Response to Peer Review Comments

for

The FSIS Risk Assessment for Risk-based Verification Sampling of *Listeria monocytogenes*

Introduction

Below are itemized responses to each of the peer review comments for the FSIS Risk Assessment for Risk-based Verification Sampling of *Listeria monocytogenes*. The risk assessment report was updated based on peer review comments and is available at the following site: http://www.fsis.usda.gov/Science/Risk_Assessments/index.asp.

Reviewer 1

Comment: The risk ranking algorithm is strongly geared towards protection of establishments that comply with the Rule, i.e., adopt a more stringent Alternative. That is achieved by application of Alternative specific weights to adjust risk related to the history of positive results (Risk3) and history of negative results (Risk4). Weights to adjust Risk4 seem to have such a strong influence on risk score rank that other risk factors may become precluded.

Reply: Sensitivity analysis using standardized regression with orthonormal independent variables on the baseline risk and the final risk rank shows that deli meat volume has the greatest influence on the baseline risk ranking (70.73% of variance explained by regression) followed by hot dog volume (4.12%) and then by other products volume (0.02%). In the overall risk ranking algorithm the baseline risk ranking has the greatest influence on the final establishment risk ranking. The percentage variation explained by regression attributed to the baseline risk in the overall risk ranking algorithm is 97.93%, which far outweighs historical risk factors in determining the final establishment risk ranking. In the example data set, Risk 1 accounts for 19.91% of the variation explained by regression, Risk 3 accounts for 3.32%, and Risk 4 accounts for 1.90%. The reason that large calculated historical weights have less than expected effect on the final ranking is that the adjusted baseline ranks can vary widely according to the historical penalty and reward weights however these differences are diminished by re-ranking due to interlacing of the weights among the alternatives and volume production classes according to the weighting scheme. The differences between baseline and final adjusted ranks can be made to vary greatly by increasing the weight reward multiplier; but the relative order of ranks varies little, irrespective of the size of the historical weights. The weights are designed to ensure that plants with positive *L. monocytogenes* cultures are always sampled. Large volume plants making high risk product may not be sampled if they are in a low risk alternative. On the other hand, low risk alternative plants may be sampled if they produce a large volume of high risk RTE product.

Comment: Results of risk based verification sampling so far seem to show that establishments that employ Alternative 3 have the highest prevalence of positive lots. However, establishments that employ Alternative 3 are also the ones that are the most heavily sampled. Therefore, comparison between prevalence of positive RTE lots produced under different Alternatives should be made with caution because there is no statistical support for that.

Reply: This is correct. It will take time to accumulate enough data to conclude that alternative 3 has the highest positive rate. In order to make prevalence comparisons among alternatives, the sample sizes need to be the same for each alternative and of sufficient size for valid statistical comparison.

Comment: While one of the general charges to this peer review was how to decrease uncertainty, to this reviewer it actually seems that the uncertainty was underestimated.

Reply: Uncertainty was re-estimated by comparing the calculated variability for the standardized regression coefficients for the full risk factor model from the data set example of 1,981 data points with the uncertainty associated with the same standardized regression coefficients from a more recent expanded data set of 2,493 data points bootstrapped for 10,000 iterations of sample size 1,981. We feel that the uncertainty estimates are improved in accuracy because of this.

Comment: Also, there are concerns related to averaging performed to derive weights for penalty and reward points and deriving the minimum and maximum for triangular distributions. It is unclear how Risk3 and Risk4 are estimated.

Reply: An improved method of estimating the average distance delta that a past positive establishment would be penalized is taken as the sum of all distances in rank units from baseline to the maximum rank eliminating present positives from establishments with positives in the past 6 months, excluding the present month divided by the number of past positive plants (excluding the present month). For the penalty weight, delta is multiplied by the establishment relative risk and divided by the maximum relative risk. For the reward weight, delta is multiplied by the negative of the establishment relative risk and divided by the maximum relative risk.

Risk3 is estimated from the number of past positives over the previous 6 months, excluding the present month for each establishment. The formula for calculating this risk is:

$$\text{Risk3} = 0.231 \text{ x } [1,0] + 0.205 \text{ x } [1,0] + 0.191 \text{ x } [1,0] + 0.186 \text{ x } [1,0] + 0.185 \text{ x } [1,0]$$

This formula was derived from an autocorrelation study of expected positives by month over the next five months after an initial positive in the first month. The model used was for an alternative 3 plant working two shifts per day over a 30 work month for one year totaling 1,000,000 iterations.

Risk4 is estimated as the sum of negative culture results for each plant over the past six months including the present month. The sum is divided by the maximum total number of samples achieved by any plant during the six month period.

Comment: While it would be very interesting and important to test the likelihood of the proposed sampling scheme of one sample/plant/month to detect L. monocytogenes in a contaminated lot, i.e., to estimate the power of the proposed sampling scheme, that sensitivity analysis was not performed and to this reviewer it seems that the algorithm structure would not allow it.

Reply: The probability of accepting a positive lot assuming a binomial probability distribution used in calculating sampling power is not dependent on the risk ranking algorithm but on the number of samples per lot taken in the identified plant. This probability is:

$$\text{Paccept} = (1\text{-prevalence})^{n}$$

However, if sampling is hypothetically increased from one sample per month incrementally, the probability of estimating the true plant prevalence rate is also increased. Using Monte Carlo sampling of the most recent 2,493 plant alternatives from December 2006 with the June 2005 sample size of 1,981, the prevalence rate was calculated in each alternative with each increasing increment in the number of risk based samples taken starting at a rate of one per plant with a constant sample size of 800. The stopping point was taken to be when the calculated prevalence in each alternative matched the prevalence in the larger data set. This was in the range of 20 samples per plant. This means that a sample size of at least 20 samples per alternative would be needed to say there was a significant difference in the alternative prevalences at p=0.05.

Comment: In this risk assessment, the correlation between the fraction of successive positive samples of RTE products and lagged days would be weaker than reported. Therefore, the history of positive and negative culture results may not really be a good predictor of L. monocytogenes contamination of RTE products. As a consequence, a risk based algorithm may not be able to target problematic establishments. Although they come from both a risk-based sampling program and a random RTE testing program, the first results of all FSIS RTE sampling, year to date, in 2005 may even confirm that. For example, the FSIS 2003 risk assessment predicted 5-7% prevalence of contaminated lots in Alternative 3 (Table 9, Appendix 3). However, the prevalence of contaminated lots among FSIS samples collected in establishments that employ the same Alternative was only 0.3% (Report, page 20).

Reply: The correlation between the fraction of successive positive samples of RTE products and lagged days was reexamined and new weights for Risk3 were developed. The basic difference is that the first two months (months 2 and 3) receive most of the weight (70.62% versus 43.6%). We feel the correlation is significant in the first few months and cannot be neglected. These weights have been incorporated into version 2 of the algorithm. With regard to the prevalence used in the 2003 deli meat model, according to 2006 estimates the 2003 estimates are high. Future versions of the model take this into account. But, with regard to the validity of using the 2003 model estimates to develop the risk ranking algorithm there is no real conflict because the algorithm is based on relative risk rather than absolute risk. That is the primary reason for using ranks rather than the absolute risk estimates.

Comment: The Agency poses very important questions, such as "In the absence of direct verification by USDA, are there any modifications that should be made to the model to take into account this lack of verification? Should the risk rank scores of plants whose data are verified, as opposed to those plants where only self-reported data are available, be adjusted in any way?" To dissipate this problem, the Agency could verify some of the responses and estimate the probability that an establishment reported the true volume of production and adopted Alternative. This probability, if added to the model, would increase the uncertainty of the model results but would represent more accurately the true state of the Agency's knowledge. Adjusting the risk score rank only for establishments whose reported data were verified would not be fair because the verification process is not available to all establishments; only some randomly selected establishments could be verified.

Reply: The Agency seeks to verify all statements made on the FSIS FORM 10,240-1 by plant management as to the expected RTE volumes, establishment risk alternative, log reduction for post-lethality processes, allowable log increases in products with antimicrobial inhibitors added, and establishment sampling rates of product and food contact surfaces. At present, there are place holders in the algorithm for unverified statements given by establishments. In the future verification teams will go into establishments for statement verification, at which time verified and non-verified statements can be compared among establishments and the place holders will become probability estimates.

Comment: A work by Nauta (2005) may help to estimate the number of samples that should be collected in this program to assure that contaminated lots will indeed be detected. Conveniently, the number of samples should be smaller in "high risk" plants compared to "low risk" plants, which will further justify the need to target "risky" plants through the risk ranking verification algorithm. To detect L. monocytogenes contamination in an unevenly contaminated lot of RTE products, more samples should be collected. How many samples should be collected depends on the total number of L. monocytogenes in the lot and clustering of L. monocytogenes.

Reply: We agree that more samples taken will result in more accurate detection of contaminated lots. It has been generally agreed that because the Agency is limited to 800 to 1,000 risk-based Lm samples per month, as many high risk plants as possible should be sampled even though that means that only one sample per plant is possible. Taking two, three, or four samples per plant would mean sampling only one-half, one-third, or one-fourth as many plants. Since the sample rate of one per RTE plant per month samples approximately one-third of the total number of plant alternatives, it may be feasible to divide the monthly sample on a rotating basis to take more than one sample in any given month. This will have to be taken in consideration in light of budgetary and laboratory constraints.

Comment: The risk ranking algorithm is structured in such a way that non-compliance with the Rule seems to be a stronger risk factor than high volume of production (Risk4 dominated over many variables in Risk2), even though the production volume is a confirmed risk factor for listeriosis in humans because of a larger potential of exposure to consumers. Gearing the algorithm towards motivating establishments to enhance control measures and in that way indirectly possibly protecting public health, but not protecting public health from direct risk factors, may trigger criticism of the risk based verification sampling program and the risk ranking algorithm.

Reply: Further sensitivity analysis on new data and data used in the draft report show that Risk2 dominates the reward and penalty variables in the risk ranking algorithm. This means that plant risk alternative, plant production volume, and type of RTE product produced explain nearly 88% of the variability in the final risk ranking. It is true as the risk-based sampling program accumulates more negative samples and positive ones per plant over the six month period these adjustments are allowed, Risk4 tends to accumulate more variability than Risk3. However, the weights for these factors are adjusted monthly so that the statistical effect of both factors is equal. The increasing accumulation of positive and negative cultures is due to increased sampling from the IVT and RLm programs and continued sampling from the ALLRTE program.

Comment: It is not clear how Risk3 and Risk4 are estimated each month.

Reply: The revised report addresses this issue extensively.

Comment: considering the uneven distribution of L. monocytogenes in the lot of RTE products, negative test results have much lower predictive value than conclusive evidence of a positive result. Contrary to that, according to sensitivity analysis, Risk4 has a stronger influence on Risk score rank than Risk3. That seems fundamentally wrong.

Reply: So that past negative results do not unduly influence the final risk ranking over past positive results, the weights for these risk factors are adjusted such that the sums of the positive versus the negative rank deviations are statistically equal. Alternatively, a tornado plot is generated for each dataset and the contribution of Risk3 and Risk4 influence to the final risk rank is equalized by adjusting the constant in the weight of

Risk4. The decision to permit a decrease in risk ranking due to past negative culture results can be justified on the basis that very few plants will attain the full reduction possible since the reward is based on achieving the same total number of negative cultures that the most heavily sampled plant can achieve. Most plants with a few negatives over the six-month period are reduced in rank less than 10 units and the weighting ensures that the total reduction is inversely related to their volume- risk alternative.

Therefore, using negative test results to reward establishments for their compliance seems inappropriate. Instead, if so desired, compliance should be introduced into the model as an additional risk factor. It would, however, be more correct to exclude establishments' compliance from the risk ranking algorithm. After the algorithm identifies "high risk" plants, the Agency could use compliance for final selection of plants to sample.

The risk ranking algorithm uses the baseline risk rank as the core risk for each establishment. The culture history of positive and negative results account for a minor adjustment to the baseline risk overall except in the case of current positive results when the establishment must be re-sampled. Usually, an establishment is sampled repeatedly following a positive result due to weighting of past positive results adding to the present risk ranking. An establishment receives a reduction in risk only when there have been no past positives for six months and can only receive the full reduction in risk if it is in the lowest risk alternative and has the same number of negative cultures as the most sampled plant in the past six months.

Comment: Weights for adjusting Risk3 (W1) and Risk4 (W2) were produced by averaging conducted in two steps. First step averaging smoothes the ranks in such a way that it "punishes" establishments that have rank smaller than average and "rewards" establishments that have rank higher than average. The second stage averaging increases the difference between average ranks of establishments in Alternatives 1 and 2a, and between establishments in Alternatives 2b and 3. On the other hand, it reduces the difference between average ranks of establishments in Alternatives 2a and 2b. In other words it diminishes the influence of Alternatives on the risk rank score. The estimated average differences between risk ranks are then used to derive weights W1 and W2, whose purpose is to differentiate establishments that employ different Alternatives.

Reply: The method of obtaining the maximum weight for adding risk and its complement for reducing risk has been modified. The average change in risk allowed, delta, is calculated as the average of the difference between baseline rank and the maximum rank achievable for each establishment having a positive result in the current month. The final Risk3 weights are obtained by multiplying delta by the standardized prevalence relative risk taken from the dynamic in-plant Lm model at the predicted high, medium, and low converted deli meat volumes for alternatives 1, 2a, 2b, and 3. The weight for Risk4 is obtained by multiplying delta by one minus the standardized prevalence relative risk for each product-volume alternative. The Risk4 weight is multiplied by an adjustment factor to equalize the effect of Risk3 and Risk4 on the final establishment risk ranking.

Comment: Rather than expressing lack of knowledge by fitting triangular distributions in such a way, it may be more appropriate to model parameters K1, K2 and Q80 as point estimates and just test the influence of these parameters on the model output in a subsequent sensitivity analysis.

Reply: This has been done with no disagreement. The sensitivity analysis has been updated such that only the volume elements for deli meat, hot dogs, and other products adjusted for product risk and risk alternative impact Risk2. The empirical volume distributions were used and bootstrapping employed to estimate variability and uncertainty using the original data set and a larger updated data set.

Comment: A small difference between contamination prevalences of sampling schemes (random and risk based) may also be a sign that risk based verification sampling was not really able to target high risk plants. To correct that, changes should be made in the design of the risk based algorithm. Specifically, risk factors should be identified that are predictive of high risk plants (such as high volume production). Adjusting the risk rank score by favoring establishments in Alternative 1 and penalizing establishments in Alternative 3 muddles up risk score ranks.

Reply: Using a perfect sampling scheme with equal numbers of samples taken in each risk alternative, and for random non-risk based samples, this determination may be made. However, the sampling plan used to produce the estimates referred to was not adequate to detect the larger expected prevalences in the high risk categories versus random sampling. Therefore, it is premature to state that the algorithm is not predictive of high risk plants. Additionally, the sensitivity analysis shows that the algorithm is heavily weighted for product type and volume of product produced. This weighting is much more than for regulatory considerations which have been included for policy requirements.

Comment: Report page 11 paragraph 2 line 1: The authors stated "If the establishment was not tested, or was tested but only negative samples were collected in the last 6 months, the establishments' risk score is reduced with reward points." This is in discrepancy with the statement on page 15, paragraph 1, line 13; where the authors stated "Obviously, establishments that have had no positive or negative laboratory results over the previous 6 month period will exhibit no change in their adjusted ranks." So, if the plant was not tested, was that information included in Risk4 or not?

Reply: If an establishment was not tested, there would be no change from its baseline risk ranking unless it had previous positive or negative test results in the past six months. If the establishment had any positive results, its risk would be increased. If it had negative results, its risk would be decreased; but any decreases in risk would be exclusive of any increases in risk.

Comment: Report page 12 Table 1: The authors stated "If the establishment was not tested, or was tested but only negative samples were collected in the last 6 months, the establishments' risk score is reduced with reward points." This is in discrepancy with the statement on page 15, paragraph 1, line 13; where the authors stated "Obviously, establishments that have had no positive or negative laboratory results over the previous 6 month period will exhibit no change in their adjusted ranks." So, if the plant was not tested, was that information included in Risk4 or not?

Reply: If a plant was not tested, there is no information gained. A negative test will add to the denominator of the Risk4 equation, with an added zero in the numerator.

Comment: Report page 18: Table 2 should probably be numbered as Table 4.

Reply: All figures and tables are renumbered.

Comment: Report page 21 paragraph 1 line 7: "A sensitivity analysis is included in Appendix X". This should probably be Appendix IX.

Reply: This error has been corrected.

Comment: Appendix I page 13 paragraph 1 line 2: Table 19 is missing.

Reply: This error has been corrected.

Comment: Appendix V page 39 Sample data column 9: Is it possible that a plant employs more than one Alternative at the same time? For example, franks are produced under Alternative 1, but deli meats under Alternative 3. Could the risk based algorithm account for that?

Reply: Yes, establishments can declare more than one alternative and provide the products and annual product volume estimates for calculating the baseline and final risk for each alternative. A risk calculation is made for each alternative an establishment may have. If the risk ranking is large enough for sampling, only the highest risk alternative is sampled and the others are disregarded.

Comment: Appendix 6 page 41 paragraph 3 line 9: Why could not the same lot of a product be sampled for more than one sample collection project (ALLRTE, RTERISK1, RTE001)?

Reply: This is a budgetary and administrative constraint. No more than one program is allowed to sample an establishment shift in a given work day.

Comment: Appendix VIII page 55 Table 4: What was the rationale of subtracting 7.6 from the penalty factor to derive the reward factor? Wouldn't it be more appropriate to multiple penalty factors with (-1) to get reward factors?

Reply: What was done was to subtract 8.01 from the relative risk to scale the Risk4 weights from -1 to -7.01. The weights are then standardized by dividing by the weight maximum, 7.01. This creates a linear system of weights.

Reviewer 2

Comment: As noted in this FSIS report, although submission of this information is required under FSIS's Interim Final Rule to Control Lm, not all establishments provided complete or accurate forms. Although submission of the Form 10,240-1 was required for an establishment to be eligible for the risk-based sampling program, this FSIS report noted that "information for approximately 350 of the 2200 establishments believed to be operating under the Interim Final Rule had critical data errors or was (sic) missing data entirely", leaving approximately 1850 establishments with complete data free of "critical data errors". However, the example of the risk-ranking for June 2005 presents results for 1981 establishments. It is not clear whether or not data were used from at least some of those establishments found to have critical data errors. Possibly even more importantly, what procedures were used (e.g., data audit of a sub-sample of establishments) to evaluate the overall validity of these self-reported data?

Reply: The 1,981 refers to the number of alternatives filed by establishments. There were only 1,820 establishments in the data set. There were approximately 30 establishments that turned in forms that had incomplete data and could not be used in the analysis. The Agency seeks to verify all information provided by establishments on the FSIS FORM 10,240-1.

Comment: This FSIS report states that "additional contamination data continues to be collected and is incorporated into the risk ranking algorithm on a monthly basis." Although the "additional contamination data" are assumed to be monthly 'updates' of Lm-testing of product samples, it is not clear if "additional" might actually mean other type(s) of contamination data.

Reply: Additional contamination data refers to other program data collecting *Listeria monocytogenes* samples from establishments falling under the Listeria Interim Final Rule. Additional contamination data can also refer to *Listeria monocytogenes* culture data from RTE products and from food contact surfaces and environmental cultures taken in these same establishments. These additional contamination data (or lack of contamination data) will be used as additional risk factors in a future version of the risk ranking algorithm.

Comment: The formula for the "Scaled plant risk score" on page 10 is straight-forward, but it is not clear how "this rescaling … makes the terms for the adjustment based on historical laboratory results easier". In fact, the formulas for the "Penalty points" and the "Reward points" based on historical lab results both lack clarity. The weights of the "Penalty points" formula are presented in Table 3, the details of which are provided in Appendix VII (not Appendix III, as cited in the FSIA report). However, the multiplier of the sum of the weight "Max penalty points" is not defined. Similarly, the "max reward points" is not defined in the formula for the "Reward points", the details of which are provided in Appendix VIII (not Appendix IX, as cited in the FSIA report).

Reply: The formulas have been re-written to reflect a non-normalized scaling. This was done to help the reader understand the risk factor and weight derivations, although final calculations are always carried out in a normalized scale for all risk factors and the risk ranks. The appendix reference has been changed to appendix VIII. Appendix VIII contains newly worded descriptions for the definitions of the penalty and reward weights.

Comment: However, the elements of the various equations on pages 11 and 12 are neither defined nor explained. In particular, is "Risk(2)$_{\text{rank}}$" the rank of the "Plant risk score" (scaled or not)? Is "Risk 3" the sum of the weights used in the "Penalty points", in which case the various multipliers are the max penalty points? Is "Risk4" the "#actual negatives/#possible tests" used in the "reward points, in which case those multipliers are the "max reward points? Obviously, the "illustration" needs considerably more explanation to be useful to the typical reader.

Reply: In the final calculation of risk ranking, all variables and weights are scaled such that risk factors range from zero to unity. Weights are also in the range from zero to one. The maximum penalty points is the average number of ranks from a positive plant's baseline rank to the maximum rank in months 2 through 6. Risk3 was derived from an assessment of estimated time-lagged contamination events from which five weights were calculated that sum to 1. These weights are multiplied by 0 or 1 depending on a positive isolation of *Listeria monocytogenes* from product. This sum of weights is Risk3, which in turn is multiplied by the risk factor weight for Risk3, an entirely different weight. Risk4 is the sum of negative culture results divided by the total cultures possible over six months.

Comment: Table 1 on page 12 (shouldn't this be Table 4?) might be more 'readable' in tabular form, thereby conveniently showing both marginal totals; for example,….

Volume Produced	Alternative				
by Establishment	1	2a(PP)	2b(GI)	3	Total
High	13	13	133	55	214
Medium	62	21	176	497	756
Low	43	23	88	857	1011
Total	118	57	397	1409	1981

Reply: Good suggestion. The table has been updated.

Comment: Table 2 on page 18 would seem to be Table 5 in the report. As noted in the FSIS report, of the 800 samples collected in June 2005, only 0.9% were from Alternative 1 establishments, which made up 6.0% of all 1981 establishments. Conversely, Alternative 3 establishments, which made up 71.1% of the 1981 establishments, accounted for 76.5% of the 800 samples for Lm. Possibly, a table that shows the number or percent of Lm samples by both volume of production (high, medium, low) and Alternative for June 2005 would be informative.

Reply: We changed the table to include percentages.

Comment: The heavy use of calibration makes interpretation of the model difficult, particularly in a short review period. It also leads to communication challenges in providing context for the total uncertainty in the estimated risk reductions presented.

Reply: We have better addressed uncertainty and concentrated on clarity of presentation in the revised draft.

Reviewer 3

Comment: First of all, the authors to some extent failed to present the structure of the model in sufficient detail for a thorough critical review. This means that an excessive amount of time was spent trying to figure out exactly what was done and how the model was constructed. For example, the reader frequently is referred to appendices for further detail about the equations and assumptions provided in the body of the main report, and hence, the main report is not prepared as a stand alone document with sufficient clarity. There are many instances in which the reviewer must refer to appendices in order to understand basic equations given in the report, or even for the definitions of key parameters of those equations, which is a burdensome task. One example is given in Appendix VII with respect to the Calculation of Weighting Factors for Historical Microbiological Results. This section actually helps a lot in understanding the general methodology used by the authors; however, this information should have been provided, in some detail, in the final Report. A better approach would be to make the structure of the model clear in the documentation with further illustrative examples, in the Report, of step-by-step execution of the model.

Reply: We have attempted to make the structure of the risk ranking algorithm clearer in the revised draft. The *Listeria* deli meat risk model is detailed in the references. It is not described in detail in this report.

Comment: lack of clarity about the model structure in terms of deterministic versus probabilistic framework. The order in which the materials and results were presented in the Report implies that the model works with some deterministic values of the parameters in order to estimate the unadjusted value of the risk-based score associated with each food establishment (see Figures 1 through 9 of Report as examples; no variability and/or uncertainty in model parameters are described in relationship to these figures).

Reply: In this draft we have separated variability and uncertainty in the risk ranking algorithm to clarify the roles of these components in the final risk ranking estimates.

Comment: Later, uncertainty and variability analysis are discussed on Page 20 of the Report (see section entitled Conclusions from the Initial Phase of Risk-Based Verification Sampling). However, it is not clear whether the results previously presented have already incorporated these concepts, and hence, they are outputs of a probabilistic simulation of the model or as stated above, are based on point estimates of the model parameters.

Reply: We have clearly outlined that Risk2 is made up of deterministic constants taken from the risk assessments cited. Variability is addressed by using the empirical distributions for deli meat, hot dogs, and other RTE products for Risk2 calculations. The empirical distributions for Risk1, Risk3, and Risk4 are also used to estimate variability in these risk factors. Components of the weights are deterministic constants taken from the risk assessments cited and the weights for Risk3 and Risk4 depend on a sample estimate of maximum allowable penalty. The weight for Risk4 is adjusted with a sample derived constant to equalize the effect of Risk3 and Risk4. All weights and risk factors are normalized by dividing by their maximum values. Uncertainty is distinguished from variability by generating uncertainty distributions for input risk distributions and the output risk distribution using bootstrapping to simulate the uncertainty components. The combined results are presented in revised Appendix IX.

Comment: Plant size designations of large, small, and very small appear in Appendix III, while these are listed as large, medium and small in the Report.

Reply: This has been corrected. Large, medium, and small refer to production volume and have been changed to high (H), medium (M), and low (L) so as to be distinguished from Large, Small, and Very Small HACCP plant size.

Comment: “raw risk score,” “plant risk score,” “min score,” and “max score,” all seem to refer at some level to a raw (non-scaled) score, but this is not entirely clear.

Reply: The “raw risk score” and “plant risk score” are now all referred to as the baseline risk score, which is the same as Risk2. The formula referred to has been changed such that the baseline risk score is normalized by dividing with the maximum baseline risk score.

Comment: Different terms in equations given on Page 11 should be defined. The reader should not need to refer to appendices for understanding the meaning of each term in given equations.

Reply: We have attempted to make these definitions clear in the body of the report.

Comment: The variable designations for the formula for calculating “Score” cited in Appendix VIII, page 53, do not match those used for basically the identical formula provided on Page 8 of the Report.

Reply: We have slightly changed the definition of “score” to be equivalent to the normalized baseline risk score or Risk2.

Comment: Most figures and tables, in both the Report and the associated appendices, need better narration in the text, along with more detailed legends and footnotes to facilitate reader interpretation. Table 4 in the Report is just one example; all of the figures and tables in Appendix VIII could use better documentation as well. The Report does not even refer to Figures 8 and 9 in the text.

Reply: The references to Figures 8 and 9 have been added and we have added notation to figures and tables for better understanding.

Comment: the Report states the reader should consult Appendix III and Appendix IX for details on how the Penalty and Reward points were calculated. This is not correct; Appendices VII and VIII actually describe these calculations.

Reply: This has been changed.

Comment: on page 6, last paragraph, the authors state that Appendix II details the modifications made to the In-Plant model version 1 to the current version 2, upon which the Risk Ranking Algorithm is built. Although Appendix II does describe the In-Plant model, a discussion of the modifications made to the In-Plant model is provided in Appendix III, not Appendix II.

Reply: This has been changed.

Comment: In terms of transparent display of data sources and values in the Report (pages 4-7 of the Report), only the deterministic values of the model parameters have been appropriately defined, and the values assigned to each justified according to logic and data availability. However, when parameters were probabilistic in nature (as described in Appendix VIII), the authors failed to provide sufficient information regarding the parameters of the distributions and the rationale for selection of such distributions. It is suggested to the authors that they tabularize the information regarding all model parameters with sufficient detail, including data source and justification.

Reply: This fault has been corrected in detail in Appendix VIII.

Comment: Justification for and determination of the “scaled plant risk score” is not well described. For example, how are the minimum and maximum scores calculated?

Reply: This has been described above.

Comment: Although the purpose of the Risk Ranking Algorithm is to rank processing plants according to their risk of producing L. monocytogenes positive products, and to use this ranking to allocate sampling resources based on risk, the objectives of the Algorithm are not entirely clear.

Reply: The objectives of the risk ranking algorithm are: 1) to rank RTE establishments falling under 9CFR430 according to *L. monocytogenes* risk of producing a lot of contaminated product at retail; 2) the algorithm will rank all alternatives within and among establishments but only the highest ranked alternative will be used for sampling any establishment; 3) to provide a basis for sample allocation that can be extended to more than one sample per establishment based on *L. monocytogenes* risk; 4) to provide regulatory incentive for RTE producers to adopt sanitation best practices that will place their establishment in the lowest risk alternatives; and 5) to provide a link between establishment risk ranking and the probable relationship of increased public health risk through a proportional number of listeriosis cases in the absence of increased *L. monocytogenes* risk-based inspection

Comment: the Report jumps between individual plant scores or ranks, to comparative ranks. For example, is the Algorithm scale designed to compare ranks WITHIN Alternatives or BETWEEN Alternatives? It appears that the answer is both (using raw plant scores for the former, and adjusted plant scores for the latter), but when and how FSIS might be interested in these two different ranking approaches should be specified.

Reply: The risk ranking algorithm ranks establishment alternatives. Establishments with one alternative and establishments with multiple alternatives are compared on the same ranking scale. Only the highest risk ranked alternative is used for sampling in any given establishment.

Comment: I am not sure why, when, or even if the scaled plant risk score is used.

Reply: The normalized baseline risk score, or "scaled plant risk score," is used as Risk2, which is ranked to give the baseline risk score rank, which is adjusted with historical risk factors: Risk1, Risk3, and Risk4.

Comment: While the calculation of reward and penalty points is straightforward enough, the Report makes virtually no mention of the "weighting factors" for the reward and penalty points. In point of fact, the calculation of these weighting factors is described on Pages 54-56 of Appendix VIII and is confusing at best. It appears that retail prevalence L. monocytogenes estimates are scaled from Alternative 1 to calculate a "relative risk" for each of the other, riskier Alternatives. These relative risk factors describe "penalty factors" (scaled from 1 to 6.6) or "reward factors" (scaled from -1 to -6.6). The penalty weights for each alternative are found by multiplying the relative risk of each alternative by an "average difference" in ranks between the four alternative categories taken in order of risk. In a similar manner, the reward weights are calculated by multiplying the derived factor times the average difference in ranks. That said, there is NO JUSTIFICATION for the scale used for the reward and penalty factors, nor for the way that the calculations were done.

Reply: The scales are from 1 to 7.01 and from -7.01 to -1. The justification is that there is a penalty weight equal to unity given to alternative 1 low volume with increasing value of the weight for each alternative volume level in proportion to the prevalence relative risk. Based on the binomial theorem, there is a proportional relationship between *L. monocytogenes* prevalence and the number of positive samples obtained, which is in accord with the expectation of the *L. monocytogenes* risk model cited in the report. The scale used sets the maximum number of ranks an establishment can be penalized as the average number of ranks from a positive plant's baseline risk score rank and the maximum rank over the six month observation period. The calculations are described in the report.

Comment: In addition, how is "average rank" calculated (see Table 4, page 55 of Appendix VIII)? Why is the "average difference" used in the calculation? Without additional information on the basis for these calculations, it is impossible to comment further on the validity of the overall Risk Ranking Algorithm.

Reply: An improved method of estimating the average distance delta, or the maximum penalty, that a past positive establishment would be penalized is taken as the sum of all distances in rank units from baseline to the maximum rank eliminating present positives from establishments with positives in the past 6 months excluding the present month divided by the number of past positive plants (excluding the present month). For the penalty weight, delta is multiplied by the establishment relative risk and divided by the maximum relative risk. For the reward weight, delta is multiplied by the negative of the establishment relative risk and divided by the maximum relative risk.

Comment: The role of the variable Risk1 is also confusing. Part of this is a wording issue, with respect to the term "current," but this may impact the validity of the Algorithm calculations as well. We know that the Algorithm is attempting to identify those food establishments that have relatively higher risk with respect to L. monocytogenes contamination and recommend them for future sampling. However, it is not clear whether they will be recommended for sampling for the current month based on the history of contamination in the last six months, or for the next month based on the testing results for the current month and the previous six months? It appears that it is the latter, and in the overall risk rank equation, the input Risk1 is designated as the number of L. monocytogenes positive cultures for the "current" month. It is not clear if we already have the testing results for the current month for each specific food establishment, how we will use the output of the risk-based sampling model to decide whether we should test that establishment or not. It is unlikely that the objective is to suggest a sampling plan for the next month because it is mentioned on Page 11 of the Report that if an establishment has a history of positive testing in the previous month, it will be automatically selected for sampling in the current month and it will not be ranked based on the risk-based sampling model. This is extremely confusing and calls into question the role of the Risk1 variable if an establishment is indeed already going to be re-sampled given a positive test result in the previous month.

Reply: Current month refers to the month from which sample results are available. Positive Risk1 culture results from an establishment in the current month mean that establishment will be sampled in the following month. The current month is termed month 1, from which months are counted backward as months 2, 3, 4, 5, and 6. If an establishment does not have a positive culture in the current month, it may be sampled in the following month if its baseline risk score rank with adjustment falls into the top 800 or 1,000, depending on the number of weeks in the month. An establishment's baseline risk score rank can only be adjusted downward if there are no penalty points assessed. An establishment with penalty points may not fall into the sampled group if its baseline risk rank is not great enough after the addition of penalty points.

Comment: On the top of page 13 of the Report, the authors state that the baseline risk score is converted to a rank by adding the number of positive sample in the "current" month. What kind of scale is used in this regard?

Reply: This statement is not in the revised report. The baseline risk score rank is adjusted by one adjustment at a time. It can be increased by Risk1 or Risk3 but not both. It can be decreased by Risk4 only when Risk1 and Risk3 are 0. The adjustment scale for the baseline risk score or Risk2 rank is in a normalized scale relative to the risk2 rank ranging from 1 divided by the maximum rank to unity since the weights of Risk2 rank are all equal to 1. The Risk adjustments for Risk1, Risk3, and Risk4 are all fractional in the normalized scale since all risks range from 1 divided by the maximum risk to unity. The generalized risk factor weights proportionally scale each risk factor contribution as 74.87% for Risk2, 19.91% for Risk1, 3.32% for Risk3, and 1.90% for Risk4 before adjustment of the Risk3 and Risk4 contributions to be equal.

Comment: In a related manner, Results for the Risk-Based Verification Sampling Program (phase 1, Jan-Sept., 2005) are provided on pages 11-13 of the Report, in which a series of equations is provided to evaluate which establishments from the June 2005 sampling frame should be chosen for sampling based on their calculated risk ranking. This also is VERY CONFUSING. All the previous narrative, including the Report and Appendix VIII, focused on a rank (or score) per facility. Now, we apparently have an overall risk ranking (inclusive of all plants), but how was this derived from the individual plant scores? The reader needs a CLEAR description of how an algorithm designed to rank individual plants can be used to do a global ranking. More detailed sample calculations would help in this regard.

Reply: We have clarified the point that the risk ranking is done on establishment alternatives and not on the total establishment risk for all alternatives.

Comment: One underlying issue that contributes to these problems is inconsistency in terminology. For example, the first time the reader sees the terms Risk1, Risk2, Risk3, and Risk4 is on page 11 of the Report. In addition, the Report makes no mention of the term Risk1, nor does it clearly provide the full equation for the Risk Ranking Algorithm. One must refer to Appendix VIII for details, and even that documentation is not entirely transparent. That said, consistency in variable names (between the Report and the Appendices) is critical, as is a full description of all equations.

Reply: These errors have been corrected.

Comment: For the quantiles (Q80) of the *L. monocytogenes* distribution at retail given in Table 2, is there any association between Q80 values and the establishment size?

Reply: In all the equations presented, size refers only to annual production volume within a risk alternative. Size does not refer to HACCP size designations of Large, Small, and Very Small plants. Size refers only to High, Medium, and Low annual RTE production.

Therefore, Q80 does reflect annual production volume within a risk alternative but not HACCP size in the equations presented.

Comment: For the equation given on Page 10, what is the maximum penalty points?

Reply: This is the delta constant calculated from the average number of ranks from each positive plant over months 2 through 5 as the difference between their unadjusted baseline risk rank and the maximum rank for that month. This is the maximum number of ranks (points) any plant can be penalized.

Comment: The Conclusions section of the Report (Page 20) actually gives a breakdown of the year 2005 FSIS L. monocytogenes testing, which includes both risk-based sampling and random testing. Basically, the authors summarize these data to conclude that the *L. monocytogenes* positivity rate for Alternative 3 establishments is 1.5 and 2.5 times higher than that for Alternative 2b and Alternative 1 plants, respectively. Little effort is made to discuss these results in light of the Risk-Based Sampling Algorithm.

Reply: This section has been revised to include a statistical test for significance. These data are only indicative of the positive association of the numbers of Lm positives from risk-base sampling and random sampling with the alternatives of increasing risk. Because the data is not entirely from risk-based sampling the result is not conclusive evidence. These mixed culture results were used because the total risk-based samples taken at this time was not large enough for a significance test.

Comment: On pages 13-19 of the Report, the authors do attempt to describe the impact of the Algorithm on Risk-Based sampling. This is done by comparing rank based on raw risk score compared to the scores calculated after adjustment for historical laboratory results. The authors demonstrate that establishments with an L. monocytogenes positive culture in the current month show no change in their ranks, and are therefore automatically sampled. They also show a change in numbers of establishments sampled after adjustment for historical laboratory data (Table 2) and the manner in which pre- and post-adjustment impacts overall plant risk scores (Figure 2) and Alternative-specific plant risk scores (Figures 3, 4, 5, and 6). In Figure 7, the authors show which plants were "chosen" for sampling (based on their scores). It is appears that a risk score of approximately $\geq$1200 (which corresponded to the top 800 risk ranks) was used as the cutoff for sampling, and that proportionally, more plants within Alternative 3 fell in this sampling range when compared to the other Alternatives (Figures 8 and 9). In this manner, the authors did demonstrate that the Algorithm, when including the impact of the adjustment factors, results in proportionally more sampling for Alternative 3 plants, and less so for the other Alternatives, with Alternative 2b establishments falling somewhere in the middle. They therefore do demonstrate that the Algorithm behaves in a manner consistent with its purpose. However, in the absence of transparency regarding the calculation of weighting factors, one cannot conclude that the results necessarily flow logically from the state model structure.

Reply: These errors have been corrected. The calculations for weighting factors have been made explicit.

Comment: It would be helpful if Figure 8 was revised to display % of plants in each Alternative sampled (both before and after Adjustment).

Reply: This has been done.

Comment: Revision of the Conclusions section is recommended. A list of significant conclusions, including the results from the sensitivity analysis, would be appropriate. For example, an important conclusion from the sensitivity analysis is that variables Risk3 and Risk4 are of relatively the same importance.

Reply: This omission has been corrected and a section on conclusions has been added.

Comment: The Report, and Appendices VIII and IX state that sensitivity and uncertainty analysis were done with respect to the Risk Ranking Algorithm. Unfortunately, this concept is not clearly explained in either of these appendices. In point of fact, Appendix VIII (Risk Ranking Model-Sensitivity Analysis) is the only place that the concept of variability, uncertainty, and the probabilistic framework of the model is explained or discussed. However, it is not clear to the reader if this application is an additional analysis in order to provide some insights with respect to the model sensitivity, or if the distributions are actually incorporated into the model structure. I suspect it is the former, and the model is in reality, mostly deterministic.

Reply: Because the risk ranking algorithm is partly deterministic and partly distribution based, the sensitivity analysis only adds information concerning the non-deterministic parts of the algorithm. These components are the empirical distributions of the risk factors that represent a nearly closed set of distribution values that vary within proscribed limits each month. By restricting the algorithm to used output distribution ranks, the effect of month to month variation is reduced to the range of the risk ranks. The variability of each risk component is evaluated using standardized regression on the ranks of the output risk distribution corresponding with the magnitude of standardized regression coefficients. The uncertainty in the output risk distribution is evaluated using the bootstrapped estimates of the standardized risk variables and their associated standardized regression coefficients.

Comment: For the sensitivity analysis, two objectives are given on Page 55 in Appendix VIII. However, these objectives are not completely clear based on the given explanations. Moreover, it is not clear whether co-mingled simulation of variability and uncertainty was performed for sensitivity analysis or these features of the model were separated using a two-dimensional Monte Carlo simulation. For example, on Page 58 it is mentioned that the model was run for 100,000 iterations. However, it is not clear specifically how many uncertainty simulations and variability iterations were performed.

Reply: In the modified sensitivity analysis there were 1,981 variability iterations and 1,000 uncertainty iterations. It was found that increasing the number of uncertainty iterations to 10,000 or greater did not improve the uncertainty estimates since the uncertainty model proved to be quite stable.

Comment: The authors mentioned that for sensitivity analysis two alternatives were performed considering the presence and absence of correlation between model parameters (Page 58, Appendix VIII). However, there is no indication of the type of correlation structure which was considered in the model.

Reply: The latent correlation structure present in the empirical distributions was not disturbed because no hypothetical fitted distributions were used. The correlation structure of the empirical distributions was defined by Spearman rank correlation matrices.

Comment: In general, the method used for evaluating the sensitivity of the model output to individual parameters was not well explained. Based on the Reviewer's understanding of the given methodology, the method evaluates the influence of one individual source of variability or uncertainty at different percentiles of its probability distribution while setting all other variability and uncertainty sources to their base-case values (point estimates). For example, when considering the impact of "Alt3 DM High Volume" on the output of the score equation (Table 5, Appendix VII), the other sources of variation are set to their point estimates. However, the authors did not try to distinguish between the key sources of variability and key sources of uncertainty as they have completely different implications.

Reply: These errors have been corrected.

Comment: Typically, little information regarding the distributions of the Risk-Based Sampling model inputs and simulation techniques was provided. Therefore, key questions that should be addressed include the following: (i) how was the value of an input altered?; and (ii) what sampling technique was used? It is necessary to clearly list the distribution assumptions and parameters and to clearly describe related simulation techniques when doing uncertainty analysis.

Reply: The simulation and sensitivity analysis methods were changed from a Monte Carlo-based sampling scheme of fitted probability distributions to one using random sampling of empirical distributions for all risk factors based on the original data. The variability distributions were estimated from the fixed data set of 1,981 alternatives. The uncertainty distributions were estimated from an expanded empirical data set of 2,493 using a random sampling of size 1,981 to estimate bootstrapped uncertainty parameters.

Comment: In general, in Appendix VIII, figures are shown with lack of proper legends, and hence, they are not self-explanatory. For example, the reader is not certain what the variables Risk 3.2, 3.3, 3.4, etc, refer to? I would assume they refer to the five weighted probability distributions used in the sensitivity analysis for the variable Risk 3?

Reply: This error has been corrected.

Comment: Tables also lack proper formatting with respect to the number of significant figures for tabulated values; it is difficult to read these values because of the unnecessarily large number of significant figures reported. In addition, no narrative at all is provided for Figures 7 and 8 and their associated tables.

Reply: These errors have been corrected.

Comment: It is my understanding from the material presented in Appendix X that the Agency itself will eventually collect information on each establishment with respect to its control measures. Based on this, documentation will be collected (by inspectors) and used to complete a decision tree that will ultimately impact the risk score of the specific establishment. It is not clear from the materials given when agency collection of such documentation would begin, or exactly how (quantitatively or qualitatively) it might be used to adjust the Risk Rank. In the current absence of such resources, the Risk Ranking Algorithm could be adjusted, but my suggestion would be to do this in a qualitative rather than quantitative manner, perhaps based on the "Conclusive," "Substantiated," and "Inconclusive" designations combined with production volume. Using a decision analysis (tree) framework is a nice touch and a sound approach.

Reply: This point is well taken. Once verification information can be collected and we have enough information as to the qualitative nature of establishment volunteered data, it will be easy to select more of those establishments for sampling that had submitted poor quality data, provided there were enough sampling resources to handle the extra sampling load for low risk establishments not already selected for sampling.

Comment: Page 2 of Report, paragraph 3, sentence 4: Suggest rewording to "…..defined groups of high-risk individuals, including…."

Reply: This have been reworded to: "Listeriosis occurs most often in certain well-defined groups of high-risk adults, including pregnant women, neonates, and immunocompromised adults, however may occasionally occur in individuals with no predisposing conditions (Slutsker et al. 1999). Illness in pregnant women can result in miscarriage, stillbirth, or severe illness or death of a newborn infant (CDC 2002). Published fatality rates for listeriosis range from 20 to 40% (Shuchat et al. 1992)."

Comment: Page 2, second full paragraph, first line: Suggest taking care of the word "risk profile," as this has a defined context to many risk assessors and managers. Is there an alternative word that can be used in its place?

Reply: This has been changed from "risk profile" to "risk factors".

Comment: Page 4 of Report, the sentence in the 4th paragraph regarding the advantage of targeted sampling plans over random sampling plans might be too strong. Random sampling plans have some advantages that should be acknowledged with respect to better capturing the variability.

Reply: The focus of the comment has been changed to stating that the allocation efficiency directed by risk based sampling compared to that directed by random sampling is greater.

Comment: Page of Report, end of 1st paragraph: There is actually quite a bit of variability in refrigeration temperatures both at domestic and retail levels, and this is alluded to in the Audits International citation. This should perhaps be noted here.

Reply: An additional qualifying sentence has been added: "It should be noted that there is significant variability and uncertainty in domestic cold storage temperatures due to insufficient monitoring when compared with the temperature controls in place for commercial cold storage, warehousing, retail, and transportation."

Comment: Table 1 on Page 12 should be labeled as Table 4.

Reply: All Tables have been renumbered.

Comment: Three numbers that are given for each alternative in Table 1 on Page 12 should be labeled. Are these numbers for large, medium, and small establishments?

Reply: This Table has been renumbered as Table 4 and the headings have been defined in terms of volume production and not HACCP size. The volume designations are High, Medium, and Low volume production.

Comment:. There should be a better explanation of the materials discussed in the first paragraph on Page 13.

Reply: The distinction between the number of establishments and the total number of alternatives as the objects for ranking is emphasized in this paragraph so as to remove previous confusion.

Comment: Page 20, 4th paragraph; the authors mentioned that "The performed analyses showed the model is quite stable in this regard". It is not clear what the authors are referring to.

Reply: This paragraph has been rewritten and the sentence does not appear in the new version.

Comment: It is not clear what the authors mean by "dividing" the volume distributions into known variability distributions. Better wording should be used.

Reply: This wording has been removed. Well-described probability distributions corresponding to the algorithm’s input risk distributions have been changed to empirical data distributions for the algorithm’s input risk distributions describing variability.

Reviewer 4

Comment: The results from the sensitivity analysis presented in Appendix VIII are difficult to interpret. Figure 1 is dominated by the differences in the 90% percentile for some establishment categories (ie. Alternative 3 and high volume). No interpretation of this finding is provided. Does this mean that for most parts of the input variable distributions (i.e. below 90%), none of the input variables has a particularly strong influence on the predicted score?

Reply: A modified sensitivity analysis shows that deli meat volume dominates the baseline risk ranking and the baseline risk ranking dominates the final risk ranking. In both the baseline risk score and adjusted based risk rank distributions no single risk factor has a dominant effect until after the 80th percentile. This effect is the result of all input distributions except Risk4 to predominantly contain null values below the 80th percentile.

Comment: Figure 2 shows the relative importance of the adjustment factors representing past testing results. Risk 4 (derived from the number of positive samples in past 6 months) and Risk 2 (the model prediction) both appear to have a strong influence on the output score. This means that the parameters used to link this data to the risk ranking algorithm need to be well justified.

Reply: Because the sensitivity analysis has been refined and more detail has been added Risk3 and Risk4 account for about 10% of the final risk ranking outcome while Risk2 accounts for over 70%. The components of Risk2 are well documented to have causal links to public health risk from the 2003 FDA/FSIS Risk Assessment on Deli Meat.

Comment: the relationship between Risk 3 (no of months with negative results) and Risk 4 is examined. The ratio was found to be near unity (plus minus 20%). The interpretation of this finding as currently presented in Appendix VIII is unclear.

Reply: In order to reduce bias for correcting baseline risk too much for past positive culture results (Risk3) increasing risk and negative culture results (Risk4) decreasing risk, analysis of the relative influence of each risk factor should reveal a ratio of effects near unity or a difference in effects near zero. We have revised the sensitivity analysis to provide the correction term to the Risk4 weights to allow the effect of both risk factors on the output risk ranking distribution to be equal. Using this adjustment on the weights the ratio of standardized regression coefficients is forced to be unity with an average difference of zero. The error on the ratio or the difference is typically less than 10% of the average.

Comment: The sensitivity analysis outputs presented in Appendix IX are also difficult to interpret. What does the data presented in Table 1 mean?

Reply: The sensitivity analysis in Appendix VIII and the uncertainty analysis in Appendix IX have been revised for clarity.

Comment: The risk ranking algorithm has been subjected to extensive sensitivity analyses, but has not been validated either.

Reply: At the time of writing there is insufficient data to validate the risk ranking algorithm. We will evaluate and perform a validation assessment when sufficient data is available.

Comment: The algorithm assumes fixed ratios between the risks of different product types. No justification for this is provided.

Reply: The 2003 FDA/FSIS risk assessment of 23 food categories for Lm risk is the source for these fixed ratios based on estimated per serving risk for deli meat, hot dogs, and other RTE products.

Comment: The algorithm used to calculate the risk score appears to be based on sets of deterministic equations. It therefore ignores the uncertainty and variability of the underlying inputs.

Reply: Including uncertainty and variability estimates for each establishment's baseline and final risk ranking will not add anything to the final risk ranking because it is a relative scale final output. If we want to estimate each establishment's absolute variability and uncertainty of risk, then this can be done. Simulation shows that Risk2 before ranking has a variability of xx and an uncertainty of xx. Also simulation shows that the adjusted risk before final risk ranking has a variability of xx and an uncertainty of xx. Including these average variability and uncertainty estimates in the risk ranking algorithm do not change the baseline or the final risk ranking.

Comment: The penalty and reward scores are based on the results from the last 6 months of testing. While the 6 months seem justified on the basis of the simulations described in Appendix VII, the algebraic calculations used to combine them have not been justified.

Reply: Risk3 is estimated from the number of past positives over the previous 6 months excluding the present month for each establishment. The formula for calculating this risk is:

$$\text{Risk3} = 0.231 \text{ x } [1,0] + 0.205 \text{ x } [1,0] + 0.191 \text{ x } [1,0] + 0.186 \text{ x } [1,0] + 0.185 \text{ x } [1,0]$$

This formula was derived from an autocorrelation study of expected positives by month over the next five months after an initial positive in the first month. The model used was for an alternative 3 plant working two shifts per day over a 30 work month for one year totaling 1,000,000 iterations.

Risk4 is estimated as the sum of negative culture results for each plant over the past six months including the present month. The sum is divided by the maximum total number of samples achieved by any plant during the six-month period. Only plants that have the most negative cultures receive the full benefit of the W4 weight.

Comment: The justification for the calculation of penalty and reward adjustments is presented in Appendix VIII and IX. Appendix VII provides part explanation and refers to Appendix IX for further information to some aspects of the calculation. This is confusing, and it is not clear how the weighting factors for these adjustments were derived. In any case, these weightings are based on model predictions, and therefore assume that these are an acceptable representation of the 'true' values.

Reply: The weights are only partially based on model predictions as prevalence relative risks for each product-volume alternative for W3 and W4. W3 and W4 are each multiplied by delta, a data-dependent value for the average number of ranks to add or subtract from the baseline risk rank for plants receiving a penalty or a reward for performance. W1 and W2 are data-dependent.

Comment: The information provided in the reference Small (2000) in Appendix II is insufficient to justify avoiding inclusion of uncertainty in the modelling process. In the same paragraph further statements are made about the rationale for excluding uncertainty from the modelling process, and one reference (Casman et al 1999) is presented. Ultimately this decision needs to be made in consultation with stakeholders and risk managers. It is stated that FSIS 'finds it reasonable, pragmatic and sufficient to use a simple, broad distribution to characterize in-plant model parameters'. This fact needs to be clearly and repeatedly communicated to stakeholders and risk managers. They need to take into consideration when interpreting the modelling outputs that the model outputs may well misrepresent the true quantitative relationships in the system which is being modelled.

Reply: At this time we feel the inclusion of only parameter variability in the 2003 deli meat model is justified and that the inclusion of parameter uncertainty would overly complicate the model and substantially reduce its computational efficiency without gaining sufficient insight into answering risk management questions. We feel that the amount of uncertainty additionally modelled would not change the qualitative conclusions of that model. Those conclusions being that there are substantial differences between the public health risks of alternative 1, 2, and 3 with alternative 3 being the riskiest. The quantitative estimates of the numbers and prevalence of Lm per lot at retail may change slightly and the total associated errors may change slightly is uncertainty is included. Since the parameter uncertainty is likely much less than the parameter variability we feel this is a valid assumption to make and emphasize to risk managers.

Comment: The modified 2003 FSIS LM in-plant risk assessment described in Appendix III is based on a very complex simulation model. There is some validation of model predictions, for example in Figure 5. This seems to suggest that the model slightly overestimates LM concentration at retail. Is this difference acceptable? How does it impact the scores produced by the risk ranking algorithm?

Reply: The risk ranking is relative and independent of the magnitude of the calculated risks, which may be different than the *actual L. monocytogenes* counts and prevalences observed. The risk ranking therefore does not depend on *the L. monocytogenes* risk model predicted magnitudes, but on the dependence is on the relative risk rank of one plant to another. There is no impact on the risk ranking.

Comment: It needs to be noted that the sensitivity analyses can only allow inferences about the internal relationships between the model variables, and the defined relationships between them, but not a validation.

Reply: This point is well taken. We agree that a validation analysis needs to be done. Once a validation analysis is done the risk ranking algorithm may be relied upon unchanged or modified so as to better reflect a valid establishment Lm risk ranking.

Comment: The results generated by the risk-ranking algorithm should include an estimate of uncertainty/variability. This is particularly critical since there are a large number of variable and uncertain parameters included in the various calculations, and it is not possible to predict how these will be 'propagated' through the model. The risk managers then have to make a conscious decision with respect to the value from that distribution (may be mean, median or upper 95% value) which will be used to define the risk management measures. After having gone through that process it may be possible to simplify the model.

Reply: We have taken additional measures to provide a complete uncertainty and variability analysis of the risk ranking algorithm output. We have evaluated the uncertainty and variability associated with the 37 input variable risk factors in the full risk ranking algorithm as to the individual effect of each on the adjusted baseline risk and its rank.

Comment: Some of the appendices (eg Appendix VIII) would benefit from better descriptions of the graphical and tabular outputs

Reply: We have provided more extensive descriptions and more complete tables and figures in the rewritten Appendices VIII and IX.

Comment: More interpretation of the findings should be provided in the appendices presenting results from model analyses (Appendices VIII and IX).

Reply: We have done an extensive rewrite of Appendices VIII and IX to satisfy this comment.

Comment: The sensitivity/specificity of the model for correctly classifying individual retail outlets needs to be determined

Reply: It is not possible to calculate the sensitivity and specificity of the model without calibration data. Such calibration data are not available for producing establishments or retail establishments.

Comment: It might be useful to assess the spatial or temporal dependence in the likelihood of identifying LM positive samples. Is it more likely to obtain positive samples in summer or winter, or within certain parts of the country? If there are such relationships, they should be incorporated in the risk-ranking algorithm.

Reply: It is a goal to incorporate *L. monocytogenes* biotype data as well as establishment location and timing of positive samples in the risk-based risk ranking algorithm.

Comment: Due to the lack of 'real' data the parameters (actual scores as well as weightings) had to be based on assumptions and model predictions. It should be possible to improve the confidence in these parameters and adjustment mechanisms once more data from establishments becomes available.

Reply: The data used in the sampling verification algorithm have been expanded. Instead of using theoretical distributions for deli meat volume, hot dog volume, other products volume, positive culture results by month, and negative culture results by month, empirical data distributions were used such that the original variable correlation matrix was maintained. Constants in the algorithm for deli meat risk per serving, hot dog risk per serving, other products risk per serving, risk alternative Q80, and risk alternative prevalence were assumed constant without a probability distribution. These constants were predicted estimates from the FDA/FSIS Lm Food Risk Categories Risk Assessment and the FSIS Deli Meat Lm Risk Assessment. The variability and uncertainty for each empirical distribution was determined in the sensitivity and uncertainty analysis as well as the variability and uncertainty of the baseline Lm risk rank distribution and the final Lm risk rank distribution. Improved estimates were made based on these data.

Comment: A validation study where the risk-ranking is compared with a detailed investigation to determine the 'true' status of each establishment would be very useful. The reliance on non-verified self-reported information is a weakness of the current approach. The first step should be to quantify the data errors through detailed auditing of a random sample of establishments. This survey needs to make sure that particularly those establishments benefiting from reduced sampling likelihood, i.e. those reporting Alternatives 1 and 2 are included in sufficient quantities to allow meaningful estimation of reporting bias.

Reply: We agree with this statement. We want to complete a validation study as soon as the data can be made available.

Comment: Normally, it would not be appropriate to reward the businesses that have been verified, since it will never be possible to verify more than a relatively small number. If risk managers feel that consumer protection is the absolute priority and the risk of misclassification is high, they may wish to consider such a measure.

Reply: Verification can occur on several levels that do not incur equal risk. Verification of the risk components used in risk ranking is only necessary when considering the present ranking system. Since all components now are self-declared on FSIS FORM 10,240-1 all establishments have the same level of verification. At the time establishments are verified as to alternative and post-lethality exposure controls in place the algorithm will be modified to label verified establishments. At that time risk managers will decide if verified establishments will qualify for reward adjustments to their risk ranking.

Comment: How might a metric of the effectiveness of the risk based verification sampling approach be developed?

Reply: The required statistical power of such a comparative investigation needs to be defined in consultation with the risk managers, who may decide that 90% confidence levels or effects significant at the 10% level are sufficient to inform the decision making process. It would also be possible to target specific risk groups to maximize the chance of obtaining useful conclusions for that category. The parameters in the self-reporting data most likely to influence a reduced score (e.g. quantities of product, or risk management procedures adopted) should be used to define such establishments to be targeted during such an audit. Statistical associations between these self-reported input data and the score for the establishments could be used for this.

Reviewer 5

Comment: From a public health perspective, will the incentive, reduction in the number of samples taken, be perceived as a necessary and sufficient means of achieving the public benefit—safer food due to more stringent control measures? This is the risk communication challenge. At this point, I am not convinced that the authors have adequately considered the multiple audiences involved in responding to the report. This is not to say that the report is flawed. Rather, I believe the project would benefit from further consideration of these risk communication factors.

Reply: The stakeholders are the producers, the consumers, government policy makers, government oversight (OMB), government inspectors, and government risk managers. Incentive for producers is created by offering decreased sampling by converting to a less risky alternative or more sampling done with negative results. The tendency should be to change alternatives if costs permit. An avenue for submitting plant culture results should be pursued to facilitate exchange of plant data without penalty for positive results they capture. Consumers need to see that plants are moving from high risk to low risk alternatives or that plant prevalence rates for alternative 1, 2a, and 2b are proportionally less than alternative 3. OMB program assessors, facility inspectors, and risk managers need to see that the data given to consumers show the sampling plan is working and risk is reduced according to the *L. monocytogenes* risk model predictions. The efforts all rely on collecting the pertinent prevalence data.

Comment: Can, for example, the algorithm inspire risk-management actions on the part of the intended audience—the producers?

Reply: That is one of the intentions, to motivate establishments in high risk alternatives to move to lower risk alternatives so that they may be sampled less.

Comment: the report could make mention of how a selected percentage of the testing opportunities could take into account the uncertainty created by unpredictable production changes or lapses. In other words, the inspection process could build in a modest amount of flexibility in order to allow inspectors to test their hunches as they see any form of unanticipated evolution in the production of RTE meat and poultry products. In this manner, the inspectors, those closest to the actual product, have the opportunity to communicate upwardly within FSIS to receive some small latitude for influencing the allocation of their time and resources.

Reply: The algorithm is amenable to frequent updates of volume and product data if they are made available. There is no mechanism for plant inspectors to record changes in volume or products produced and then pass them to personnel updating the database.

Comment: My background in risk communication leaves me wondering, however, if the communication exchange between FSIS and the establishments producing the relevant products is sufficient. Is there a mechanism for evaluating the communication between

FSIS and the establishments as this procedure is adopted? Will there be cases of reporting errors? Will there be a misinterpretation by an establishment as it attempts to institute Alternative 1, 2, or 3? These are questions that are not answered in the report. If such problems do arise, how will they be noted? What actions can be taken?

Reply: This area needs work. Cases of reporting errors and their correction can now only be handled through the establishment district office. A questionnaire checklist has been completed for administering to individual establishments to verify responses on FSIS FORM 10,240-1.

Comment: Are there any studies available on the rate of compliance? Have the actual producers been interviewed or surveyed to see how they respond to the proposed measures for controlling Listeria monocytogenes. Without input from procedures, I cannot be certain that the proposed control measures are not in some way resented by producers. Such resentment, if it exists at all, could influence compliance rates and reporting accuracy.

Reply: The most accurate data on compliance with the Listeria rule is the establishment culture result history. This data is collected by CSOs, so there is little bias from the establishment.

Comment: I have mentioned above that, from a communication perspective, I am always a bit apprehensive about self-reported data. Thus, I would be in favor to giving more influence to verified data than to unverified data. There are simply too many opportunities for organizations to apply or assume ambiguity in a strategic sense if verification is absent. Any additional strategies for enhancing or expediting the verification process would improve the overall project.

Reply: At this point all plant supplied data are unverified. There is no basis for assigning weights for verified versus unverified data.

Comment: If a plant shows unusually high risk, there should not be limits on how many samples can be taken from that plant. Placing limits on samples should no longer be necessary now that the standards for sampling frequency have been clarified and appropriate alternatives for risk reduction have been provided. At this point, forcing sampling limits in extreme cases would be inappropriate and arbitrary. Forced limits create an unfortunate situation since the ultimate goal of the project is to create a sampling process that matches the level of risk.

Reply: We agree.