation for disaggregating checklists into many subitems often stems from a desire to reduce discretion to rule (Bardach & Kagan 1982; Ho, 2017). But our evidence suggests that aggregating violations has no detectable effect on citation rates, thereby supporting calls to simplify inspection score sheets, which can improve consistency (Braithwaite & Braithwaite 1995). As an exemplar of how inspections could be reinvented, Appendix E presents a score sheet we explored as part of a project for streamlined inspections of risk factors. In more than 400 inspections conducted as part of a different study, we found a version of this abbreviated checklist to work well both for interrater reliability and efficiency of the inspection process. Such simplification should be evidence based, with peer

[30] As Appendix F spells out, the specific labels vary across jurisdictions. We use the term "critical" here to refer to a jurisdiction's classification of the training/certification requirement in the highest class by severity.

[31] Evidence is stronger that training improves food safety knowledge as measured by surveys, but the evidence of the effect on behavior is much more limited.

review serving as a natural place to assess interrater reliability and optimal simplification of code items. In the medical context, Peter Pronovost describes the value of checklists as distilling the Centers for Disease Control complex 120-page guideline for preventing central line catheter infections into simple, actionable practice of five steps (Pronovost & Vohr 2010:25–27). Applied to law and regulation, the *Checklist Manifesto* may be more appropriately titled the *Check-Less Manifesto*.

Last, as a methodological matter, our study suggests a novel path forward for improving the evaluation of checklists. Stepped-wedge randomized designs typically roll out an intervention across areas, but such designs can be complex and infeasible. King County's application of a checklist to only half of code items suggests an alternative design feasible with only a single jurisdiction or institution: a (stepped-wedge) randomized rollout of the checklist across code items. Not only is such a design trivially easy when score sheets are digital in form, but such a design also has the virtue of sequencing potentially time-consuming training as the checklist is rolled out, enabling us to understand how training and checklists interact.

In sum, we find no evidence that checklists can improve regulatory enforcement behavior. Contrary to widespread calls for checklist-style reform, checklists alone are unlikely to address core challenges in the administration of law.

References

Aaronovitch, David (2010) "Review of *The Checklist Manifesto*: How to Get Things Right by Atul Gawande," January 23 *Times*.

Arriaga, Alexander F., Angela M. Bader, Judith M. Wong, Stuart R. Lipsitz, William R. Berry, John E. Ziewacz, David L. Hepner, Daniel J. Boorman, Charles N. Pozner, & Douglas S. Smink (2013) "Simulation-Based Trial of Surgical-Crisis Checklists," 368(3) *New England J. of Medicine* 246.

Bardach, Eugene, & Robert A. Kagan (1982) *Going by the Book: The Problem of Regulatory Unreasonableness*. Philadelphia, PA: Temple University Press.

Benjamini, Yoav, & Yosef Hochberg (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," 57 *J. of the Royal Statistical Society. Series B (Methodological)* 289.

Bergs, Jochen, Johan Hellings, Irina Cleemput, Ö Zurel, Vera De Troyer, Monique Van Hiel, J.-L. Demeere, Donald Claeys, & Dominique Vandijck (2014) "Systematic Review and Meta-Analysis of the Effect of the World Health Organization Surgical Safety Checklist on Postoperative Complications," 101(3) *British J. of Surgery* 150.

Bibas, Stephanos (2011) "Regulating the Plea-Bargaining Market: From Caveat Emptor to Consumer Protection," 99(4) *California Law Rev.* 1117.

Boorman, Daniel (2001) "Today's Electronic Checklists Reduce Likelihood of Crew Errors and Help Prevent Mishaps," 56(1) *ICAO J.* 17.

Bosk, Charles L., Mary Dixon-Woods, Christine A. Goeschel, & Peter J. Pronovost (2009) "Reality Check for Checklists," 374(9688) *Lancet* 444.

Braithwaite, John, & Valerie Braithwaite (1995) "The Politics of Legalism: Rules Versus Standards in Nursing-Home Regulations," 4 *Social & Legal Studies* 307.

Chamberlain, Gary (1980) "Analysis of Covariance with Qualitative Data," 47(1) *Rev. of Economic Studies* 225.

Conference for Food Protection (2002) *CFP Inspection Forms Committee (Risk Factors, Interventions and Inspection Form)*. Nashville, TN: Conference for Food Protection.

—— (2004) *2004 Issues Submitted*. Chandler, AZ: Conference for Food Protection.

—— (2012) "The 2009 FDA Food Code Introduced New Confusing Terms." *2012 Issue Form*. Indianapolis, IN: Conference for Food Protection.

Cutler, Stephen M. (2004) "Tone at the Top: Getting it Right," in *Second Annual General Counsel Roundtable*. Washington, DC. Available at: https://www.sec.gov/news/speech/spch120304smc.htm.

Daughtrey, Taz, & Sue Carroll (2007) *Fundamental Concepts for the Software Quality Engineer*, Vol. 2. Milwaukee, WI: ASQ Quality Press.

Davidoff, Frank (2010) "Checklists and Guidelines: Imaging Techniques for Visualizing What to Do," 304(2) *JAMA* 206.

Degani, Asaf, & Earl L. Wiener (1990) "Human Factors of Flight-Deck Checklists: The Normal Checklist." *NASA Contractor Report* 177549.

De Jager, Elzerie, Chloe McKenna, Lynne Bartlett, Ronny Gunnarsson, & Yik-Hong Ho (2016) "Postoperative Adverse Events Inconsistently Improved by the World Health Organization Surgical Safety Checklist: A Systematic Literature Review of 25 Studies," 40(8) *World J. of Surgery* 1842.

Department of Public Health Environmental Health, County of Los Angeles (2014) *Reference Guide for the Food Official Inspection Report*. Los Angeles, CA: Department of Public Health.

De Sitter, H., & S. Van de Haar (1998) "Governmental Food Inspection and HACCP," 9(2–3) *Food Control* 131.

Douglas, Joshua A. (2015) "A Checklist Manifesto for Election Day: How to Prevent Mistakes at the Polls," 43 *Florida State Univ. Law Rev.* 353.

Duclos, A., J. L. Peix, V. Piriou, P. Occelli, A. Denis, S. Bourdy, M. J. Carty, A. A. Gawande, F. Debouck, & C. Vacca (2016) "Cluster Randomized Trial to Evaluate the Impact of Team Training on Surgical Outcomes," 103(13) *British J. of Surgery* 1804.

Egan, M. B., M. M. Raats, S. M. Grubb, A. Eves, M. L. Lumbers, M. S. Dean, & M. R. Adams (2007) "A Review of Food Safety and Food Hygiene Training Studies in the Commercial Sector," 18(10) *Food Control* 1180.

Findley, Keith A. (2010) "New Perspectives on Brady and Other Disclosure Obligations: Report of the Working Groups on Best Practices," 31(6) *Cardozo Law Rev.* 1961.

Food and Drug Administration (2010) *FDA Trend Analysis Report on the Occurrence of Foodborne Illness Risk Factors in Selected Institutional Foodservice, Restaurant, and Retail Food Store Facility Types (1998–2008)*. Washington, DC: FDA National Retail Food Team.

Ford, Cristie L. (2005) "Toward a New Model for Securities Law Enforcement," 57(3) *Administrative Law Rev.* 757.

—— (2008) "New Governance, Compliance, and Principles-Based Securities Regulation," 45(1) *American Business Law J.* 1.

Gawande, Atul (2007) "The Checklist," 83(39) *New Yorker* 86.

—— (2009) *The Checklist Manifesto: How to Get Things Right*. New York: Henry Holt and Company.

Gillespie, Brigid M., Wendy Chaboyer, Lukman Thalib, Melinda John, Nicole Fairweather, & Kellee Slater (2014) "Effect of Using a Safety Checklist on Patient Complications After Surgery—Systematic Review and Meta-Analysis," 120(6) *J. of the American Society of Anesthesiologists* 1380.

Gordon, Suzanne, Patrick Mendenhall, & Bonnie Blair O'Connor (2012) *Beyond the Checklist: What Else Health Care Can Learn from Aviation Teamwork and Safety*. Ithaca, NY: Cornell Univ. Press.

Griffin, Lissa (2011) "Pretrial Procedures for Innocent People: Reforming Brady," 56 *New York Law School Law Rev.* 969.

Hale, Andrew, & David Borys (2013) "Working to Rule or Working Safely? Part 2: The Management of Safety Rules and Procedures," 55 *Safety Science* 222.

Hales, Brigette, & Peter J. Pronovost (2006) "The Checklist—A Tool for Error Management and Performance Improvement," 21(3) *J. of Critical Care* 231.

Hales, Brigette, Marius Terblanche, Robert Fowler, & William Sibbald (2008) "Development of Medical Checklists for Improved Quality of Patient Care," 20(1) *International J. for Quality in Health Care* 22.

Haynes, Alex B., Lizabeth Edmondson, Stuart R. Lipsitz, George Molina, Bridget A. Neville, Sara J. Singer, Aunyika T. Moonan, Ashley Kay Childers, Richard Foster, & Lorri R. Gibbons (2017) "Mortality Trends After a Voluntary Checklist-Based Surgical Safety Collaborative," 266(6) *Annals of Surgery* 923.

Haynes, Alex B., Thomas G. Weiser, William R. Berry, Stuart R. Lipsitz, Abdel-Hadi S. Breizat, E. Patchen Dellinger, Teodoro Herbosa et al. (2009) "A Surgical Safety Checklist to Reduce Morbidity and Mortality in a Global Population," 360(5) *New England J. of Medicine* 491.

Hersch, Matthew H. (2009) "Checklist: The Secret Life of Apollo's 'Fourth Crewmember'," 57 *Sociological Rev.* 6.

Ho, Daniel E. (2012) "Fudging the Nudge: Information Disclosure and Restaurant Grading," 122(3) *Yale Law J.* 574.

—— (2017) "Does Peer Review Work: An Experiment of Experimentalism," 69 *Stanford Law Rev.* 1.

Ho, Daniel E., & Sam Sherman (2017) "Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement," 13(1) *Annual Rev. of Law & Social Science* 251.

Institute of Education Sciences & National Center for Education Statistics, Department of Education (2014) "Agency Information Collection Activities; Submission to the Office of Management and Budget for Review and Approval; Comment Request; Impacts of a Detailed Checklist on Formative Feedback to Teachers," *Federal Register.*

Jammer, I., T. Ahmad, C. Aldecoa, D. Koulenti, T. Goranović, I. Grigoras, B. Mazul-Sunko, R. Matos, R. Moreno, & G. H. Sigurdsson (2015) "Point Prevalence of Surgical Checklist Use in Europe: Relationship with Hospital Mortality," 114(5) *British J. of Anaesthesia* 801.

Jones, Timothy F., Boris I. Pavlin, Bonnie J. LaFleur, L. Amanda Ingram, & William Schaffner (2004) "Restaurant Inspection Scores and Foodborne Disease," 10(4) *Emerging Infectious Diseases* 688.

Koski, William S. (2009) "The Evolving Role of the Courts in School Reform Twenty Years After Rose," 98 *Kentucky Law J.* 789.

Langevin, James R. (2007) *Beyond the Checklist: Addressing Shortfalls in National Pandemic Influenza Preparedness.* Washington, DC: Government Printing Office.

Lau, Christine S. M., & Ronald S. Chamberlain (2016) "The World Health Organization Surgical Safety Checklist Improves Post-Operative Outcomes: A Meta-Analysis and Systematic Review," 7(04) *Surgical Science* 206.

Levine, Kate (2016) "Police Suspects," 116(5) *Columbia Law Rev.* 1197.

Lyons, Vanessa E., & Lori L. Popejoy (2014) "Meta-Analysis of Surgical Safety Checklist Effects on Teamwork, Communication, Morbidity, Mortality, and Safety," 36(2) *Western J. of Nursing Research* 245.

Maggs, Gregory E. (1992) "Reducing the Costs of Statutory Ambiguity: Alternative Approaches and the Federal Courts Study Committee," 29 *Harvard J. on Legislation* 123.

Mashaw, Jerry L. (1985) *Bureaucratic Justice: Managing Social Security Disability Claims.* Yale Univ. Press.

Mayer, Erik K., Nick Sevdalis, Shantanu Rout, Jochem Caris, Stephanie Russ, Jenny Mansell, Rachel Davies, Petros Skapinakis, Charles Vincent, & Thanos Athanasiou (2016) "Surgical Checklist Implementation Project: The Impact of Variable WHO Checklist Compliance on Risk-Adjusted Clinical Outcomes After National Implementation: A Longitudinal Study," 263(1) *Annals of Surgery* 58.

McAllister, Lesley K. (2012) "Regulation by Third-Party Verification," 53(1) *Boston College Law Rev.* 1.

Nichols, Elizabeth, & Aaron Wildavsky (1987) "Nuclear Power Regulation: Seeking Safety, Doing Harm," 11 *Regulation* 45.

Noonan, Kathleen G., Charles F. Sabel, & William H. Simon (2009) "Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform," 34(3) *Law & Social Inquiry* 523.

Pickering, S. P., E. R. Robertson, D. Griffin, M. Hadi, L. J. Morgan, K. C. Catchpole, S. New, G. Collins, & P. McCulloch (2013) "Compliance and Use of the World Health Organization Checklist in UK Operating Theatres," 100(12) *British J. of Surgery* 1664.

Powell, Douglas A., S. Erdozain, Charles Dodd, R. Costa, K. Morley, & Benjamin J. Chapman (2013) "Audits and Inspections Are Never Enough: A Critique to Enhance Food Safety," 30(2) *Food Control* 686.

Pronovost, Peter, & Eric Vohr (2010) *Safe Patients, Smart Hospitals: How One Doctor's Checklist Can Help Us Change Health Care from the Inside Out*. New York: Penguin.

Reason, James (1997) *Managing the Risks of Organizational Accidents*. Burlington, VT: Ashgate Publishing.

Ross, Catherine J. (2003) "The Tyranny of Time: Vulnerable Children, 'Bad' Mothers, and Statutory Deadlines in Parental Termination Proceedings," 11 *Virginia J. of Social Policy & the Law* 176.

Simon, William H. (2012) "Where Is the Quality Movement in Law Practice," 2012 *Wisconsin Law Rev.* 387.

—— (2015) "The Organizational Premises of Administrative Law," 78 *Law & Contemporary Problems* 61.

Stuart, Diana, & Michelle R. Worosz (2012) "Risk, Anti-Reflexivity, and Ethical Neutralization in Industrial Food Processing," 29(3) *Agriculture & Human Values* 287.

Thaler, Richard H., Cass R. Sunstein, & John P. Balz (2012) "Choice Architecture," in E. Shafir, ed., *The Behavioral Foundations of Public Policy*, pp. 428–39. Princeton, NJ: Princeton Univ. Press.

U.S. Food and Drug Administration (2001) *FDA Food Code 2001*. Washington, DC: U.S. FDA.

—— (2005) *FDA Food Code 2005*. Washington, DC: U.S. FDA.

Van Klei, W. A., R. G. Hoff, E. E. H. L. Van Aarnhem, R. K. J. Simmermacher, L. P. E. Regli, T. H. Kappen, L. Van Wolfswinkel, C. J. Kalkman, W. F. Buhre, & L. M. Peelen (2012) "Effects of the Introduction of the WHO 'Surgical Safety Checklist' on In-Hospital Mortality: A Cohort Study," 255(1) *Annals of Surgery* 44.

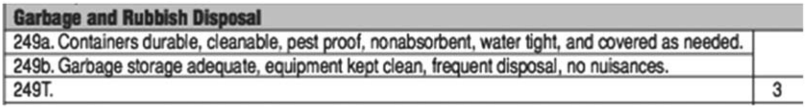
York, Valerie K., Laura A. Brannon, Carol W. Shanklin, Kevin R. Roberts, Amber D. Howells, & Elizabeth B. Barrett (2009) "Foodservice Employees Benefit from Interventions Targeting Barriers to Food Safety," 109(9) *J. of the American Dietetic Association* 1576.

Appendix A: Data Cleaning

We describe our process for cleaning the database of King County inspections. For violations, the raw data correspond to a code item cited as a violation during an inspection. Each violation carries a number of "violation points" depending on its severity. Pre-2005, as Figure 4 shows, the inspection score sheet listed the point score below applicable violations with multiple subviolations.

In instances like those shown in Figure 4, the violations data assign a 0-point violation to the substantive lines, but a 3-point violation to a line with an empty description. The scope of this problem was small but nontrivial, occurring in 8,097 of 188,365 total inspections in the raw data. Of 5,882,271 violations, 8,987 rows (0.15 percent) presented this challenge. Because the problem uniquely affected the precode change data, inferences could be biased without accounting for this problem.

Figure 4: Example of a quirk in the pre-2005 inspection form.



Garbage and Rubbish Disposal	
249a. Containers durable, cleanable, pest proof, nonabsorbent, water tight, and covered as needed.	
249b. Garbage storage adequate, equipment kept clean, frequent disposal, no nuisances.	
249T.	3

Notes: Inspectors cited two distinct code items using the same row and assigning the same point total.

Table 6: Distribution of Empty and 0–Point Violations as a Percentage of 8,097 Inspections Where the Problem Exists

	0-Point Violations								
Empty Violation	**0**	**1**	**2**	**3**	**4**	**5**	**7**	**8**	**9**
1	0.36	87.65	2.38	0.05	0.00	0.00	0.00	0.00	0.00
2	0.01	0.05	7.90	0.54	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.64	0.12	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.17	0.02	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00

To clean these data, we classify each affected inspection by the number of code items affected. As Table 6 demonstrates, the vast majority of instances are easily resolved, as there is a one-to-one correspondence between a violation with an empty description and a 0-point violation. The inspections with the same number of empty violations and 0-point violations (N-N cases) constituted roughly 97 percent of affected inspections. Roughly 3 percent of inspections had unequal numbers of empty violations and 0-point violations (N-M cases). In a very small number of inspections (0.37 percent), empty violations were accompanied with no 0-point violation. In these latter cases, we apply no correction, as it is likely due to input error.

We used the expected violation points from the official inspection form to match 0-point violations to empty violations based on point totals. (Violation points did not appear to change during the relevant time period.) When code items were manually written for a 0-point violation, we first matched the inspector's written description to an official violation on the inspection form, and assigned expected point totals to 0-point violations based on this match. For example, we would match descriptions such as "Containers [garbage] durable, cleanable, pest proof, nonabsorbent, water tight, and covered as needed" and "Garbage, refuse properly disposed; facilities maintained" to the official 2004 Violation 249a, "Containers durable, cleanable, pest proof, nonabsorbent, water tight, and covered as needed," shown in Figure 4. In N-M cases, when multiple identically scored violations existed, we matched empty violations to the directly preceding violation or consecutive grouping of subviolations. For an example of the latter, if 249a and 249b occurred consecutively before an empty violation, we matched the empty violation to both 249a and 249b. We did so on the theory that in these cases, inspectors intended to cite a grouping of subviolations using one empty violation.

For all matched 0-point violations, we imputed the expected point totals based on the official inspection form. After matching violations, 0-point violations in all N-N cases and the great majority of N-M were uniquely matched, allowing us to correct over 8,000 0-point violations. For the small remaining set of cases with no matches, one explanation is that an inspector wanted to draw the establishment's attention to a violation, without an actual citation (hence the 0-point violation).

As a sensitivity analysis, we matched and scored all 0-point violations in the dataset. Due to aggregation of violations postchecklist and the dropping of all nonequivalent code items from our analysis, the more liberal imputation rule under the sensitivity analysis changed a total of seven violations from un-scored to scored. Thus, rerunning the regression models from Table 5 using the full period data with this imputation rule yielded nearly identical results.

Appendix B: Classification of Equivalent Code Items

This appendix provides additional details on the classification of equivalent code items. For each code item, the following table provides a comparison of code and score sheet language before and after the intervention in 2005. For ease of reference, we cite to the King County Food Code and Washington State Retail Food Code, rather than the codified provisions. The similarity of language across years illustrates the reason we classified these code items as an equivalent match.

Type	*Aggregation*	*Short Name*	*2000 Food Code Language/2004 Inspection Form Language*	*2005 Food Code Language/2006 Inspection Form Language*
Elevated	Same	Worker ID	Requires that an establishment "comply with the provisions of Chapter 69.06 RCW and Chapter 246-217 WAC" (5.18.060A1); food worker cards are displayed or filed where they are available for inspection "upon request" (5.18.060B). *235—"Food and beverage service workers permits current for all food workers"*	Requires that an establishment "comply with the provisions of Chapter 69.06 RCW and Chapter 246-217 WAC" (2–103.12A); food worker cards are displayed or filed where they are available for inspection "upon request" (2–103.12B). *0200—"Food Worker Cards current for all food workers; new food workers trained"*
Red	Same	Proper Food Cooling	"When potentially hazardous foods require cooling or cold holding after preparation, rapid methods of cooling from one-hundred forty degrees Fahrenheit (140° F) to forty-five degrees Fahrenheit (45° F) shall be used" (5.16.070). *109—"Potentially hazardous food cooled from 140F to 45F or below by approved methods (size reduction; uncovered shallow pans or ice bath with stirring)"*	"Cooked potentially hazardous food shall be cooled within 6 hours from 140° F to 41°F or less, or to 45°F or less" "if existing refrigeration equipment is not capable of maintaining the food at 41°F or less" (3–501.14A(2) & 3–501.16(A)(2)(b)). [Temperature change was phased in over five years.] *1600—"Proper cooling procedures"*

Appendix B *Continued*

Type	*Aggregation*	*Short Name*	*2000 Food Code Language/2004 Inspection Form Language*	*2005 Food Code Language/2006 Inspection Form Language*
		Raw Meat Contamination	"All utensils and food contact surfaces of equipment used in preparation, service, display, or storage of potentially hazardous food ["Includes any food of animal origin" 5.04.550] shall be sanitized . . . Following any interruption of operations during which contamination of food contact surfaces may have occurred and whenever contamination has occurred" (5.12.020D; 5.22.010B-C). *103a—"Food contact surfaces of equipment used with raw meats; aquatic foods; or poultry thoroughly cleaned and sanitized before contacting foods."*	"Equipment food-contact surfaces shall be cleaned: Except as specified in (B) of this section, before each use with a different type of raw animal food . . . Each time there is a change from working with raw foods to working with ready to eat foods; Between uses with raw fruits and vegetables and with potentially hazardous foods" (4–602.11A). *1300—"Food contact surfaces used for raw meat thoroughly cleaned and sanitized"*
		Bare Hand Contact	Employees should minimize "hand contact with foods by using appropriate utensils; providing tongs, bakery papers, scoops, spatulas, ladles, and similar utensils for handling foods during display or service; and/or using single service food service gloves when appropriate" (5.08.010G). *107a—"Food workers use proper utensils (tongs; scoops; disposable gloves) to minimize direct hand contact with RTE foods"*	"[E]xcept when washing fruits and vegetables as specified under 3–302.15 or when otherwise approved, food employees may not contact ready-to-eat food with their bare hands and shall use suitable utensils such as deli tissue, spatulas, tongs, single-use gloves, or dispensing equipment" (3–301.11). *0500—"Proper methods used to prevent bare hand contact with RTE foods"*
		Nonapproved Source	Water shall be "[a]dequate in quantity and quality" (5.26.010A1). Water shall be "supplied directly from a source approved under WAC 246–290 [which defines and regulates the operation and maintenance of public and nonpublic water systems])" (5.26.010A2).	The quality of water "shall meet drinking water standards" (5–102.11). Water shall be supplied from "an approved source that is a public water system or a nonpublic water system that is constructed, maintained, and operated according to law" (5- 1-1.11A-B).

Appendix B *Continued*

Type	*Aggregation*	*Short Name*	*2000 Food Code Language/2004 Inspection Form Language*	*2005 Food Code Language/2006 Inspection Form Language*
			Bottled water from an approved source (5.26.010B). Ice is to be “made from an approved water source; and manufactured, stored, transported, and handled in a sanitary manner” (5.26.101C). *101c—“Non-regulated foods (including water and ice) from sources approved by H.O.”*	Bottled water from an approved source (5- 101.13). “Ice used for food or a cooling medium shall be made from drinking water” (3–202.16). *0800—“Water supply, ice from approved source”*
		Adequate Food Condition	Food shall be “[c]lean, wholesome, and free from spoilage and adulteration; Protected from becoming adulterated; Safe for human consumption” (5.06.010C-E). *102a—“Foods wholesome; free from spoilage; not adulterated”*	“[F]ood shall be safe, unadulterated, and, as specified under 3–601.12, honestly presented” (3–101.11). *1000—“Food in good condition, safe and unadulterated; approved additives”*
Blue	Same	Adequate Room Lighting	“Lighting of at least thirty (30) foot candles” in areas where food is prepared/stored; utensils are washed; hands are washed; in bathrooms; and when cleaning is occurring, and “proper shields or guards for lights in the food preparation areas and areas where unwrapped food is stored and displayed” (5.32.030A-B); “ensure design, installation, and maintenance of ventilation systems” (5.32.040A). *253—“Lighting and ventilation provided as required. Ventilation hoods, —, filters cleaned and maintained properly”*	Light intensity shall be: at least 110 lux in walk-in refrigeration units and dry food storage areas (06–303.11A); at least 220 lux on consumer self-service surfaces, inside equipment, or in area used for handwashing/warewashing (6–303.11B); at least 540 lux where a food employee is working with food or potentially dangerous equipment (6–303.11C). “If necessary to keep rooms free of excessive heat, steam, condensation, vapors, obnoxious odors, smoke, and fumes, mechanical ventilation of sufficient capacity shall be provided” (6–304.11). *4900—“Adequate ventilation, lighting; designated areas used”*
		Non-Food Contact Surfaces Maintained	“[N]on-food contact surfaces of equipment are cleaned at such intervals to keep them clean and in a sanitary condition” (5.22.060C). *241—“Non-food contact surfaces maintained and clean”*	“Nonfood-contact surfaces of equipment shall be cleaned at a frequency necessary to preclude accumulation of soil residues” (4–602.13). *4300—“Non-food-contact surfaces maintained and clean”*

Appendix B *Continued*

Type	*Aggregation*	*Short Name*	*2000 Food Code Language/2004 Inspection Form Language*	*2005 Food Code Language/2006 Inspection Form Language*
		Proper Refrigeration of Food	“Equip each refrigeration unit with a numerically scaled thermometer accurate to within three degrees Fahrenheit” (5.16.010D). *229—“Accurate thermometers present in refrigerators”*	“Food temperature measuring devices shall be provided and readily accessible for use in ensuring attainment and maintenance of food temperatures” (4–302.12A). *2900—“Adequate equipment for temperature control”*
		Toilets	Toilets within 200-feet of building for employees (5.26.050), toilets for patrons if there is seating at the establishment (5.26.060, 5- 203.12). *248—“Toilet facilities for employees/patrons available, adequate, convenient, clean and in good repair”*	Toilets within 200-feet of building for employees (5–203.12), toilets for patrons if there is seating at the establishment (5.26.060, 5–203.12). *4600—“Toilet facilities properly constructed, supplied, cleaned”*
		Proper Use of Wiping Cloths	Wiping cloths should be “kept in a clean, sanitary condition at all times, moistened with an approved sanitizing solution at all times when in use; and stored in a proper concentration of sanitizing solution between uses” (5.22.040). 1995 marking instructions make clear that “dirty wiping cloths” is a violation (#53). *239—“Wiping cloths clean, moistened with an approved sanitizer, restricted in use”*	Wiping cloths “used with raw animal foods shall be kept separate from cloths used for other purposes,” that cloths shall be “wet and cleaned . . . stored in a chemical sanitizer at a concentration specified in 4–501.11, and used for wiping spills” (3–404.11). “[D]ry wiping cloths shall be free of food debris and visible soil” (3–304.11). *3400—“Wiping cloths properly used, stored”*
	Aggregated	Food Protected from Contamination	Food must be covered (5.08.010A); stored off the floor (5.08.010C); in sealed containers (5.08.010A); out of reach of water, (05.08.010B); not in toilets or garbage rooms (05.08.010D). *224—“Food protected from actual and potential contamination; properly covered (except during cooling); no double stacking; sneeze guards provided.”* *226—“Foods stored off the floor”*	Food must be covered (3–305.11); stored off the floor (3–305.11); in sealed containers (3–305.11); out of reach of water, (3–305.11); not in toilets or garbage rooms (3–305.12). *3300—“Potential food contamination prevented during preparation, storage, display”*

Appendix B *Continued*

Type	*Aggregation*	*Short Name*	*2000 Food Code Language/2004 Inspection Form Language*	*2005 Food Code Language/2006 Inspection Form Language*
		Proper Garbage Disposal	Garbage storage “durable, easily cleanable, insect and rodent proof, nonabsorbent, in sound condition, watertight, kept covered” (5.28.010). Establishments should ensure “frequent disposal” (5.28.050) and clean the receptacles frequently (5.28.050). *249a—“Containers [garbage] durable, cleanable, pest proof, nonabsorbent, water tight, and covered as needed”* *249b—“Garbage storage adequate, equipment kept clean, frequent disposal, no —“ [We could not make out the last few words on the form]*	Garbage storage “durable, cleanable, inset- and rodent-resistant, leakproof, and nonabsorbent” (5–501.13). Receptacles “shall be kept covered” (5–501.113). Establishments should ensure “frequent disposal” (5–502.11) and clean the receptacles frequently (5–501.116). *4700—“Garbage, refuse properly disposed; facilities maintained*
		Proper Sanitation	Manual dishwashing to take place in water hotter than 170 degrees (5.22.010) or a mechanical dishwasher that automatically dispenses sanitizer (5.22.010B3), has a temperature measuring device/thermometer (5.22.010B2). *238—“Sanitizing dish solutions at proper temperature for penetration. Correct dishwashing procedures”* *242a—“Utensils prewashed as needed. Wash water clean and proper temperature”* *242b—“Accurate thermometers, chemical test kits and pressure gauges present and functional to monitor dishwashing sanitizing”*	Manual dishwashing to take place in water hotter than 170 degrees (4–204.116) or a mechanical dishwasher that automatically dispenses sanitizer (4 −204.117), has a temperature measuring device/thermometer (4-204.115). Warewashing machines must have a water pressure gauge (4–204.118). *4100—“Warewashing facilities properly installed, maintained, used; test strips available and used”*
		Proper Food-Contact Surface	Food-contact surfaces shall be “made of food-grade material,” nontoxic,” “corrosion resistant,” and “nonabsorbent” (5.18.020), “smooth” (5.20.020), “durable” (5.20.101), and “easily cleanable” (5.20.010).	Food-contact surfaces shall be “made of food-grade material,” nontoxic,” “corrosion resistant,” and “nonabsorbent” (4–101.11-4- 101.19), “smooth” (4–202.11), “durable” (4- 101.11), and “easily cleanable” (4–101.11).

Appendix B *Continued*

Type	*Aggregation*	*Short Name*	*2000 Food Code Language/2004 Inspection Form Language*	*2005 Food Code Language/2006 Inspection Form Language*
			Non-food contact surfaces "smooth, easily cleanable, durable, in good repair" (5.20.010) and installed properly (5.20.040). *237a—"Food contact surfaces smooth, easily cleanable, properly constructed, and non-toxic"* *237b—"Non-food contact surfaces properly constructed and installed"* *237c—"Plastic used as food contact surface food grade"*	Non-food contact surfaces shall be "corrosion- resistant, non-absorbent, and smooth" (4-101.111) and "constructed to allow easy cleaning" (4–202.16). *4000—"Food and non-food surfaces properly used and constructed; cleanable"*

Appendix C: Individual Violation Models

In this appendix, we report results from fixed effects logistic regressions fit separately to each code item. We model the data as follows:

$$\log \frac{\mathrm{P}\left(y_{jkl}=1\right)}{1-\mathrm{P}\left(y_{jkl}=1\right)}=\beta_0+\beta_1\cdot \mathrm{Post}_j+\alpha_l,$$

where y_{jkl} is a binary indicator for whether the code item in inspection j of establishment k is scored by inspector l, Post_j is an indicator for whether the violation occurred in the postchecklist period, and α_l are inspector fixed effects.

Table 7 fits the model to control (blue) code items. The results suggest no pre-post time differences in citation rates, which implies that the pre-post difference for each treated (red or elevated) code item will not be substantially different from a difference-in-differences estimate. We hence fit model the each treated code item.

Table 7: Fixed Effects Logistic Regression on Equivalent Blue Code Items, 2001–2009

	Blue								
Parameter	*Food Protected from Contam.*	*Proper Garbage Disposal*	*Adequate Room Lighting*	*Non-Food Contact Surfaces Maintain.*	*Proper Sanitation*	*Proper Food-Contact Surface*	*Toilets*	*Proper Use of Wiping Cloths*	*Proper Refrig. of Food*
Post	−0.05	0.05	−0.07	0.12	0.09	0.12	−0.29	−0.21	−0.04
	(0.10)	(0.19)	(0.15)	(0.11)	(0.10)	(0.12)	(0.28)	(0.10)	(0.21)
N	95,087	95,087	95,087	95,087	95,087	95,087	95,087	95,087	95,087

Notes: Inspector fixed effects included. "Post" represents the coefficient of interest, that is, the pre-post difference on the code item citation rate. Coefficients are presented with standard errors in parentheses, clustered at the inspector level. *P* values from models corresponding to blue code items are corrected for multiple testing using the Benjamini and Hochberg (1995) procedure. */**/*** denote that the coefficients are significant under the Benjamini and Hochberg (1995) procedure at false discovery rates of 0.1, 0.05, and 0.01, respectively.

Table 8: Fixed Effects Logistic Regression on Equivalent Red or Elevated Code Items, 2001–2009

	Elevated	*Red*				
Parameter	*Worker ID*	*Raw Meat Contam.*	*Proper Food Cooling*	*Bare Hand Contact*	*Nonapproved Source*	*Adequate Food Condition*
Post	0.49**	0.75	0.02	−0.36***	1.01	0.39
	(0.24)	(0.59)	(0.09)	(0.08)	(0.78)	(0.20)
N	95,087	95,087	95,087	95,087	95,087	95,087

NOTES: Inspector fixed effects included. "Post" represents the coefficient of interest, that is, the pre-post difference on the code item citation rate. Coefficients are presented with standard errors presented in parentheses, clustered at the inspector level. *P* values from models corresponding to red code items are corrected for multiple testing using the Benjamini and Hochberg (1995) procedure. */**/*** denote that the coefficients are significant under the Benjamini and Hochberg (1995) procedure at false discovery rates of 0.1, 0.05, and 0.01, respectively.

Results are comparable to the pooled models (see Table 8). We find a statistically significant increase in the citation rate for the elevated code items, but none for other code items, except for a bare hand contact violation. Violations for that code item, if anything, appear to decrease after the introduction of the checklist.

Appendix D: Mean and Variance Effect Models

In this appendix, we report results from a hierarchical logistic regression that allows the variance of inspector random effects to vary as a function of the checklist. We model the data as follows:

$$\log \frac{\mathrm{P}\left(y_{jkl}=1\right)}{1-\mathrm{P}\left(y_{jkl}=1\right)} = \beta_1 \cdot \mathrm{Pre}_j + \beta_2 \cdot \mathrm{Post}_j$$

$$\beta_1 = \gamma_{10} + u_{1m}$$

$$\beta_2 = \gamma_{20} + u_{2m}$$

$$u_{1m} \sim \mathrm{N}\left(0, \tau_1^2\right)$$

$$u_{2m} \sim \mathrm{N}\left(0, \tau_2^2\right),$$

where y_{jkl} is a binary indicator for whether the code item in inspection j of establishment k is scored by inspector l, Pre_j is an indicator for whether the violation occurred in the prechecklist period, and Post_j is an indicator for whether the code item occurred in the postchecklist period. The γ parameters allow for a mean shift in citation rates after the intervention and the τ parameters allow for the variance of inspector random effects to be different before and after the intervention.

Table 9 confirms that there are no mean or variance effects for control (blue) code items. This again justifies modeling the simple pre-post difference in the treated (elevated and red) code items.

Table 9: Hierarchical Logistic Regression of Blue Code Items Allowing the Variance of Inspector Random Effects to Vary Before and After the Intervention

	Blue								
Parameter	*Food Protected from Contam.*	*Proper Garbage Disposal*	*Adequate Room Lighting*	*Non-Food Contact Surfaces Maintain.*	*Proper Sanitation*	*Proper Food-Contact Surface*	*Toilets*	*Proper Use of Wiping Cloths*	*Proper Refrig. of Food*
Pre	−2.77	−5.54	−5.11	−3.77	−3.06	−4.75	−5.82	−2.96	−4.41
	(0.12)	(0.26)	(0.31)	(0.22)	(0.21)	(0.20)	(0.22)	(0.32)	(0.33)
Post	−2.92	−5.24	−5.44	−3.46	−3.01	−5.05	−6.62	−2.92	−4.57
	(0.15)	(0.21)	(0.29)	(0.23)	(0.22)	(0.27)	(0.27)	(0.28)	(0.31)
Mean-Diff	−0.15	0.30	−0.33	0.31	0.05	−0.30	−0.80	0.04	−0.16
Var(Pre)	0.49	1.85	3.32	1.73	1.33	1.22	1.11	3.55	3.40
	(0.12)	(0.53)	(0.86)	(0.36)	(0.50)	(0.31)	(0.37)	(0.94)	(1.10)
Var(Post)	0.76	1.26	3.40	1.71	1.39	2.13	1.97	2.55	3.33
	(0.16)	(0.28)	(0.74)	(0.57)	(0.48)	(0.54)	(0.69)	(0.71)	(0.91)
Var-Diff	0.27	−0.59	0.08	−0.02	0.06	0.91	0.86	−1.00	−0.07
N	95,087	95,087	95,087	95,087	95,087	95,087	95,087	95,087	95,087

NOTES: Coefficients are presented with standard errors in parentheses, clustered by inspectors. Data only include establishments inspected in both the pre- and postintervention periods. *P* values for both mean and variance effects from models corresponding to blue code items are corrected for multiple testing using the Benjamini and Hochberg (1995) procedure.*/**/*** denote that the coefficients are significant under the Benjamini and Hochberg (1995) procedure at false discovery rates of 0.1, 0.05, and 0.01, respectively.

Table 10: Hierarchical Logistic Regression of Elevated and Red Code Items Allowing the Variance of Inspector Random Effects to Vary Before and After the Intervention

	Elevated	*Red*				
Parameter	*Worker ID*	*Raw Meat Contam.*	*Proper Food Cooling*	*Bare Hand Contact*	*Nonapproved Source*	*Adequate Food Condition*
Pre	−3.31	−7.50	−4.01	−3.70	−9.35	−6.22
	(0.27)	(0.38)	(0.24)	(0.16)	(0.82)	(0.22)
Post	−2.32	−7.11	−3.73	−3.96	−9.16	−5.73
	(0.19)	(0.34)	(0.15)	(0.13)	(0.68)	(0.24)
Mean-Diff	0.99***	0.39	0.28	−0.26	0.19	0.49
Var(Pre)	2.44	2.09	1.56	0.69	3.89	1.00
	(0.83)	(0.84)	(0.60)	(0.22)	(2.32)	(0.31)
Var(Post)	1.14	3.37	0.64	0.40	5.16	0.97
	(0.37)	(1.19)	(0.22)	(0.15)	(2.76)	(0.38)
Var-Diff	−1.30	1.28	−0.92	−0.29	1.27	−0.03
N	95,087	95,087	95,087	95,087	95,087	95,087

NOTES: Coefficients are presented with standard errors in parentheses, clustered by inspectors. Data only include establishments inspected in both the pre- and postintervention periods. *P* values for both mean and variance effects from models corresponding to red code items are corrected for multiple testing using the Benjamini and Hochberg (1995) procedure. */**/*** denote that the coefficients are significant under the Benjamini and Hochberg (1995) procedure at false discovery rates of 0.1, 0.05, and 0.01, respectively.

As Table 10 illustrates, we find no detectable evidence of variance effects resulting from the checklist. Again, the robust effect is the increase in the citation rate of the elevated code item.

Appendix E: Abbreviated Checklist of Core Items

The following is an abbreviated checklist we piloted for an alternative inspection scheme in King County. We distill items down to 10 categories, and present the core question here, abstracting from some of the more complex situations.

Hand-washing	Did employees use proper hand-washing techniques? Did employees wash hands at appropriate times?
Bare hand contact	Are employees using suitable utensils or gloves to prevent bare hand (or arm) contact with ready-to-eat foods?
Cross-contamination	Are raw animal foods separated from ready-to-eat foods during storage, preparation, and display?
Cleaning/sanitization	Are food contact surfaces cleaned and sanitized before uses with different types of meat or when changing from raw animal foods to ready-to-eat food?
Temperature control	Is potentially hazardous food held at proper refrigeration or hot holding temperatures? If potentially hazardous food is held out of temperature control, is it being properly handled using Time as a Public Health Control?
Cooling	Are any cooling methods (time/temperature, shallow pan, or pieces of meat) out of compliance?
Approved source	Are there any food or beverage items that are not from an approved source?
Animal food	Is animal food being cooked to correct temperature and time?
Reheating	Is potentially hazardous food reheated to proper temperature and time prior to hot holding?
Toxic substances	Are toxic substances properly identified, stored, and used?

Appendix F: Food Safety Training Requirements

We researched health codes and inspection score sheets in the top 20 metropolitan areas by population. We code these training requirements along two dimensions. First, we classify for whom training is required: managers or employees. Our principal source for determining training requirements is the health code of each jurisdiction. To confirm the absence of a requirement, we also conducted web searches and consulted the health department websites and inspection forms. Following the FDA Model Food Code, every jurisdiction's health code requires manager food safety training or certification. Jurisdictions vary, however, in whether training or certification is required for all employees.[32]

[32] For our coding criteria, we do not distinguish between types of nonmanagerial employees. For instance, the food worker training in King County is required for all "food employees," defined as "individual[s] working with unpackaged food, food equipment or utensils, or food-contact surfaces." Washington Administrative Code § 1–201(B)(35).

Second, we code the severity of the violation. This coding was more complex, as the terminology to classify violations and the number of categories (typically two to three) vary across jurisdictions. Florida uses the labels "high priority," "intermediate," and "basic." Michigan uses "priority," "priority foundation," and "core." Los Angeles uses "major," "minor," and "good retail practices." In all jurisdictions, we rank order the violations classes by severity of risk to determine the severity of violating the training requirement for employees and/or managers. In many instances, the classification alone was clear about the rank order of severity of violations (e.g., critical vs. noncritical). In other instances, we consulted the jurisdiction's description of the classification to determine the category with highest risk. In Los Angeles, for instance, a major violation "warrants immediate correction and may require an immediate action"; a minor violation "does not pose an imminent health hazard, but does warrant correction"; and good retail practices are described as "low risk violations that do not require immediate action (Department of Public Health Environmental Health, County of Los Angeles 2014:1).

Table 11 summarizes our findings about the food safety training requirements. Jurisdictions are sorted by the stringency of training requirements, and each circle indicates the existence of a requirement, shaded by the severity of the violation. A solid

Table 11: Training and Certification Requirements Across Top 20 Metropolitan Jurisdictions (by Population)

	Jurisdiction	*Manager*	*Employee*
All employees	Seattle, WA	●	●
	Los Angeles, CA	◍	◍
	Houston, TX	●	○
	Phoenix, AZ	●	○
	San Francisco, CA	◍	○
	Riverside, CA	○	○
	San Diego, CA	○	○
Manager only	New York, NY	●	
	Dallas, TX	●	
	Washington, DC	●	
	Philadelphia, PA	●	
	Boston, MA	●	
	Detroit, MI	●	
	Minneapolis, MN	●	
	St. Louis, MO	●	
	Chicago, IL	◍	
	Miami, FL	◍	
	Tampa, FL	◍	
	Atlanta, GA	○	
	Denver, CO	○	

NOTES: The left column indicates whether managers are required to be trained. The right column indicates whether all employees or food handlers are required to be trained. Dots indicate the presence of the requirement and shading indicates the severity of the violation. A hollow dot (○) indicates that the training violation is in the lowest severity class; a gray dot (◍) indicates that the violation is an intermediate severity class (where more than two classes exist); and a black dot (●) indicates that it is in the highest severity class. The length and type of training can vary by employee status. Except for King County, jurisdictions requiring manager and employee training require more intensive training for the former.

black circle indicates that the jurisdiction classifies the violation in the most severe violation class (e.g., critical). A hollow circle indicates the violation is in the least severe violation class. Gray circles indicate an intermediate classification (e.g., priority foundation). The top panel represents jurisdictions that require all employees to be trained and the bottom panel represents jurisdictions that require only managers to be trained.

As Table 11 shows, there is substantial heterogeneity in both whether employees are required to be trained or certified, as well as the severity of the violation. King County's elevation of food handler training from noncritical to a critical violation makes it a clear outlier. The county is not only in the minority of jurisdictions requiring employee-wide training, but it is also the only jurisdiction classifying the item as the highest risk. That said, counties requiring manager and employee training typically require more intensive training for managers. In Houston, Los Angeles, and San Diego, for instance, managers take an eight-hour certification course, while employees only take a two- to three-hour course. King County requires managers and employees to go through the same training.