Appendix S:
Analysis Plan for the Behavior Change Study

S.1 Meal Preparation Experiment

Research Aims and Hypotheses. The goal of the meal preparation experiment is to provide information on which of four SHI labels (current label as the control and three alternative SHI labels as the treatments) is more likely to lead to consumers following the safe handling instructions on the SHI label for handwashing, using a food thermometer, keeping raw meat/poultry separate from ready-to-eat (RTE) product, and cleaning/sanitizing kitchen surfaces and equipment. The study uses a randomized experimental design with random assignment to each of the four study conditions. The study will answer the following research question: Of the four SHI labels included in the study, does one lead to participants' greater adherence to safe handling instructions compared with the others? We will use a one-way analysis of variance (ANOVA) to examine the null and alternative hypotheses:

- Ho: There is no statistically significant difference among the label adherence score for the 4 labels (current SHI and three treatment options).
- Ha: There is a statistically significant difference, indicating that the label adherence score for at least 1 of the 4 SHI labels differs from the remaining adherence scores.

Coding. Trained coders will watch the video observations and conduct the coding using the video observation rubric to assess adherence to the safe handling instructions found on the SHI label (described in more detail below). The decision trees used for the coding are provided as Appendix M. The coders will work in pairs during training to help ensure inter- and intracoder reliability and then will work separately to code the study data. Coders will meet periodically to compare notes and discuss areas where they do not agree on codes to reach an agreed-upon outcome. Coders will be trained and evaluated on the coding rubric before data collection.

Analysis Procedures. We will conduct statistical analyses comparing the label adherence scores among the four groups (i.e., current SHI label and three alternative SHI labels). Adherence scores will be based on observed behaviors related to four instructions on the SHI labels (handwashing, using a food thermometer, keeping raw meat/poultry separate from RTE product, and cleaning/sanitizing kitchen surfaces and equipment). The results of this analysis, along with the findings from the eye-tracking study, will help to inform the design of the final SHI label.

Table S-1 describes our approach for how we will code and score adherence to each of the four instructions. We will assign each participant an overall index by summing the scores for each of the four instructions. Our primary analysis will use a continuous measure (0–4) that includes "partial" credit for participants who follow some but not all steps associated with an instruction (e.g., 0.50 points for using a thermometer for the fresh and frozen product but cooking to less than 160°F).

For the primary analysis, we will use one-way ANOVA with the Tukey-Kramer adjustment for post-hoc estimates to compare all pair-wise condition means. For the secondary analysis, we will use chi-square goodness of fit test to assess differences in the frequency of participants who are fully adherent across the four SHI label groups.

Table S-1. Coding and Scoring for the Meal Preparation Experiment

Contributing Factor	SHI Instructions	Message Adherence Index for Message #3
Improper/inadequate handwashing	**1. Wash and dry hands** **2. Wash hands** with soap and water for 20 seconds, and then **dry** **3. Wash hands** with soap and water for 20 seconds before cooking and after touching this product, and then **dry**	.25 points per action (total score of 0 to 1 for each HW event) ■ Use water (when initially wetting hands or during rinsing) ■ Use soap ■ 20 seconds ■ Dry with clean towel or single-use towel =0 if HW required and not attempted *nested approach since multiple observations per participant; we will code each HW event and the score for each HW event, which will be reflected in the overall score for HW
Improper/inadequate thermometer use--frozen meatballs	**1. Use food thermometer** **2. Use food thermometer.** Cook to temperature shown below. **3. Use food thermometer** at end of cooking. Cook to temperature shown below.	.25 points for each action ■ Use thermometer (by placing in the item, protective sheath on or off) ■ Temperature = 160°F or above (from eye tracking video)
Improper/inadequate thermometer use--raw ground beef	**1. Use food thermometer** **2. Use food thermometer.** Cook to temperature shown below. **3. Use food thermometer** at end of cooking. Cook to temperature shown below.	■ Use thermometer(by placing in the item, protective sheath on or off) ■ Temperature = 160°F or above (from eye tracking video)
Cross-contamination between raw and RTE foods--raw ground beef	**1. Keep uncooked meat and poultry separate** 2. This product is uncooked: **Keep separate** from other foods 3. This product is uncooked: **Keep this product separate** from other foods until cooked.	.5 = no direct touch (for entire observation) 0 = One event of not-separate (i.e., a touch)
Cross-contamination between raw and RTE--frozen meatballs	**1. Keep uncooked meat and poultry separate** 2. This product is uncooked: **Keep separate** from other foods 3. This product is uncooked: **Keep this product separate** from other foods until cooked.	.5 = no direct touch (for entire observation) 0 = One event of not-separate (i.e., a touch)
Improper/inadequate cleaning *and* sanitizing	**1. Clean utensils and surfaces, and then sanitize** **2. Clean utensils and surfaces** that contact this product with soap and water **and then sanitize** **3.** Always **clean utensils and surfaces** with soap and water after contact with this product **and then sanitize**	0-1 points for each trigger of contact 0 no action .5 point for clean[a] .5 point for sanitize after cleaning[b] *nested approach since multiple observations per participant; we will code each trigger and the score for each trigger, which will be reflected in the overall score for cleaning and sanitizing
Total score		**Index – continuous measure, 0-4**

Notes: HW = handwashing

[a] Clean is defined as an event in which the participant thoroughly washed the surface with hot, soapy water and dried it with clean one-use towel.

[b] Sanitize is defined as an event in which the participant used one of the provided sanitizers (Clorox, Clorox wipes, 409 spray, Lysol disinfectant spray) to spray the surface and wiped it dry with a clean, one-use towel or to place the utensils in the dishwasher and use heat (after cleaning).

S.2 Eye-Tracking Study

Research Aims and Hypotheses. The goal of the eye-tracking study is to provide information on the ability of various SHI and RTE/not-ready-to-eat (NRTE) labels to capture consumers' attention. This goal will be achieved by gathering and analyzing data on participant gaze patterns and comparing patterns across different SHI labels. The study will answer the following research questions:

- Which version of the SHI label is most often attended to when participants look at a busy food package?
- Can participants properly distinguish between RTE and NRTE products?

Description of Eye-Tracking Tasks. A trained lab assistant/interviewer will use script to guide participants through three tasks. Task A is a distractor task and will use one of the RTE products. The use of a distractor will help ensure that participants are not aware of the primary research goal and disguise the fact that the true intent of the study is to capture a participant's gaze data on the SHI label via eye-tracking technology. Participants will be asked to consider how healthy the product is, then asked to answer two questions about fat and salt content.

The purpose of Task B is to collect information on participants' attendance to the SHI label. Participants will be asked to consider how they would safely prepare the products at home while they examine three of the NRTE products in a randomized sequence (raw ground beef, frozen packaged hamburger patties, and NRTE chicken cordon bleu). This prompt will cue the participants to pay attention to the product labels but will not draw attention to a particular aspect of the label. After examining the three products, participants will be asked to point to the location on the product packaging where they recall seeing a warning about food safety.

The purpose of Task C is to collect information on how participants distinguish between RTE and NRTE products. Participants will be asked to look at the six products in a randomly assigned sequence and instructed to indicate for each product, whether the product is already cooked or raw (or not sure) by placing the appropriate sticker on the product.

We will ask all participants the same questions across all conditions; however, we will use block randomization to randomize the order in which the products are shown.

After the eye-tracking study is completed, a trained lab assistant/interviewer will use a semistructured interview guide to collect information on the following topics:

- reason for indicating which products are already cooked vs. raw
- awareness of current SHI label before study
- debriefing questions on meal preparation study
- participants' usual behavior at home when handling raw or uncooked meat and poultry products

Trained interviewers will follow the questions in the interview guide while allowing flexibility for dialogue and ensuring that participants feel comfortable providing more detail when needed. Each question will also include follow-up questions and probes to gather further details from participants. At the end of the study, we will inform participants of the true nature of the study and provide them the opportunity to request that their data not be used in the study if they have any concerns.

Analysis Plan. We will conduct statistical analyses to address the study's primary research questions. Analytic procedures will be based on expectations associated with distribution of the dependent variable. For variables that assess proportion and rates, we will use nonparametric procedures. For variables that assess differences in lengths of time, we assume that inferential statistical procedures suitable for large samples and dependent variables that are IID (normal) are appropriate. This assumption will be evaluated for each outcome. Preliminary analyses will include univariate examination of each variable to be used in scales and descriptive statistics examining the demographic characteristics of the study participants. For outcomes that do not meet the assumptions of IID normal distributions, we will employ appropriate nonparametric procedures.

Additionally, we will use two data visualization techniques—gaze plots and heat maps—to describe the temporal-spatial distribution of attention. Gaze plots reveal the temporal sequence of visual attention. Visual attention, most commonly expressed as fixation duration, is indicated by the diameter of the fixation circles. A gaze plot consists of a sequential set of circles that indicate the path of visual attention. Heat maps are used to display the distribution of visual attention across the stimulus. Heat maps use color variation to convey the amount of attention received, with deeper colors (i.e., red) indicating greater amounts of attention to a given area. Gaze plots will be used to assess individual patterns of visual attention, while heat maps will be used to assess aggregate attention across large numbers of participants.

Primary Outcome: Proportion of Participants who Fixate on the Area of Interest (AOI). The analysis will be based on the eye-tracking data collected in Task B of the eye-tracking script. The proportion of participants who fixate on the SHI is an indicator of attentional capture. The dependent variable for this analysis will be a binary indicator that dichotomizes participants into those who fixate on the AOI and those who do not fixate on the AOI. The AOI will be defined as the package space that conforms to the SHI label. We will use chi-square goodness-of-fit tests to examine the null and alternative hypotheses:

- Ho: There will be no difference in proportion of participants who attend to any of the SHI labels.
- Ha: The proportion of participants who attend to the SHI labels will be significantly different for at least one of the SHI labels.

The chi-square is an appropriate statistical test when the purpose of the research is to examine the relationship between two nominal-level variables. In this case, those variables are SHI label (four levels) and visual attention to the AOI (yes/no). To conduct the chi-square test, the data are summarized in a 4-by-2 frequency table defined by the categories of the nominal

variables. The observed value in each cell (O_{ij}) is then compared with the expected value for each cell (E_{ij}), which is the product of the row (R_i) and the column (C_j) divided by the total sample (N), and the chi-square statistic is generated with the following formula:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

To evaluate significance of the results, the calculated chi-square coefficient (X^2) and the critical value coefficient will be compared. When the calculated value is larger than the critical value, with alpha of 0.05, the null hypothesis will be rejected (suggesting a significant relationship). To determine the degrees of freedom (df) for a chi-square, it is necessary to use the following equation:

$$df = (r-1)(c-1)$$

The r value equals the number of rows, and the c value equals the number of columns. For a chi-square to provide a valid result, several conditions and assumptions must be met. The data must be random samples of multinomial mutually exclusive distributions, and the expected frequencies should not be too small. If the value of the resulting chi-square coefficient is sufficiently large, the null hypothesis of independence is rejected.

Secondary Outcomes: Time to First Fixation on the AOI. Time to first fixation is a temporal measure of attention. The dependent variable for this analysis will be calculated as the number of milliseconds (ms) between the beginning of the exposure period and the point in time the participant's gaze fixates on the AOI. The AOI will be defined as the package space that conforms to the SHI label. ANOVA will be used to examine the null and alternative hypothesis:

- Ho: Time to first fixation will not differ among any of the four SHI labels.
- Ha: At least one of the four SHI labels will demonstrate a time to first fixation that is significantly different from the mean time to first fixation of the other SHI labels.

One-way ANOVA is an appropriate statistical analysis when the purpose of research is to assess if mean differences exist on one continuous dependent variable by an independent variable with two or more discrete groups. The dependent variables for this analysis will be created by summing the intervals before the first fixation interval. The independent variable will be a four-level indicator representing the three revised SHI labels and the current SHI label (as reference level). The assumptions of normality and homogeneity of variance will be assessed. Normality assumes that the scores are normally distributed (bell-shaped) and will be assessed using the One-Sample Kolmogorov-Smirnov test. Homogeneity of variance assumes that both groups have equal error variances and will be assessed using Levene's Test for the Equality of Error Variances. If the data fail either of these tests, we will employ data transformation or nonparametric statistical methods (e.g., Mann-Whitney U test).

The t test will be two-tailed with the probability of rejecting the null hypothesis when it is true set at $p < .05$. This ensures a 95% certainty that the differences did not occur by chance.

Following the main hypothesis test, we will conduct planned contrast analyses to examine each paired (two-group) comparison.

Secondary Outcomes: Total Dwell Time on AOI. Dwell time is a spatial measure of attention. The dependent variable for this analysis will be calculated as the total number of milliseconds between the beginning and termination of all fixation periods within the AOI. The AOI will be defined as the package space that conforms to the SHI label. ANOVA will be used to examine the null and alternative hypothesis:

- Ho: Total dwell time will not differ among any of the four SHI labels.
- Ha: At least one of the four SHI labels will demonstrate a total dwell time that is significantly different from the mean dwell times of the other SHI labels.

One-way ANOVA is an appropriate statistical analysis when the purpose of research is to assess if mean differences exist on one continuous dependent variable by an independent variable with two or more discrete groups. The dependent variables for this analysis will be created by summing the intervals before the first fixation interval. The independent variable will be a four-level indicator representing the three revised SHI labels and the current SHI label (as reference level). We will assess the assumptions of normality and homogeneity of variance. Normality assumes that the scores are normally distributed (bell-shaped) and will be assessed using the One-Sample Kolmogorov-Smirnov test. Homogeneity of variance assumes that both groups have equal error variances and will be assessed using Levene's Test for the Equality of Error Variances. If the data fail either of these tests, we will employ data transformation or nonparametric statistical methods (e.g., Mann-Whitney *U* test).

The *t* test will be two-tailed with the probability of rejecting the null hypothesis when it is true set at $p < .05$. This ensures a 95% certainty that the differences did not occur by chance. Following the main hypothesis test, we will conduct planned contrast analyses to examine each paired (two-group) comparison.

Analysis of Data for Secondary Research Question on Whether Participants Can Properly Distinguish between RTE vs. NRTE Products. The analysis will be based on the data collected in Task C and will be descriptive in nature. A summary of this approach is described below.

1. For each of the six products, percentage of participants who correctly identify the product type (RTE vs. NRTE); test for differences between four SHI labels using chi-square.
2. For each of the six products, provide heat maps for participants who correctly identified the product type (RTE vs. NRTE) and separate heat maps for participants who did not correctly identify the product type. This will provide descriptive information on where participants looked on the package to complete the task and if where participants looked varied between those who provided the correct answer and those who did not.
3. For each of the six products, provide individual gaze plots for a random selection of 10 participants in each of the two groups (correct answer vs. incorrect answer). This will provide similar descriptive information as above but for individual participants.

S.3 In-depth Interviews

We will organize the interview data using QSR International's NVivo 11 qualitative data analysis software (QSR International, 2015). NVivo software allows users to store, classify, sort, and arrange text-based and multimedia information to examine relationships and patterns across study data. NVivo software also includes TranscribeMe, a competitively priced and integrated transcription service that automatically downloads professional transcriptions into the NVivo project file for analysis. For the Spanish interviews, because it is very costly to translate the interviews into English and then transcribe the interviews, we will prepare a typed summary of these interviews in English instead of a word-for-word translation for use in NVivo. We will provide FSIS with the verbatim transcripts of the English interviews and a summary of the Spanish interviews.

A team of five qualitative analysts from RTI will code and analyze written data using a thematic content analysis approach. This common qualitative practice involves searching through data to find recurring patterns and themes. For quality assurance purposes, the RTI qualitative task leader will develop and train analysts on a comprehensive coding framework. This framework will list key coding categories (organized around the key interview domains) and their definitions, as well as provide examples of text from transcripts and videos where each code may apply. To increase the quality of the codings and the credibility of analyses, especially given the large number of interviews and that more than one coder is involved, the team will examine intercoder agreement. For this process, the team will independently code 10% of study transcripts (n = 48) and meet to assess agreement among coders, where the same sections of text are coded with the same codes, and any discrepancies are reconciled through discussion. The goal of these intercoder agreement meetings will be to iteratively refine and develop a coding framework and process that achieve high agreement among coders and, thus, reliably capture the range of key themes of interest in the interviews.

Once a final coding scheme is established, analysts will independently code the remaining transcripts. The task leader will hold regular meetings to discuss their coding progress, challenges and/or concerns, and any feedback on emerging themes.

Following completion of the coding process, a final codebook and report listing all codes and associated data (quotes) will be created and stored as an NVivo file. The task leader will review code reports and develop a report of key themes. We will also ground and clarify these interpretations of the data with exemplary quotations provided by participants.