

INSPECTION DESIGN  
AND INSPECTOR BEHAVIOR\*

David Becker  
School of Public Health, University of Alabama  
DBecker@ms.soph.uab.edu

Ginger Jin  
Department of Economics, University of Maryland and NBER  
jin@econ.umd.edu

Phillip Leslie  
UCLA Anderson School of Management and NBER  
pleslie@anderson.ucla.edu

December, 2012

---

\*Thanks to the Los Angeles County Department of Health Services for providing the data used in this study.

# 1 Introduction

Audits, inspections and reviews are essential devices for determining many outcomes of economic significance: e.g. tax audits, aircraft safety inspections, nuclear powerplant safety inspections, immigration reviews, home inspections, vehicle emission inspections, workplace safety inspections, childcare certifications, nursing home inspections, university accreditation reviews, and student grading to name just a few. In all cases a primary concern is the consistency of inspectors (or auditors, graders or reviewers)—the degree to which the inspection outcome depends on the identity of the inspector. Inspectors may differ in many aspects, including experience, training, ability, harshness, and willingness to exert effort. In addition, the same inspector may perform inconsistently over time: e.g. accounting auditors may become lenient on firms they repeatedly audit, exam graders may become tired, or an inspector is simply in a bad mood. Even when the inspection outcome is based on ostensibly quantitative tests, if an inspector is needed to administer the test then invariably this person can influence the outcome in some way, either deliberately or inadvertently. In the extreme, if inspection outcomes are arbitrary then the process is wasteful at best, and may even be harmful to the individuals or firms being inspected as well as the general public who rely on the results (e.g. restaurant food safety inspections).

However, in any inspection system there will be potential to make design changes that can influence inspectors' performance. Consider the familiar example of exam grading. First, the exam may be written in such a way that there is an objectively right answer which helps graders to be consistent. The professor may write detailed solutions for the graders to use, with directions for how many points to award each component of each question. The completed exams may include only a student ID and no name, in order to eliminate biases from knowing the examinee's identity. Professors can re-grade a subset of the exams themselves to verify the grading. Students may also appeal their grade (Prendergast, 2007, emphasizes the effect an appeals process can have on inspector bias). These are all design features of an inspection process that can be structured with the purpose of promoting consistent assessment. It is easy to imagine the same kind of design choices being important for achieving reliable airline safety inspections, say, where the stakes are much higher. How much impact do these kind of inspection design changes have on the objectivity of inspectors?

In this study we analyze a dataset with unique features allowing us to analyze these issues based on actual inspections by government officials. Our data cover all 330,000 restaurant food-safety inspections performed by roughly 500 inspectors from the Los Angeles County Department of Health Services between 1995 and 2002. We observe the restaurant identity, inspector identity and outcome for each inspection. Most importantly, during this time frame there is variation in

the design of the inspection system, not dissimilar to the design of exam grading outlined above, at least some of which were aimed at promoting inspector consistency.

The most dramatic change in the inspection system to occur in the dataset is the introduction of restaurant hygiene grade cards in 1998. Prior research shows that the grade cards caused a 20% decrease in hospitalizations for food-related illnesses (Jin and Leslie, 2003). However, an important concern is whether the grade card policy also caused inspectors to become more lenient. Indeed, Figure 1 shows basic evidence that is highly suggestive of inspectors engaging in a form of leniency: bumping-up in which restaurants that are marginally below grade cutoffs are bumped up to the cutoff for a higher grade. Another concern that is often raised in jurisdictions considering the adoption of grade cards is that inspection outcomes are too arbitrary to shine a spotlight on them, because of the inconsistent assessments of inspectors. However, it is conceivable that grade cards will also cause inspectors to behave more consistently. This could be due to increased pressure from restaurant managers who now care much more about their inspection grade, or because inspectors exert more effort when the stakes are higher for the inspectees. We examine this issue in our analysis.

Economists have long presumed that regulatory enforcement is imperfect (e.g., Becker and Stigler, 1974), and a number of empirical papers have provided evidence that the behavior of inspectors depends on the design of the inspection system. For example, Nelson and Lynch (1984) provide evidence that teaching evaluations may contribute to grade inflation, and Blank (1991) shows that acceptance rates at the American Economic Review are lower for papers that are subject to double-blind refereeing, relative to single-blind reviews. More recent examples include the study by Garicano, Palacios-Huerta and Prendergast (2005) on home team biases of soccer referees, a study by Doyle (2007) highlighting the heterogeneity of child protection investigators, and Pierce and Snyder’s (2008) study of inspector heterogeneity in vehicle emission testing.<sup>1</sup>

The most relevant prior study, however, is Feinstein’s (1989) seminal analysis of the behavior of government safety inspectors at nuclear power plants. Central to his approach is the distinction between non-compliance (or actual violations) and detection by inspectors, which we also apply in our study. Based on data from about 1,000 inspections at 17 power plants over 3 years, Feinstein finds that variation in detection rates across inspectors is on a par with non-compliance rates at the power plants. In other words, the observed variation in inspection outcomes across nuclear power plants is equally driven by differences in actual plant safety and differences in inspector behavior. The main difference in our study is that we examine how changes in inspection design affect inspector heterogeneity—we consider whether the problem highlighted in

---

<sup>1</sup>See also Bar and Zussman (2012), Forbes, Lederman and Tombe (2012), Hodgson (2008), Iyengar (2007) and Li (2012). There are also studies showing the potential importance of decision fatigue on the part of an inspector (or judge, say). E.g., Danziger, Levav, Avnaim-Pesso (2011).

Feinstein’s work can be mitigated by changing the design of the inspection system.

In our data, as in Feinstein (1989), we observe only detected violations, as opposed to actual violations. This gives rise to the key identification issue: how do we separate compliance from undetected non-compliance? Since we observe the identity of the inspector and the restaurant in each inspection, the dataset is a multi-dimensional panel, allowing us to separately estimate inspector and restaurant fixed effects. Similar to Feinstein (1989), we then estimate a specification in which inspector heterogeneity is formally modeled as differences in detection probabilities, and restaurant heterogeneity is modeled as their compliance with the hygiene code. A perfect inspector reveals the level of compliance. As long as there is enough variation in inspector-restaurant matches (assumed to be exogenous), then we can identify the detection probability of each inspector.

## 2 Background

The data cover all restaurant hygiene inspections conducted by Los Angeles County Department of Health Services (DHS) inspectors between July 1995 and September 2002. Starting from a maximum score of 100, inspectors deduct points for each violation according to a set of criteria specified by the DHS. From the DHS inspection level data we observe the date of each inspection, the numerical inspection score, and unique inspector and restaurant identification codes. These identifiers allow us to link our inspection level data to other DHS files containing a number of inspector characteristics such as experience, gender and race. More importantly, knowledge of the inspector identity allows us to assess the role of restaurant and inspector behavior in determining hygiene scores.

We study five design regimes for the restaurant inspection system in LA. These regimes are summarized in the following table.

Regime	Time Frame	Num. Days	Description
1	1-Jul-1995 to 30-Jun-1997	730	Subjective component
2	1-Jul-1997 to 16-Nov-1997	139	No subjective component
3	17-Nov-1997 to 15-Jan-1998	58	News exposé aftermath
4	16-Jan-1998 to 18-Aug-1998	214	Grade cards
5	19-Aug-1998 to 30-Sep-2002	1,503	Owner-initiated inspections

All regimes involve point scoring—the inspector completes a form that details all the violations

and has a pre-specified point deduction associated with each violation. The final score is always computed by subtracting the total point deductions from a maximum score of 100. In addition to these violations, under Regime 1 inspectors also determine their overall subjective evaluation of the restaurant’s hygiene quality.<sup>2</sup> In Regime 2 the subjective assessment is removed, leaving all the specific violations intact.

We speculate that the subjective component was included in Regime 1 because the inspection system prior to Regime 1 (before our sample period) was an entirely subjective assessment. The introduction of a point-scoring system could have been seen as a major change by the inspectors who emphasized the need for subjective/discretionary assessment. This may explain why Regime 1 was a hybrid of objective and subjective criteria, and was subsequently phased out in Regime 2.

The most dramatic change in the inspection design during our sample period is the introduction of restaurant hygiene grade cards that restaurants are required to prominently post in their windows near the entrance. An A-grade is given for scores above 90, a B-grade for scores between 80 and 89, a C-grade for scores between 70 and 79, and a card bearing the actual numerical score for scores below 70. As with the prior research, we treat this policy change as exogenous, because it was a rapid regulatory response to a surprise hidden-camera news exposé.

Disclosure policies are designed to improve welfare through demand-driven incentives for high quality (better performance). The existing empirical studies into report cards have focused exclusively on their effects on consumer and firm-level behavior. Studies of the demand response have examined public reporting in a variety of contexts, including HMOs (Dranove and Dafny, 2008, Jin and Sorenson, 2006), hospitals (Mennemeyer et al., 1997), public schools (Hastings and Weinstein, 2006) and restaurants (Jin and Leslie, 2003). Collectively, this research demonstrates that report cards do influence behavior, although the effects vary with the accessibility of the information, the underlying uncertainty regarding product quality, and the salience of the reported product characteristics. On the supply side, researchers have found evidence that firms respond to information disclosure in ways that can both enhance (Jin and Leslie, 2003) and reduce (Dranove et al, 2003) welfare. The existing research demonstrates the importance of specific design features of quality reporting programs to their social value. This study expands upon the previous work by examining how restaurant report cards impact the behavior of the inspectors themselves.

The design change for Regime 5 stemmed from restaurant’s complaint that sometimes they are able to quickly fix the problems that led to a B or C-grade, but they are brandished with a

---

<sup>2</sup>Inspectors could select between five possible values for their subjective *establishment status score*: Excellent (no deduction), Good (5 points), Average (20 points), Fair (30 points) and Poor (40 points).

lower grade for months until their next regular inspection. In response, the DHS introduced a modification allowing restaurants to request an inspection—a so-called owner-initiated inspection. The restaurant pays a fee to help cover the cost (less than \$200 at the time) and the inspector should arrive randomly within the next two weeks. Prendergast (2007) suggests that inspectors may be harsher when there is the possibility of appeal by the inspectee.

Note that Regime 3 does not represent a design change for the inspection system. Regime 3 is the brief period between the news exposé and the implementation of the grade cards. During this time there is no formal change in policy, although anecdotally we know that inspectors were harsher during this time. While it would be interesting to document the changed behavior of inspectors in this period, the shortness of the time frame makes it infeasible.

Finally, it is important to recognize that the change from Regime 1 to Regime 2, and the change from Regime 4 to Regime 5, are both relatively invisible to consumers. In contrast, grade cards are an increase in the provision of information to consumers about restaurant’s hygiene quality. As the prior research shows, this caused consumers to become sensitive to restaurant hygiene quality, in turn causing restaurants to improve their hygiene quality.<sup>3</sup> It is conceivable that restaurants may also react to the grade cards by seeking to influence the inspectors. This need not involve bribery, but simply that restaurant managers are much more attentive to the inspector, trying to influence their assessment through personal conduct. Indeed, inspectors have told us that their jobs became more difficult when grade cards were introduced.

## 2.1 Overview of Data

Across all regimes there are 505 unique DHS inspectors, who performed an average of 656 inspections on 280 unique restaurants. There is significant variation in total inspector workload, which is partly attributable to differences in the duration of inspector employment. There are 28,105 unique restaurants in our sample. The average restaurant was inspected 11.8 times by 5 different inspectors. There are significant differences in the degree of inspector turnover across restaurants that are likely related to differences in restaurant density and staffing patterns across Los Angeles County. Restaurants average slightly more than 2 inspections per year (conditional on appearing in the data in a given year).

Table 1 provides summary statistics for the inspectors and restaurants, for each regime. Regime 3 is excluded from the table, as it is for all of the analysis in the paper, due to the brief period of time this regime covered. We included it in the above discussion only for completeness.

---

<sup>3</sup>See Jin and Leslie (2003).

Roughly 25% of the 331,289 total inspections are from the pre-grade era, with the bulk of these inspections occurring during Regime 1. With the elimination of the subjective point deduction in regime 2, the average hygiene score increased from 77.3 to 85.9. Although restaurants were not assigned letter grades, the fraction of restaurants receiving "A" and "B" level scores increased significantly between regimes 1 and 2. Average hygiene scores continued to increase following the implementation of grade cards, but in a much different manner than the pre-grade card period. A relatively small 3.4 point increase in the average score between regime 2 and 3 was associated with a 20 percentage point increase (43.2% to 63.2%) in the fraction of restaurants with scores above 90 (A-grade), and a large decrease (25.3 to 10.2%) in the fraction of restaurants receiving scores below 80 (C-grade or worse). These trends continue in regime 4, which provides over two-thirds of our inspection data. The average score increased by 1.5 points to 90.8, while the percentage of restaurants receiving A-grades rose to 77.2% and the fraction receiving C-grades or worse fell to just 3.4%.

The table also shows that the dispersion in the distribution (across inspectors) of inspectors' average inspection score has been narrowing over time. In Regime 1 the interquartile range is about 20 points (from 69 to 89), this narrows under Regime 2 to 10 points, then 5 points under Regime 4, and finally 3 points under Regime 5. Some of this narrowing is driven by restaurants becoming less heterogeneous over time, so one must be cautious in interpreting these statistics.

Figure 1 shows significant changes in the distribution of hygiene scores across the four inspection regimes. The figure shows only the densities for scores above 40, which is the key region of interest. The most striking difference across the distributions is the spiking that is apparent under Regimes 4 and 5. Whereas restaurants were equally likely to receive a score of 89 or 90 in Regime 2, restaurants were over 4 times more likely (9.46% vs. 2.05%) to receive a score of 90 (vs. 89) in Regime 4. This spiking became even more pronounced in Regime 5, as restaurants were 26 times more likely to receive a score of 90 (18.5%) than a score of 89 (0.7%). This massing of scores at the grade-thresholds provides basic evidence that inspector behavior is influenced by grade cards and raises questions about the long-run utility of the report card program.

## 2.2 Basic Evidence of Inspector Heterogeneity

We assume there are three factors contributing to an inspection score: the hygiene effort by the restaurant, the ability of the inspector to detect violations, and random hygiene shocks. In the next section we detail a model for how these three components interact, which we then estimate. Before doing that we can examine straightforward evidence which is suggestive of the relative

importance of these three factors in explaining the distribution of inspection scores. Separately for each design regime, we regress hygiene scores on inspector fixed effects, restaurant fixed effects and a dummy variable for whether an inspection was conducted by a repeat inspector. The motivation is to compare the dispersion in the estimated inspector and restaurant fixed effects, and to see how these distributions have changed under different regimes.

Table 2 reports the standard deviations of the various fixed effects and the residuals (hygiene shocks). As shown in Table 1, there are few restaurants with multiple inspections under Regimes 2 and 4, and these restaurants are undoubtedly non-representative of the broader population of restaurants. We therefore focus our attention on the estimates for Regime 1 and Regime 5. The standard deviation of the estimated inspector fixed effects declines from 13.4 under Regime 1 to 5.8 in Regime 5. At face value, this suggests inspectors have become dramatically less heterogeneous after grade cards. We also find a dramatic reduction in dispersion in the restaurant fixed effects, from 14.8 to 4.5. Indeed, the relative decline in the dispersion of restaurant fixed effects is greater than for inspectors, as shown in the second to last row of Table 2. The table also reports that the variance of the residuals has also declined, but not as dramatically (from 7.4 to 4.3). Based on this specification, in Regime 5 the variance in the hygiene shocks are on a par with the variation in restaurant fixed effects (across restaurants), which is also not much less than the variation in inspector fixed effects.

The results in Table 2 suggest that inspectors and restaurants have both become less heterogeneous in their hygiene quality over time—indeed, to a dramatic degree. The analysis here also helps to highlight the multidimensional panel nature of the dataset, in which we observe restaurants over time, inspectors over time, and variation in the matching of inspectors to restaurants (as well as variation in the inspection design). However, we must be cautious in our interpretation of the estimated fixed effects from this linear panel model, for at least a few reasons. First, since the maximum score is 100, as restaurants improve hygiene quality (the mean of the hygiene score distribution increases) it is inevitable that the dispersion in the score distribution will also shrink, which confounds improvements in average hygiene quality with reduced restaurant and inspector heterogeneity.

Second, the model assumes that the score for any given inspection is the additive sum of the inspector fixed effect, the restaurant fixed effect, and a random hygiene shock (plus a repeat inspector effect). Hence, the impact on the score of any given inspector is the same absolute effect at a restaurant that tends to have good hygiene as it is for a restaurant that tends to have bad hygiene. That is unrealistic. Third, this basic model does not explicitly allow for differences in inspector behavior when the score is near a threshold (such as the bumping up from 89 to 90). The model we present in the next section is designed to address both of these concerns, allowing for a more realistic interpretation of the role of inspector heterogeneity—heterogeneity

across inspectors in their tendency to detect food safety violations.

### 3 Model

In this section we present a model of restaurant hygiene scoring. Fundamental to our approach is the distinction between actual, detected, and reported violations. Actual violations are dependent on each restaurant's hygiene quality (and a random hygiene shock). Detection of violations depends on inspector behavior. Only actual violations can be detected and any actual violation may or may not be detected. A hygiene score is assigned to a restaurant based on how many violations are detected and whether the inspector reports all violations.

The scoring mechanism works as follows. When an inspector  $i$  steps into restaurant  $r$ , the restaurant's actual hygiene violations ( $\lambda_r$ ) depends on the degree of hygiene quality ( $\gamma_r$ ) that the restaurant can control, plus a random hygiene shock  $\epsilon$  that is out of the restaurant's control. We model hygiene quality ( $\gamma_r$ ) as a function of the restaurant's observable characteristics ( $X_r$ ) and the restaurant manager's unobservable effort:  $\alpha_r$ . We assume  $\alpha_r$  conforms to a log normal distribution with mean  $\mu_r$  and variance  $\sigma_r^2$ .

Let  $\delta_i$  denote the probability that inspector  $i$  detects all violations. We assume this detection probability depends on an inspector's observable characteristics,  $X_i$ , as well as her unobservable stringency:  $\alpha_i$ . We assume  $\alpha_i$  conforms to a normal distribution with mean  $\mu_i$  and variance  $\sigma_i^2$ . Beginning with the maximum score of 100, inspector  $i$  deducts points corresponding to *detected* violations. For simplicity, we assume equal weight per violation (one point).

It follows that the detected score  $s_{ir}$  is given by:

$$\begin{aligned} s_{ir} &= 100 - \lambda_r \delta_i \\ &= 100 - (\alpha_r + \beta_r \cdot X_r + \epsilon) \cdot \frac{\exp(\alpha_i + \beta_i \cdot X_i)}{1 + \exp(\alpha_i + \beta_i \cdot X_i)}, \end{aligned}$$

where

$$\begin{aligned} \alpha_r &\sim \text{lognormal}(\mu_r, \sigma_r^2) \\ \alpha_i &\sim N(\mu_i, \sigma_i^2) \\ \epsilon &\sim N(0, \sigma_\epsilon^2). \end{aligned}$$

The score is bounded between 0 and 100, which we implement by assuming  $s_{ir}$  is censored at 0 if the inspector reports more than 100 violations (never happened in our data); or censored at

100 if the restaurant is so clean that its actual dirtiness is negative. To be consistent with the data, we also assume the detected score is rounded up to the closest integer between 0 and 100.

In the above description we suppress the time dimension which we now introduce. As described above, our data cover multiple regimes (i.e., periods in which the design of the inspection system has varied). Also, within any given regime we observe multiple inspections at any given restaurant, and multiple inspections by each inspector. Let  $t$  index inspections, so that  $s_{irt}$  denotes the score detected by inspector  $i$  at restaurant  $r$  during inspection  $t$  (in the corresponding regime).

In addition to detecting violations an inspector may choose to inflate the detected score, perhaps due to pressure from the restaurant manager, because she feels bad for giving out a low number, or some other reason. There are many potential forms of score inflation and here we focus on a particular form: bumping up a score from just below a grade threshold. Prior to grade cards, hygiene inspection outcomes are not readily available to consumers and restaurants have a minimal incentive to influence the inspector for a higher score. That said, some restaurant managers may take pride in earning a perfect score, which could explain why we observe very few 99s and a spike at 100 in Regime 2. In comparison, no inspection gets an imputed score of 99 or 100 in Regime 1, probably due to imputation. To reflect these data patterns under Regime 2 we incorporate a probability of bumping up from 97 to 100 ( $\pi_{from97}$ ), from 98 to 100 ( $\pi_{from98}$ ), and from 99 to 100 ( $\pi_{from99}$ ). All these probabilities are assumed to be zero in Regime 1. Hence, the expected reported score ( $\hat{s}_{irt}$ ) conditional on the detected score  $s_{irt}$  can be written as:

$$\text{Regime 1: } E(\hat{s}_{irt}|s_{irt}) = s_{irt}$$

$$\text{Regime 2: } E(\hat{s}_{irt}|s_{irt}) = \begin{array}{ll} s_{irt} & \text{if } s_{irt} < 98 \text{ or } s_{irt} = 100 \\ s_{irt} + (100 - s_{irt}) \cdot \text{prob}_{from97} & \text{if } s_{irt} = 97 \\ s_{irt} + (100 - s_{irt}) \cdot \text{prob}_{from98} & \text{if } s_{irt} = 98 \\ s_{irt} + (100 - s_{irt}) \cdot \text{prob}_{from99} & \text{if } s_{irt} = 99. \end{array}$$

Following the introduction of grade cards, consumer responsiveness to hygiene quality provides a strong incentive for restaurants to obtain a high letter grade. Since almost all inspections result in a C-grade or better, we focus on threshold bumping from C to B, from B to A, and from near 100 to 100. We model bumping up as a sequence of two independent actions: one is whether to bump-up from a particular score, and the other is which score to bump-up to. We therefore specify the probability of bumping occurring, the scores that can be bumped up, and the scores that an inspection can be bumped up to. Let  $\pi_{from_s}$  denote the probability of bumping up from  $s$ . We denote the lowest scores within each grade that may be bumped up

as  $\underline{s}_C$ ,  $\underline{s}_B$  and  $\underline{s}_A$ . The probability of bumping up to score  $y$  is denoted by  $\pi_{to_y}$ . Finally, the highest A-grade (B-grade) score that an inspector may bump-up to is denoted by  $\bar{y}_A$  ( $\bar{y}_B$ ). We can now specify the expected score under Regimes 4 and 5:

$$\text{Regime 4 or 5: } E(\hat{s}_{irt}|s_{irt}) = \begin{array}{ll} s_{irt} & \text{if } \underline{s}_B \leq s_{irt} \leq 89 \\ s_{irt} + \sum_{y=90}^{\bar{y}_A} (y - s_{irt}) \cdot \pi_{to_y} \cdot \pi_{from_s} & \text{if } \underline{s}_C \leq s_{irt} \leq 79 \\ s_{irt} + \sum_{y=80}^{\bar{y}_B} (y - s_{irt}) \cdot \pi_{to_y} \cdot \pi_{from_s} & \text{if } \underline{s}_A \leq s_{irt} \leq 99. \\ s_{irt} + (100 - s_{irt}) \cdot \pi_{from_s} & \end{array}$$

To summarize, our model is designed to focus on the behavior of inspectors, allowing us to examine how inspector behavior changes under different inspection design regimes. Our approach incorporates the notion that actual hygiene is partly due to restaurant behavior and partly due to nature, although the inspector does not distinguish between these factors. The role of the inspector centers on translating actual hygiene into a reported score, that involves inspector effort to detect violations and inspector willingness to bump-up scores.

## 4 Estimation, Identification and Results

### 4.1 Estimation

The central focus of our analysis is to examine how the distribution of inspectors' detection probabilities ( $\delta_i$ ) changes according to the design of the inspection system, while controlling for changes in the distribution of restaurants' underlying hygiene quality ( $\lambda_r$ ). We therefore allow the parameters of these distributions to vary across regimes.

Each regime includes three types of parameters. First,  $\beta_i$  and  $\beta_r$  capture the influence of observable inspector or restaurant characteristics on detected scores. Observable inspector characteristics include gender and race (time invariant), and whether the inspector is a repeat inspector at this restaurant (varies by inspection). We have limited restaurant observables, although we do observe the type of inspection: routine, consumer complaint, suspected food poisoning or owner initiated.<sup>4</sup> We presume that the type of inspection may influence the behavior of either inspectors or restaurants, which would not be separately identified (accept by functional form assumptions). We therefore include the inspection type on the restaurant-side of the specification with the caveat that we do not interpret this as a pure impact on restaurant behavior.

---

<sup>4</sup>Owner initiated inspections are only possible under Regime 5.

The second type of parameter is intended to capture unobservable heterogeneity across inspectors and across restaurants. While it would be ideal to estimate inspector fixed effects and restaurant fixed effects under each regime, as we did above in the descriptive analysis, such an approach is infeasible for so many inspectors and restaurants under this nonlinear model, that we estimate via simulated GMM (as detailed below). We therefore take a random effects approach. In each regime we assume that the random effect associated with each inspector ( $\alpha_i$ ) remains time-invariant for that inspector. Similarly, in each regime we assume that the random effect associated with each restaurant ( $\alpha_r$ ) remains time-invariant for that restaurant. Changes in the distribution of  $\alpha_i$  tell us whether inspectors are more consistent across regimes in their ability to detect violations, and changes in the distribution of  $\alpha_r$  across regimes tell us whether restaurant owners improve actual hygiene and become more similar to each other over time. By assuming time-invariant behavior within each regime, the approach attributes all other variation in scores to the random hygiene shock ( $\epsilon$ ). This may include a restaurant manager changing their cleaning effort within a regime (due to labor turnover, say), or an inspector varying their behavior due to their mood or tiredness. By estimating the variance of the hygiene shock ( $\sigma_\epsilon^2$ ) separately for each regime, we can infer whether the within-regime randomness is sensitive to regime change.

The third type of parameter relates to the probability of bumping up the reported score. As modeled above, we assume no bump-up behavior in Regime 1, potentially some bump-up from 97, 98 or 99 to 100 in Regime 2, and bump-ups around the grade thresholds of 80 and 90, and also 100 in Regimes 4 and 5. In Regime 5 we have the most data and can allow the most flexibility: we allow potential bump-up from 78 and 79 to 80, 81, 82 or 83; from 85, 86, 87, 88 or 89 to 90, 91, 92 or 93; and from 98 or 99 to 100. Regime 4 is a much shorter time frame and has much fewer observations, and so we restrict Regime 4 bump-up behavior to bumping from 78 or 79 to 80, 81 or 82; from 87, 88 or 89 to 90 or 91; and from 98 or 99 to 100.<sup>5</sup>

Our estimation method is simulated GMM. For each regime we search for parameters that best match a weighted sum of differences between data moments and simulated moments. The objective function is:

$$Q_N = \left[ \frac{1}{N} \sum_{i=1}^N h_i(\theta) \right]' W_N \left[ \frac{1}{N} \sum_{i=1}^N h_i(\theta) \right]$$

where

$$h_i(\theta) = \hat{m}_i(\theta) - m_i(\text{data}),$$

$N$  is the number of observations,  $k$  is the number of moments,  $h_i$  is a  $k \times 1$  vector for the difference between the simulated moments ( $\hat{m}_i$ ) and the data moments ( $m_i$ ) for observation  $i$ ,

---

<sup>5</sup>Within each regime and around a particular grade cutoff, the probability of bumping up to allowed scores must add up to be one. This implies  $\pi_{to80} + \pi_{to81} + \pi_{to82} + \pi_{to83} = 1$  and  $\pi_{to90} + \pi_{to91} + \pi_{to92} + \pi_{to93} = 1$  for Regime 5. And under Regime 4:  $\pi_{to80} + \pi_{to81} + \pi_{to82} = 1$  and  $\pi_{to90} + \pi_{to91} = 1$ .

$W_N$  is a  $k \times k$  weighting matrix, and  $\theta$  is a vector of parameters.

We use three sets of moments. The first set focuses on summarizing the overall score distribution, including mean score of all inspections, the second moment of mean score per inspector, and the second moment of mean score per restaurant.<sup>6</sup> The second set of moments captures correlations between inspection scores and observable characteristics. If a characteristic is categorical (e.g., inspector race, gender, and whether an inspector is repeat for the restaurant in an inspection) we compute the first and second moments of score for each category. If the characteristic is continuous (e.g., inspector’s hourly wage rate, job tenure, and the number of inspections this inspector has done in our dataset before this inspection) we compute the correlation between inspection scores and the characteristic. The third set of moments refer to the score density on a particular range, or a particular point of the score distribution.

In total we utilize 71 moments in Regime 1, 73 moments in Regime 2, 79 moments in Regime 4 and 83 moments in Regime 5. Regime 1 has two fewer moments than Regime 2 because we do not include the density at 99 and 100 as the raw data for Regime 1 show zero density on these two scores. Regime 4 has six more moments than Regime 2 because in Regime 4 we add dummies for whether the inspector is a new hire after grade card are implemented, and whether an inspection is sparked by suspected food poisoning.<sup>7</sup> There are additional dummies in Regime 5 for whether an inspection is owner-initiated or DHS-initiated. In each regime the number of moments is greater than the number of parameters, implying over-identification.

To perform simulated GMM for a specific regime we take draws of inspector random effects ( $\alpha_i$ ), restaurant random effects ( $\alpha_r$ ), and random hygiene shocks ( $\epsilon_{irt}$ ) from a standardized lognormal or normal distribution, and then transform them according to the starting values of their mean and variance. We compute simulated scores according to our model and conjectured parameter values, then compute simulated moments and compare them to the data moments in order to calculate the objective function. To determine the optimal weighting matrix we initially assume equal weight on each moment and search for optimal parameters that minimize the objective function. Then we compute optimal weighting matrix from the first-stage estimates, use the optimal weighting matrix to redefine the GMM objective function, and redo the estimation. Both stages use simplex search. Appendix A describes how we numerically obtain standard errors for the second-stage estimates. Appendix B describes how we define the optimal weighting matrix from the first-stage estimates. The above procedure is repeated for each regime separately.

---

<sup>6</sup>By definition, the mean of mean score per inspector, and the mean of mean score per restaurant, is always equal to the mean of all inspections.

<sup>7</sup>There were too few inspections for suspected food poisoning in Regime 2 to be able to identify this coefficient.

## 4.2 Identification

Identification of inspector heterogeneity in our model stems from the multi-dimensional panel nature of the data. Within each regime, each inspector performs multiple inspections, at many different restaurants and sometimes on multiple occasions at the same restaurant. Restaurants are also inspected multiple times by various different inspectors and sometimes by the same inspector. These are the features of the data that provide very intuitive identification of the simple linear fixed effects specification discussed in the background section. For example, the average inspection score of a restaurant (controlling for some loosely defined inspector heterogeneity in that specification) reveals the restaurant manager’s preference for hygiene and/or the cost of maintaining good hygiene. The same intuitive source of identification carries over to the structural model, although the model now delivers a particular interpretation for inspectors’ heterogeneity—differences in detection probabilities across inspectors.

The approach also relies on a normalization. Consider again the linear model with fixed effects. When there are two sets of fixed effects (inspector and restaurant) a complete list of restaurant dummies will be collinear with a complete list of inspector dummies. Therefore, at least one inspector (or restaurant) dummy has to be dropped and the remaining fixed effects are relative to that dropped identity. Similar logic applies to our structural model. The average of the mean score per inspector is by definition equal to the average of the mean score per restaurant, and both of them are equal to the average of all inspections. Therefore, one can always readjust the mean of restaurant random effects and the mean of the inspector random effects to match the average score in the raw data. In other words, when we compare across regimes, changes in the average score across regimes can be explained by either changes in the mean of restaurant random effects or changes in the mean of inspector random effects. To overcome this under-identification problem, we choose to normalize on the side of inspectors.

Since our aim is to separately identify the distribution of inspector random effects for each regime, we must maintain a consistent normalization across regimes. Otherwise, if we implement a different normalization in each regime, we may falsely infer that inspector behavior is changing on average, even though it may not be. The solution we adopt is to select an inspector (or set of inspectors) that are present in all regimes, and for which we can reasonably assert that their behavior does not change from regime to regime. This implies that any conclusions about changes in the average level of inspector behavior are conditional on the assumption that this set of inspectors did not change their behavior. That is a strong assumption, but it is innocuous with respect to changes in the dispersion (or heterogeneity) of inspector behavior, which is our key focus.

We use a data-driven method to determine the specific normalization. To define a set of benchmark inspectors we focus on Regime 1 and Regime 5 because they contain the most data. In each regime we regress raw score on inspector fixed effects and restaurant fixed effects. We then match the inspector fixed effects in Regime 1 and Regime 5 by inspector identity. This reveals a set of inspectors that have valid inspector fixed effects in both Regimes 1 and 5. Within this set of inspectors we standardize each inspector’s fixed effect within each regime, sum up the two fixed effects of the same inspector and rank inspectors according to this sum. The top five inspectors that have the highest sum of inspector fixed effects represent the most stringent inspectors during the first and last regimes of our data. Taking these five inspectors as the benchmark, we normalize their detection probability as 0.9 in every regime.

More generally, a model of compliance and detection always faces the problem of how to separately identify these two components—how to separate compliance from undetected non-compliance? This was the focus on Feinstein (1989), and we apply his explanation. A perfect inspector (or at least the inspector with the highest draw of random effect) reveals the level of compliance. As long as we have enough variation in inspector-restaurant matches (which we also assume to be exogenous), and assuming restaurants level of compliance is fixed (on average, within regime), then we can identify the detection probability of each inspector, whose distribution gives us estimates for the mean and variance of inspector random effects.

As described above, our model incorporates two types of inspector behavior: one is imperfect detection that depends on inspector random effects and observable characteristics, and the other is grade inflation in the form of score bump-ups. The former is identified from the (smooth) shape of score distribution, with the help of the panel data structure. The latter is identified by empirical deviations at particular points of the score distribution. To separate the latter from the former, we have to restrict score bump-ups to a subset of the score distribution, namely scores around grade cutoffs (80 and 90) or the upper limit (100). These particular restrictions are partly driven by the economic incentive from grade cards (as demonstrated in Jin and Leslie, 2003), and partly by the spikes right around the grade cutoffs in the raw data (Figure 1). Hence, while inspector’s detection behavior is identified by the overall score distribution, the tendency for inspectors to bump-up scores is identified by data frequency in the immediate vicinity of grade thresholds.

### 4.3 Results

## 5 Discussion

5.1 Why do inspectors vary in their performance?

5.2 How large is the inconsistency of inspection outcomes?

5.3 Does grade inflation in restaurant inspections matter?

5.4 Does grade inflation become exacerbated over time?

5.5 How could a change in the inspection design reduce grade inflation?

## 6 Conclusion

An interesting question is to what degree does inconsistency in the behavior of inspectors cause firms to change their effort levels? Conceivably, this effect may be either positive (if firms are risk averse, say) or negative (if the inconsistency of inspection outcomes is so severe that firm effort has a negligible impact). Our approach does not explicitly model restaurants effort choices.

Inspection design is not just an issue for government inspections. Many firms also provide inspection services such as product reviews (e.g. movie or wine reviews), certification of collectibles (e.g. baseball cards), house inspections, and vehicle emissions tests. In some cases competing firms may even seek to differentiate themselves by offering higher quality inspections than their competitors, providing more consistent or reliable results. This study shows that the design of the inspection system plays an important role in attaining high quality inspection services.

## Bibliography

- : Bar, T. and A. Zussman (2012): “Partisan Grading,” *American Economic Journal: Applied Economics*, 4(1), 30–48.
- Becker, W.E. (1982): “The Educational Process and Student Achievement Given Uncertainty in Measurement,” *American Economic Review*, 72(1), 229–36.
- Blank, R.M. (1991): “The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review,” *American Economic Review*, 81(5), 1041–67.
- Danziger, S., J. Levav and L. Avnaim-Pesso (2011), “Extraneous Factors in Judicial Decisions,” *Proceedings of the National Academy of Sciences USA*, 108:6889-6892.
- Doyle, J.J. (2007): “Child Protection and Child Outcomes: Measuring the Effects of Foster Care,” *American Economic Review*, 97(5), 1583–610.
- Feinstein, J.S. (1989): “The Safety Regulation of U.S. Nuclear Power Plants: Violations, Inspections, and Abnormal Occurrences,” *Journal of Political Economy*, 97(1), 115–54.
- Forbes, S.J., M. Lederman and T. Tombe (2012): “Quality Disclosure Programs with Thresholds: Misreporting, Gaming, and Employee Incentives,” *Mimeo*.
- Garicano, L., I. Palacios and C. Prendergast (2005): “Favoritism Under Social Pressure,” *The Review of Economics and Statistics*, 87(2), 208–16.
- Goldin, C. and C. Rouse (2000): “Orchestrating Impartiality: The Impact of ‘Blind’ Auditions on Female Musicians,” *American Economic Review*, 90(4), 715–41.
- Glaeser, E.L. and A. Shleifer (2001): “A Reason for Quantity Regulation,” *American Economic Review*, 91(2), Papers and Proceedings of the 113th Annual Meeting of the American Economic Association, 431–35.
- Hodgson, R.T. (2008): “An Examination of Judge Reliability at a Major U.S. Wine Competition,” *Journal of Wine Economics*, 3(2), 105–13.
- Iyengar, R. (2007): “An Analysis of the Performance of Federal Indigent Defense Counsel,” *Mimeo*.
- Jacob, B.A. and S.D. Levitt (2003): “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating,” *The Quarterly Journal of Economics*, 118(3), 843–77.
- Jin, G. and P. Leslie (2003): “The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards,” *Quarterly Journal of Economics*, 118(2), 409–51.
- Jin, G. and P. Leslie (2007): “Reputational Incentives for Restaurant Hygiene,” *American Economic Journal: Microeconomics*, 1(1), 236–67.

- Li, D. (2012): “Information, Bias, and Efficiency in Expert Evaluation: Evidence from the NIH,” *Mimeo*.
- Macher, J.T., J.W. Mayo and J.A. Nickerson (2006): “Exploring the Information Asymmetry Gap: Evidence From FDA Regulation,” *Mimeo*.
- Medeiros, P. and A. Wilcock (2006): “Public Health Inspector Bias and Judgement,” *Food Protection Trends*, 26(12).
- Pierce, L. and J. Snyder (2008): “Ethical Spillovers in Firms: Evidence from Vehicle Emissions Testing,” *Management Science*, 54(11), 1891–903.
- Prendergast, C. (2007): “The Motivation and Bias of Bureaucrats,” *American Economic Review*, 97(1), 180–96.
- Prendergast, C. and L. Stole (1996): “Impetuous Youngsters and Jaded Oldtimers: Acquiring a Reputation for Learning,” *Journal of Political Economy*, 104(6), 1105–34.
- Sabot, R. and J. Wakeman-Linn (1991): “Grade Inflation and Course Choice,” *Journal of Economic Perspective*, 5(1), 159–70.
- Simon, P., P. Leslie, G. Run, G. Jin, R. Reporter, A. Aguirre and J. Fielding (2005): “Impact of Restaurant Hygiene Grade Cards on Foodborne Disease Hospitalizations in Los Angeles County,” *Journal of Environmental Health*, 67(7), 32-8.

Table 1: Score Statistics, Restaurants and Inspectors Characteristics by Regime

	Regime 1	Regime2	Regime4	Regime5
<b><i>Inspection Statistics</i></b>				
Mean Score	77.33	85.91	89.22	90.78
Score (25 <sup>th</sup> percentile)	69	79	85	90
Score (50 <sup>th</sup> percentile)	80	88	91	91
Score (75 <sup>th</sup> percentile)	88	95	95	94
Number of Inspections	71,934	11,333	19,984	228,007
% Due to Complaints/Foodpoisoning	5.65%	7.49%	7.89%	3.21%
<b><i>Inspector Characteristics</i></b>				
# of Inspectors	240	201	239	433
# of New Inspectors after Grade Card	/	/	22	230
Mean # of Inspections per Inspector	299.73	56.38	83.62	526.58
# of Inspections per Inspector (25 <sup>th</sup> percentile)	61	25	19	42
# of Inspections per Inspector (50 <sup>th</sup> percentile)	260.5	47	80	267
# of Inspections per Inspector (75 <sup>th</sup> percentile)	487	84	134	909
Mean Score per Inspector (25 <sup>th</sup> percentile)	69.15	81.48	87.06	89.67
Mean Score per Inspector (50 <sup>th</sup> percentile)	76.62	86.34	89.33	90.98
Mean Score per Inspector (75 <sup>th</sup> percentile)	89.545	91.039	91.666	92.424
% Hispanic Inspectors	4.17%	4.48%	5.02%	6.24%
% Inspections by Hispanic Inspectors	5.86%	4.46%	5.31%	4.95%
% Asian Inspectors	2.50%	2.99%	3.77%	6.93%
% Inspections by Asian Inspectors	1.15%	1.31%	1.37%	7.30%
% Female Inspectors <sup>1</sup>	29.17%	30.85%	28.87%	28.18%
% Inspections by Female Inspectors	28.97%	31.66%	28.77%	29.80%
Tenure	4.35	4.43	4.52	4.07
Hourly Wage Rate	34.93	33.21	33.63	37.664
<b><i>Restaurant Characteristics</i></b>				
# of Restaurants	19,257	9,959	15,402	26,475
Mean # of Inspections per Restaurant	3.76	1.14	1.36	8.55
# of Inspections per Restaurant (25 <sup>th</sup> percentile)	2	1	1	5
# of Inspections per Restaurant (50 <sup>th</sup> percentile)	4	1	1	9
# of Inspections per Restaurant (75 <sup>th</sup> percentile)	5	1	2	12
Mean Score per Restaurant (25 <sup>th</sup> percentile)	70.167	80	85	89.143
Mean Score per Restaurant (50 <sup>th</sup> percentile)	79	88	91	91.667
Mean Score per Restaurant (75 <sup>th</sup> percentile)	84.5	93	94	93.5

<sup>1</sup>Around 23% of the inspectors in regime 1 to 4 and 27% of the inspectors in regime 5 missed records of gender.

Table 2: Comparison of SD of Inspector and Restaurant Fixed Effects and Error by Regime

	Regime1	Regime2	Regime4	Regime5
<i>Regression of Restaurant AND Inspector Fixed Effects</i>				
SD Inspector FE (A) <sup>1</sup>	13.391	19.313	9.353	5.757
SD Restaurant FE (B)	14.771	22.416	10.023	4.475
SD Residuals (C)	7.445	2.548	2.874	4.262
A/B	0.907	0.862	0.933	1.286
A/C	1.799	7.580	3.254	1.351

<sup>1</sup> SD computes the standard deviation of fixed effects estimated from the linear regressions

$$S_{irt} = \gamma D(\text{repeated inspection}) + \alpha_i + \alpha_r + \varepsilon_{irt}$$

Table 3: Main Model Estimation Results (Parameters for RE and Inspector Characteristics)

	Regime1	Regime2	Regime4	Regime5
<i>Restaurant and Inspector Random Effect Distribution</i>				
Restaurant Mean	47.7577 (0.1932)***	19.4691 (0.1578)***	15.332 (0.1079)***	14.6074 (0.0172)***
Restaurant Std Dev	19.3279 (0.1032)***	9.0354 (0.3271)***	9.5745 (0.2477)***	6.4408 (0.0247)***
Inspector Mean	0.6464 (0.0083)***	1.8933 (0.0475)***	1.6716 (0.05)***	1.6067 (0.0916)***
Inspector Std Dev	0.9266 (0.0022)***	1.4618 (0.0134)***	0.4604 (0.0219)***	0.312 (0.0266)***
Noise Variance	5.4276 (0.2574)***	10.2995 (0.1196)***	6.821 (0.0266)***	5.7933 (0.0354)***
<i>Effects of Inspector Characteristics on Detection Probabilities</i>				
Repeated Inspection	-0.196 (0.0056)***	-0.0634 (0.0069)***	-0.1093 (0.0335)***	-0.1834 (0.0087)***
New Inspector post GC			-0.2377 (0.3559)***	-0.1234 (0.0115)***
# of Inspections	-0.0005 (0.0002)***	-0.0007 (0.0000)***	-0.0003 (0.0001)***	-0.0002 (0.0000)***
Job Tenure	-0.0047 (0.0002)***	-0.0006 (0.0008)	0.0034 (0.0037)***	0.0134 (0.0011)***
Hourly Wage Rate	-0.0312 (0.0002)***	-0.0256 (0.0012)***	-0.0307 (0.0013)***	-0.0301 (0.0027)***
Race (Asian)	-0.0417 (0.0299)	-0.0405 (0.2393)	-0.0752 (0.0755)***	-0.0907 (0.0163)***
Race (Hispanic)	-0.3384 (0.0141)***	-0.3433 (0.0371)***	0.0869 (0.1303)	-0.0389 (0.0193)***
Female <sup>1</sup>	0.3318 (0.0088)***	0.304 (0.0213)***	0.4699 (0.054)***	0.1408 (0.0119)***
Male <sup>1</sup>	0.0974 (0.004)***	-0.0103 (0.0032)***	-0.1023 (0.036)***	-0.0707 (0.0106)***
Most Visited Division	0.3722 (0.0069)***	0.7991 (0.0289)***	0.1776 (0.0379)***	0.4774 (0.0151)***
Division2	-0.0339 (0.0041)***	-0.2326 (0.0169)***	-0.0313 (0.0489)	-0.229 (0.0194)***
Division3	0.1829 (0.0084)***	-0.1686 (0.0143)	0.4262 (0.0596)***	0.0021 (0.018)
Division4	0.5327 (0.0143)***	-0.069 (0.0113)***	0.426 (0.0777)***	0.2949 (0.025)***
Division6	0.2171 (0.0074)***	-0.5925 (0.0307)***	0.0795 (0.0498)	0.2742 (0.0126)***
Division7	0.0006 (0.001)	0.0011 (0.0026)	0.0007 (0.0419)	0.0007 (0.0112)
Division8	0.4926 (0.013)***	0.3602 (0.0327)***	0.2523 (0.0704)***	0.0247 (0.0111)***

<sup>1</sup> The default set includes inspectors whose race are not recorded.

<sup>2</sup> Division 9 is used as the default regime.

Table 4: **Main Model Estimation Results (Parameters for Inspection Type)**

	<b>Regime1</b>	<b>Regime2</b>	<b>Regime4</b>	<b>Regime5</b>
<i>Effect of Inspection Type on Restaurant Negative Hygiene Score</i>				
Consumer Complaint	7.2162 (0.3297)***	6.6227 (0.4291)***	5.9525 (0.3965)***	3.3037 (0.1429)***
Food Poisoning				-0.4549 (0.7506)
Owner Initiated				-0.6015 (0.105)***
DHS Initiated				-0.0049 (0.1387)

Table 5: Main Model Estimation Results (Parameters for Inspector Bump Up Probability)

	Regime2	Regime4	Regime5
<b><i>Bump from C to B</i></b>			
Prob(bumpfrom78)	<i>(constrained to 0)</i>	0.3505 (0.0417)***	0.6841 (0.0029)***
Prob(bumpfrom79)	<i>(constrained to 0)</i>	0.6414 (0.0527)***	0.8175 (0.0024)***
Prob(bumppto80)	<i>(constrained to 0)</i>	0.515 (0.0315)***	0.9475 (0.0331)***
Prob(bumppto81)	<i>(constrained to 0)</i>	0.227 (0.0115)***	0.0324 (0.0152)***
Prob(bumppto82)	<i>(constrained to 0)</i>	0.2579 (0.2185)	0.0125 (0.0226)
Prob(bumppto83)	<i>(constrained to 0)</i>	<i>(constrained to 0)</i>	0.0077 (0.0429)
<b><i>Bump from B to A</i></b>			
Prob(bumpfrom85)	<i>(constrained to 0)</i>	<i>(constrained to 0)</i>	0.3551 (0.0026)***
Prob(bumpfrom86)	<i>(constrained to 0)</i>	<i>(constrained to 0)</i>	0.5083 (0.0039)***
Prob(bumpfrom87)	<i>(constrained to 0)</i>	0.3612 (0.042)***	0.6473 (0.0017)***
Prob(bumpfrom88)	<i>(constrained to 0)</i>	0.3849 (0.0345)***	0.7888 (0.0019)***
Prob(bumpfrom89)	<i>(constrained to 0)</i>	0.6068 (0.0200)***	0.8915 (0.0006)***
Prob(bumppto90)	<i>(constrained to 0)</i>	0.6022 (0.0102)***	0.6907 (0.1422)***
Prob(bumppto91)	<i>(constrained to 0)</i>	0.3978 (0.0102)***	0.1943 (0.0593)***
Prob(bumppto92)	<i>(constrained to 0)</i>	<i>(constrained to 0)</i>	0.0698 (0.1195)
Prob(bumppto93)	<i>(constrained to 0)</i>	<i>(constrained to 0)</i>	0.0452 0.1950
<b><i>Bump to 100</i></b>			
Prob(bumpfrom97)	0.2937 (0.0294)***	<i>(constrained to 0)</i>	<i>(constrained to 0)</i>
Prob(bumpfrom98)	0.5649 (0.0374)***	0.0829 (0.6601)	0.0000 (0.0159)
Prob(bumpfrom99)	0.8652 (0.0102)***	0.5004 (0.0127)***	0.5006 (0.0069)***

Figure1. Score Distribution in Data

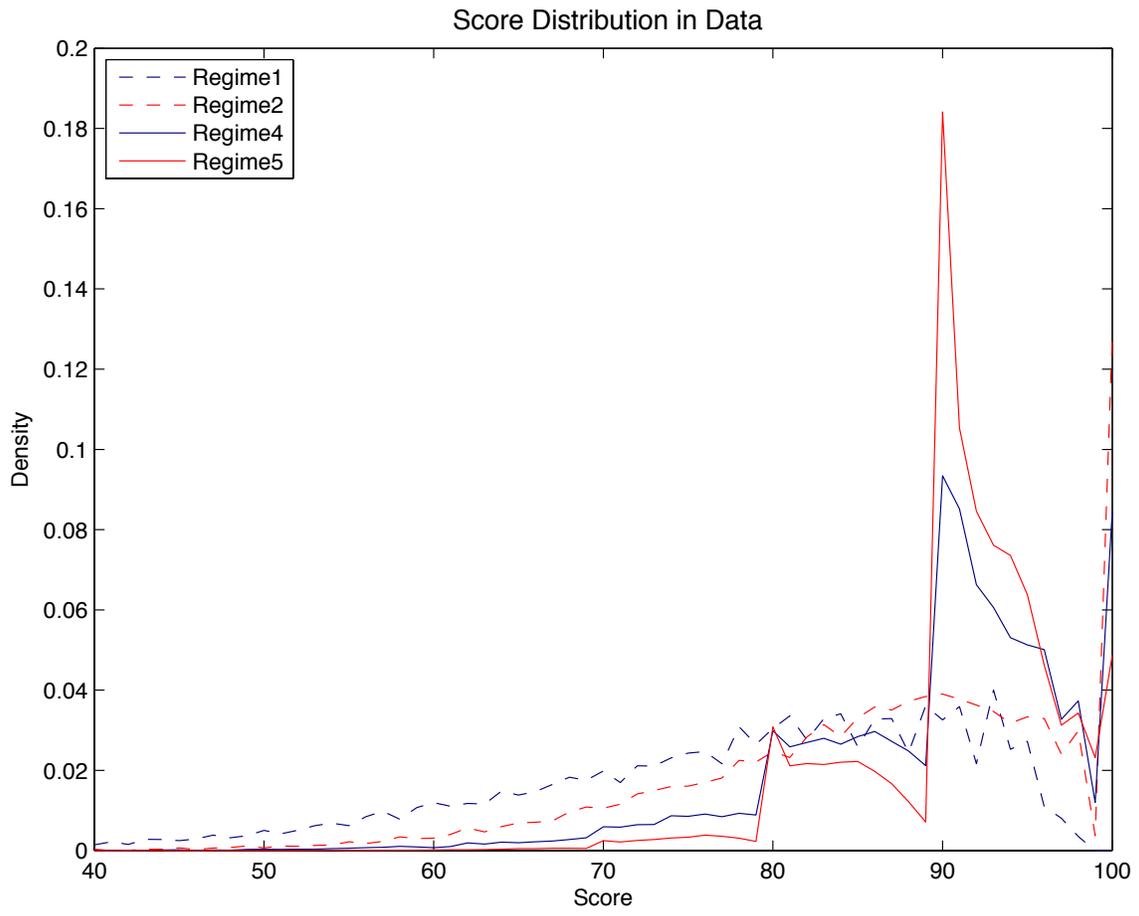


Figure2. Simulated and Data Score Distribution by Regime

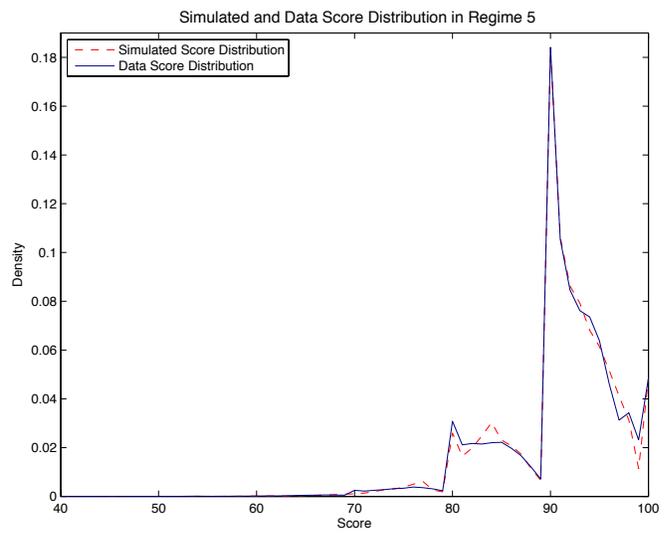
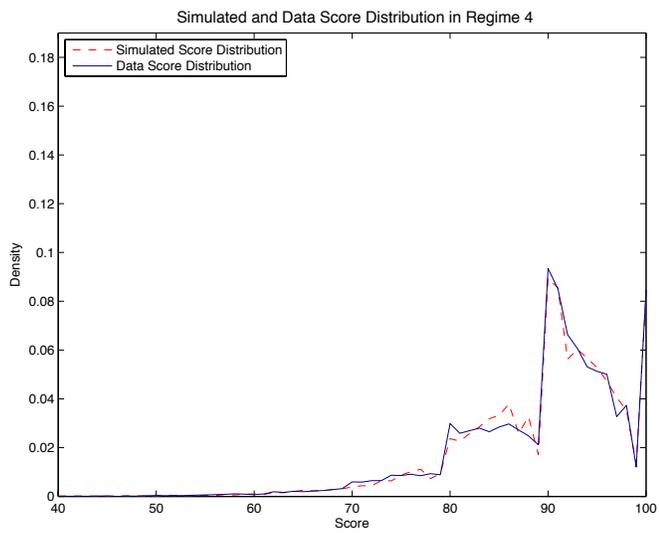
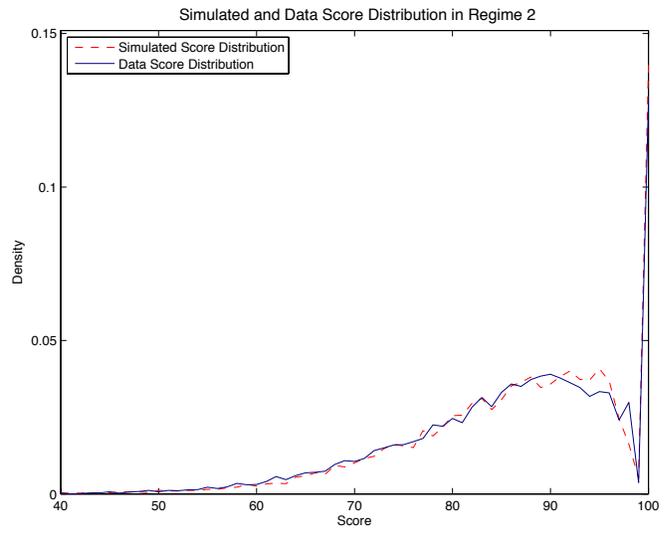
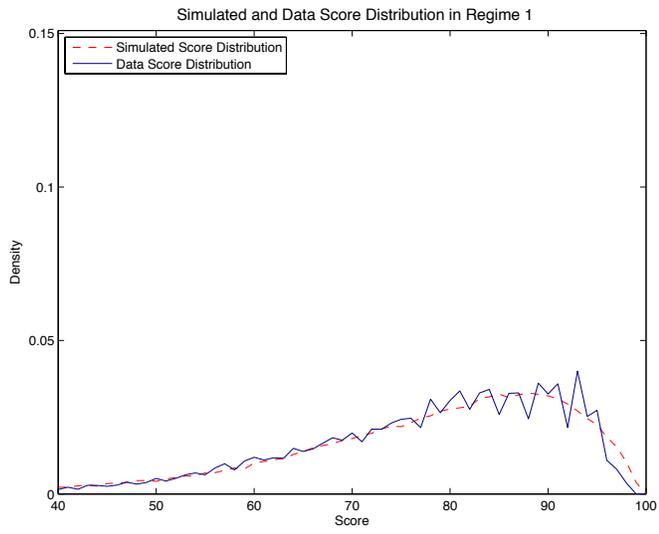


Figure 3. Simulated Detection Probability and Probability Index Function  
(Varying inspector characteristics set fixed at the total sample mean)

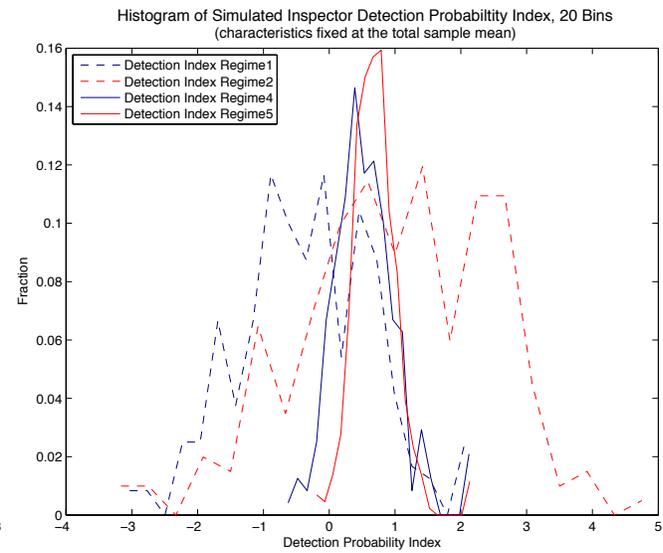
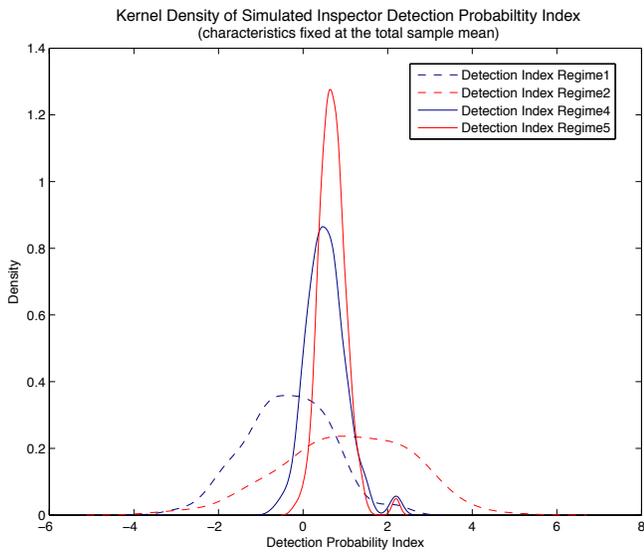
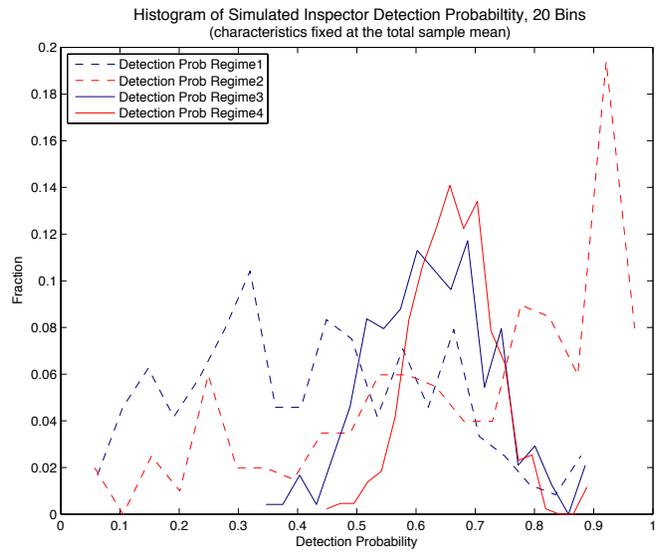
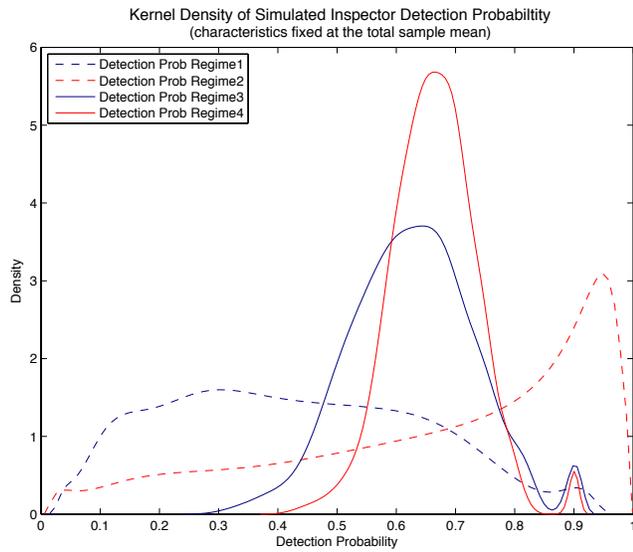


Figure 4. Simulated Detection Probability and Probability Index Function  
(Varying inspector characteristics set fixed at the total sample mean)

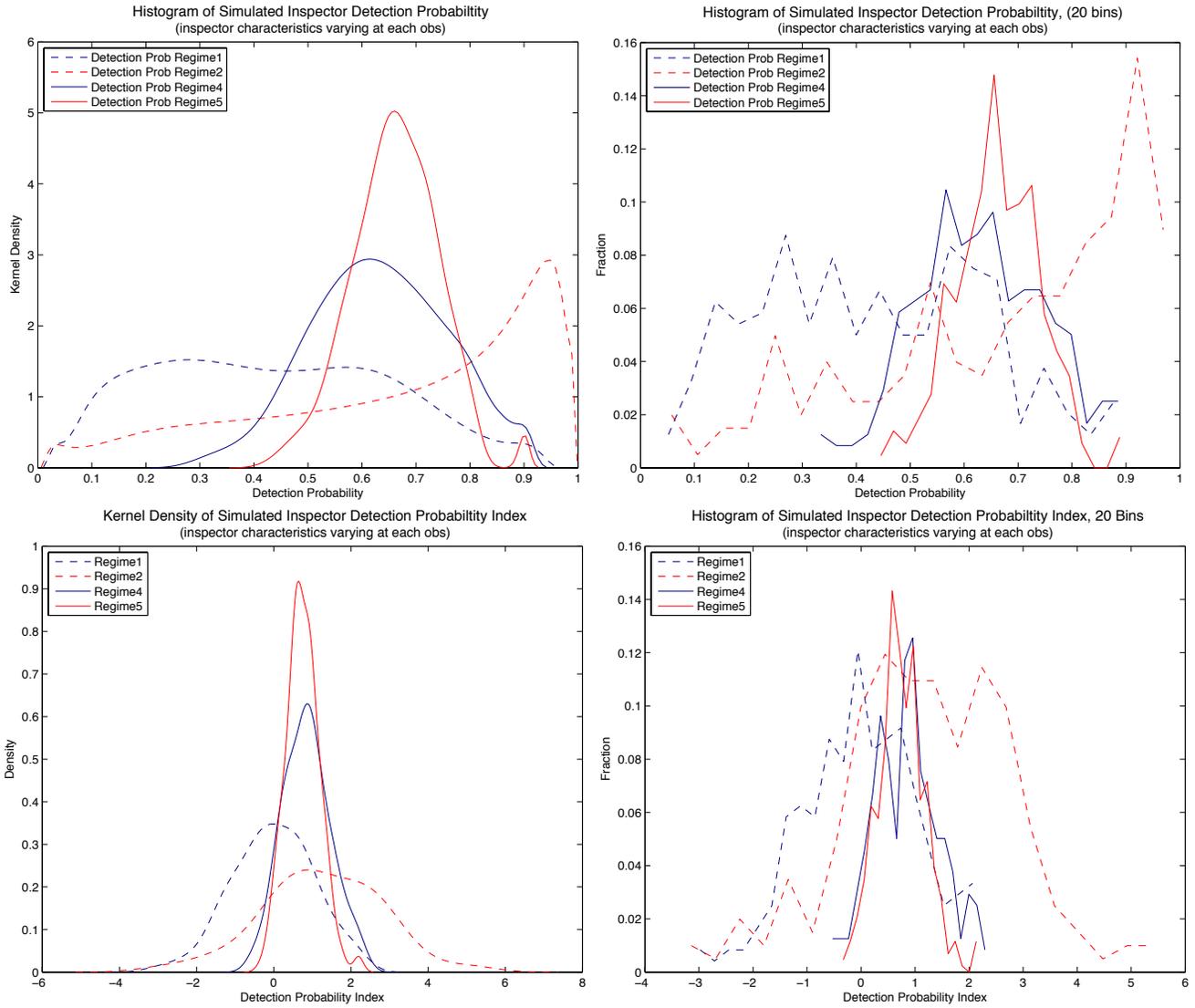


Figure 5. Simulated Restaurant Hygiene Distribution

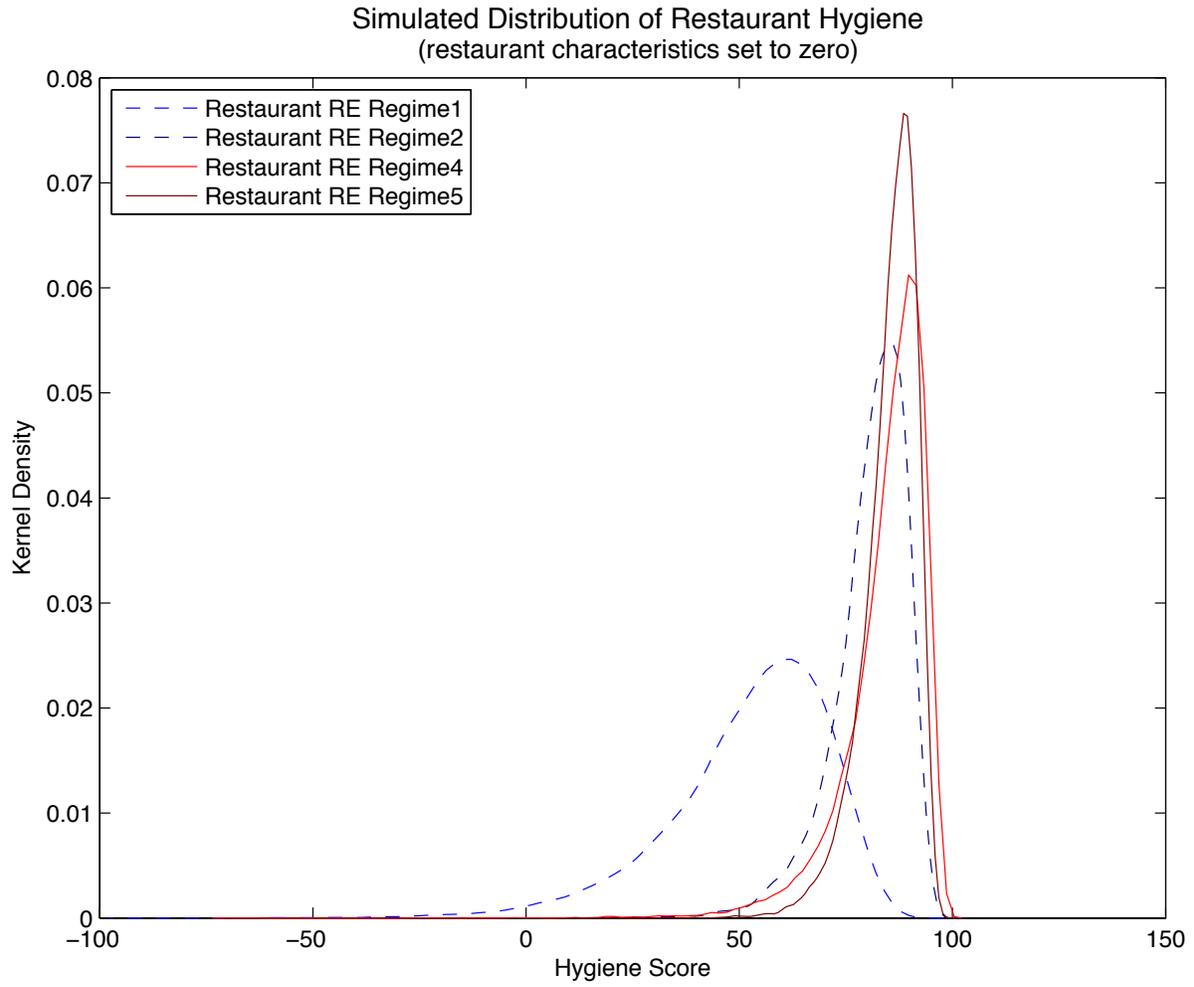


Figure 6. Simulated Detection by Gender  
(Inspector level detection probability averages across inspections of each inspector)

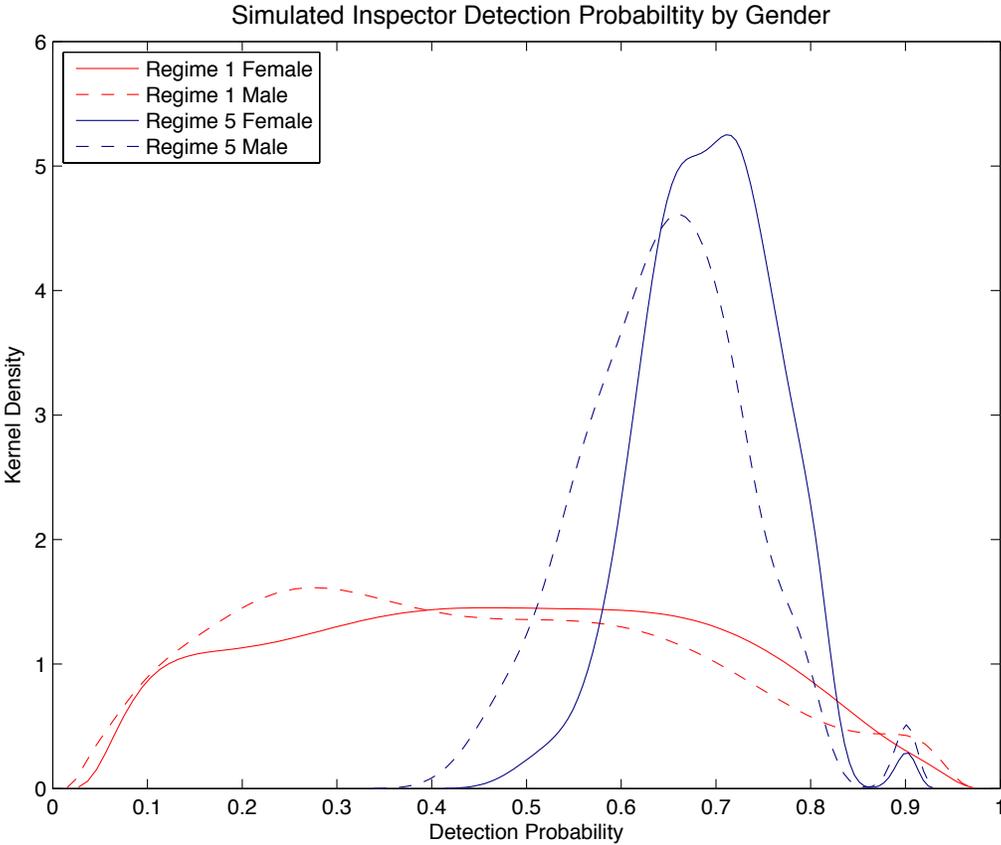


Figure 7. Simulated Detection by New Inspector after Grade Card  
(Inspector level detection probability averages across inspections of each inspector)

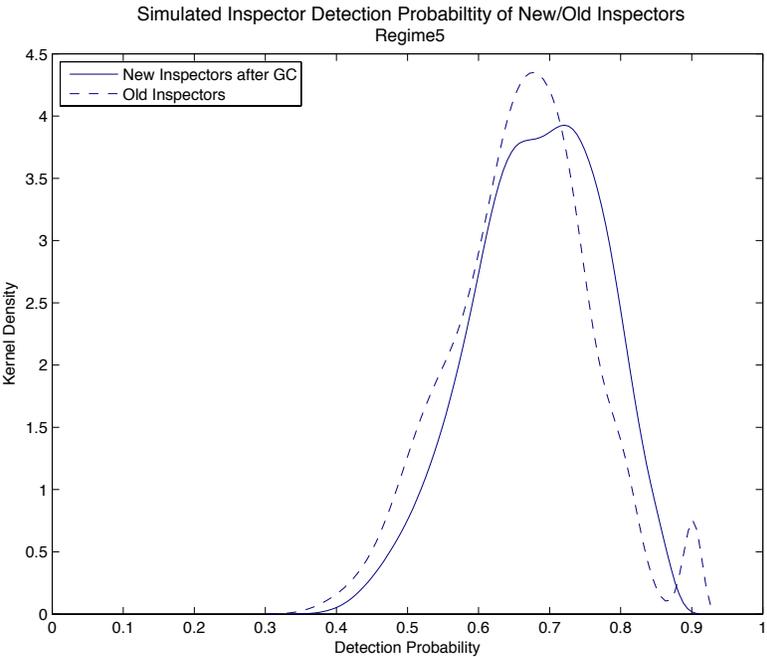
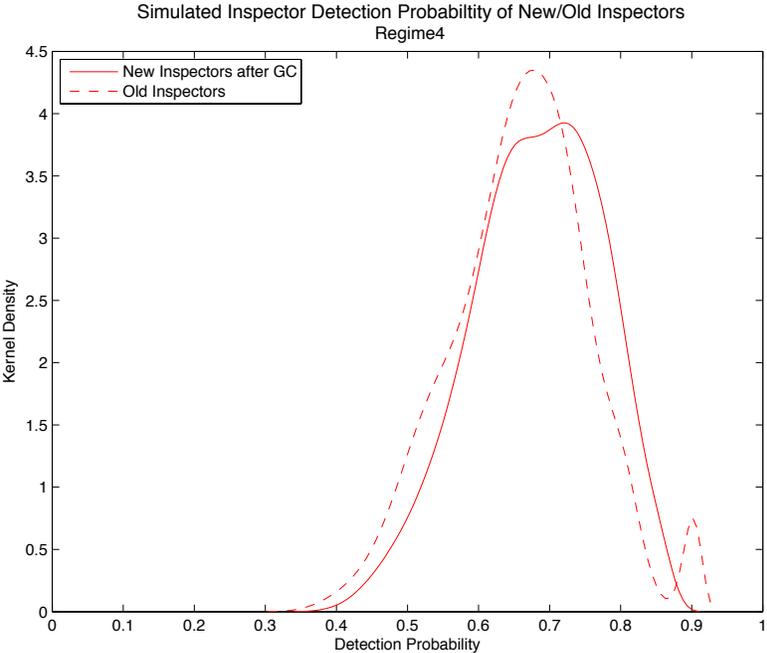


Figure 8. Counterfactual Score Distribution and Detection Probability: I  
(Every inspection is a non-repeated inspection)

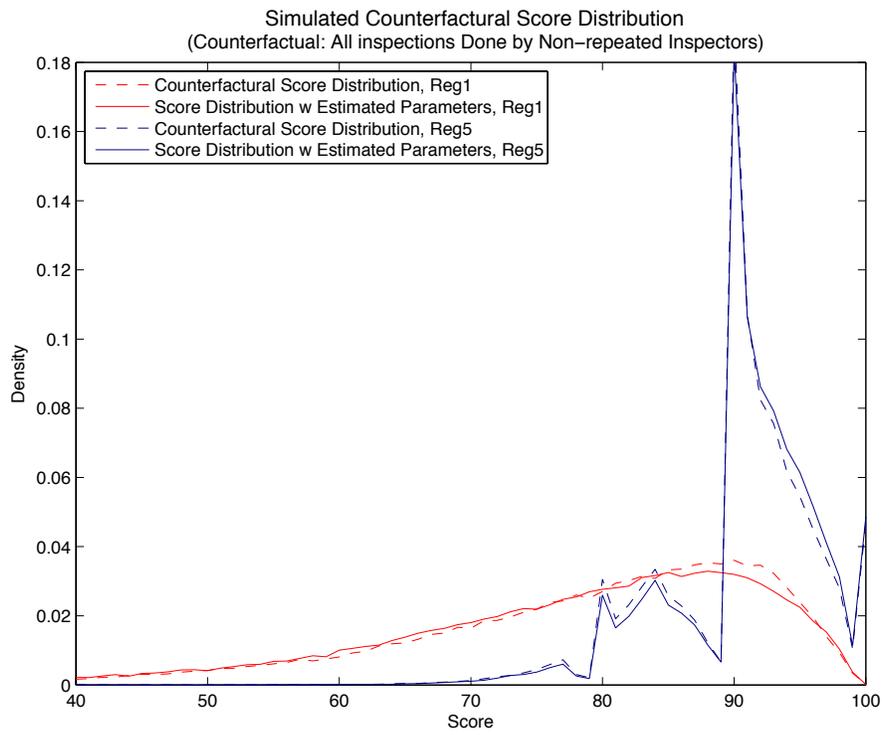
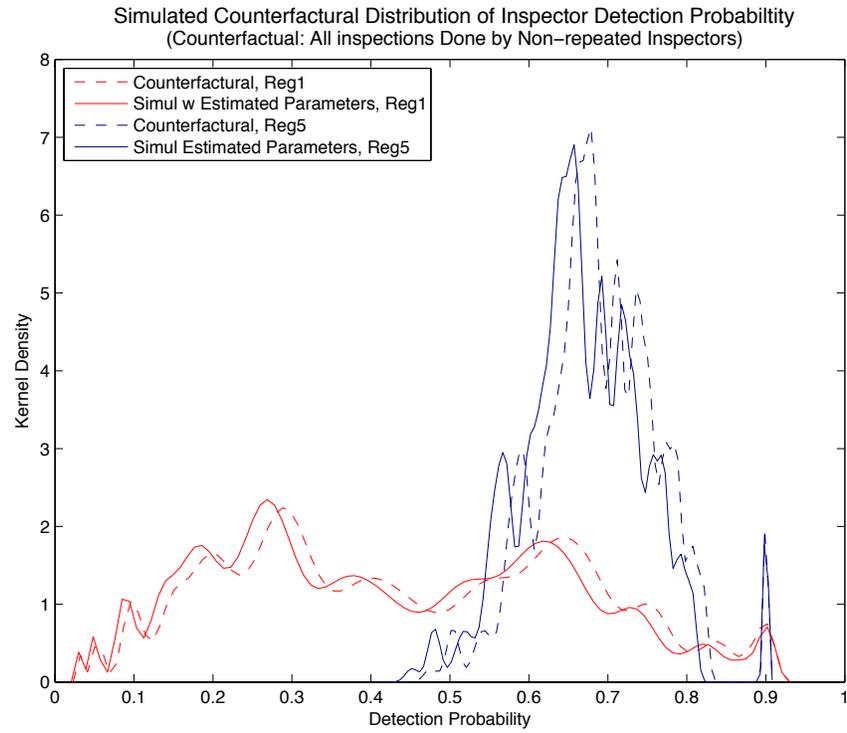


Figure 9. Counterfactual Score Distribution II  
(All Inspector Detection Probability Fixed at 0.9)

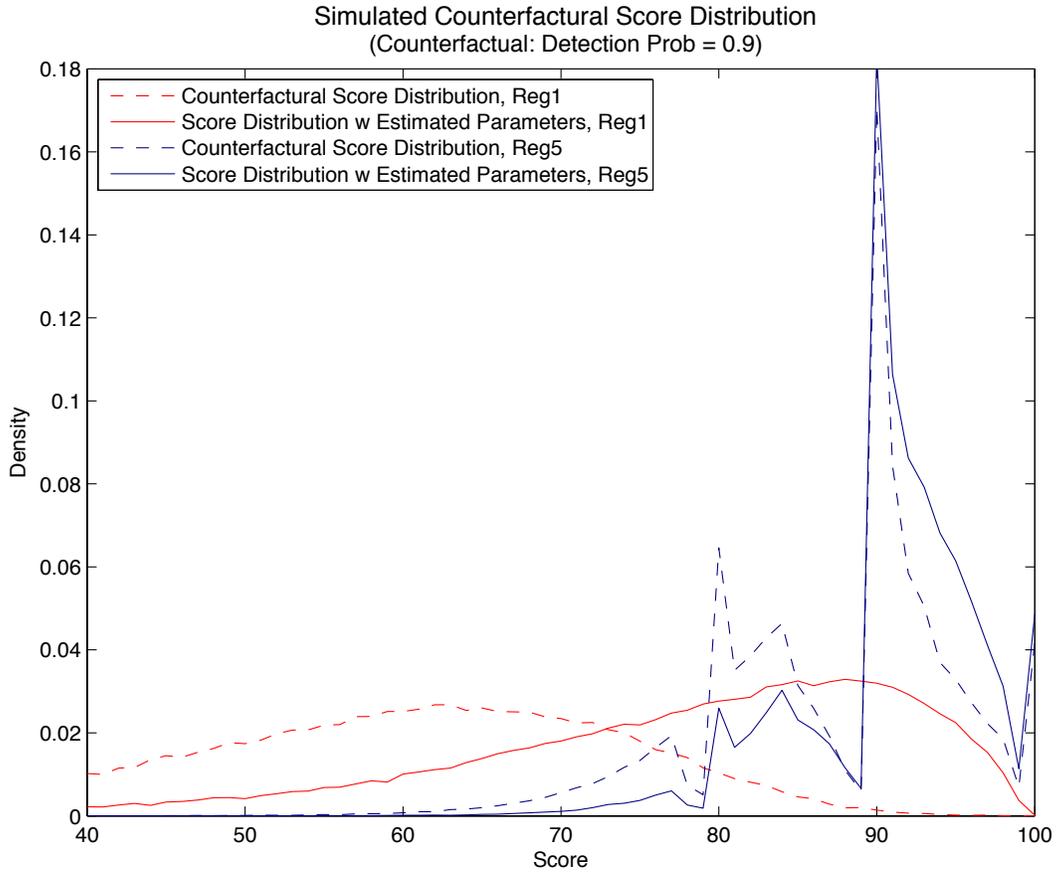


Figure 10. Counterfactual Score Distribution III  
(Inspector RE Variance Fixed at 0.1)

