EXPLANATION OF THE STATISTICAL MODEL for STANDARD 4

This is an explanation of the thinking that determined the statistical model relating to the criteria used for evaluating the inspectional performance of jurisdictions.

Evaluation of the performance of large jurisdictions

For large jurisdictions (jurisdictions with 10 or more inspectors), the evaluation is based on direct oversight of two inspections per inspector, with respect to 10 items of performance. If 10 or more inspectors are being evaluated in the program, then we will see 20 or more scores of satisfactory or unsatisfactory for each item. The standard for approval of the inspection performance is a passing score of 75% on each of the 10 items. An individual item receives a passing score if at least 75 percent of the instances of observation are completed in a satisfactory manner. For example, with 10 inspectors, we must have at least 15 (that is 75 percent of 20 inspections) completed correctly for item number 1. Similarly, for item number 2, we would need to see at least 15 inspections done correctly. In order for the program to pass the evaluation successfully with respect to inspection performance, all of the 10 items would be required to show satisfactory completion of at least 15 out of the 20 ratings. For those jurisdictions with more than 10 inspectors, we simply apply the 75 percent rule as we did for the jurisdiction with 10 inspectors. Using two overseen inspections for each inspector, record the observations for each item, figure the percent correct for each item, and round up to the next higher whole number when the percent is not a whole number.

The 75 percent per item rule was determined by the consensus of several highly experienced individuals working in the retail food safety team. We view the set of overseen inspections as a sample from a much larger set of total inspections performed. In this approach to program evaluation, the statistical measure does not evaluate any individual inspector. The emphasis is on the overall performance of the team, with respect to any item. Even if an inspection is observed in which one inspector fails all 10 items, the program would not necessarily fail.

The jurisdiction's quality assurance program, however, must address individual inspector's performance to ensure a standard of uniformity among the team. If each inspection were successful only 75 percent of the time for each item, the team as a whole would almost always fail. This is because they would almost always dip below 75 percent on at least one of the 10 items. For example, a team that scored 70, 70, 70, 75, 75, 75, 75, 80, 80, and 80 on each of the 10 items would be successful 75 percent of the time, but they would fail three times over since three items scored below 75. However, for a team with 10 inspectors exactly, if their chance of getting each item right improved to 88 percent at each inspection, then they would have a much better chance of keeping all 10 results at 75 percent or higher. Under the simple statistical assumption of independent sampling, a team achieving 88 percent at each inspection would pass the evaluation 75 percent of the time. Therefore, this 88 percent level of performance was used as a simple representation of a team that is good enough that we want them to have a

good chance of passing, but not so good that they would not find it advantageous to improve.

Evaluation of performance of small jurisdictions

A statistical issue was to determine a reasonable standard for those jurisdictions with less than 10 inspectors. When the sample gets this small, the relative error in the estimated fractions gets so large that the "each of 10 items rule" will fail good programs too frequently. Therefore, the 88 percent level of performance at each inspection was the feature of the standard that was kept constant in designing the sample sizes for the smaller jurisdictions

In jurisdictions with less than 10 inspectors, the statistical solution is to group all of the individual ratings, disregarding the individual items. For 5 inspectors we would review 5 x 2 = 10 inspections, with respect to all 10 items combined. This gives 100 observations. It is not possible to make a total observation test mimic exactly a 10 item test, but the minimum passing rates will be about as stringent as the 75 percent for each of 10 aspects test:

For 4 to 9 inspectors, conduct two joint inspections for each inspector. Chart 4-1 shows the lowest total passing score out of the complete set of combined items that would give at least a 75 percent chance of passing for a team with an 88 percent chance of getting any particular observation correct. For a team of three or less, it is recommended that extra oversight inspections be performed to produce a total of 8 inspections. This is an intuitive judgment call that any set smaller than 8 could randomly turn out to be odd enough to produce an unfair rating.

Chart 4-1
Method of Calculation for Jurisdictions with Less Than Ten Inspectors

# of inspectors	# inspections needed	# of items needed to be marked IN compliance in order to meet Standard 4 criteria
<4	8minimum	**65** (out of 80 possible Items)
4-9	2 per inspector	4 inspectors = **65** (out of 80 possible Items) 5 inspectors = **82** (out of 100 possible Items) 6 inspectors = **99** (out of 120 possible Items) 7 inspectors = **116** (out of 140 possible Items) 8 inspectors = **133** (out of 160 possible Items) 9 inspectors = **150** (out of 180 possible Items)