U.S. FOOD & DRUG ADMINISTRATION

FDA's Response to External Peer Review of Quantitative Consumer Research on Front of Pack Nutrition Labeling, May 17, 2024

I. INTRODUCTION

The U.S. is continuing to face an epidemic of diet-related chronic diseases, many of which are experienced disproportionately by racial and ethnic minority groups, those with lower socioeconomic status, and those living in rural areas. To help address this problem, FDA has continued to prioritize its nutrition activities by leveraging its authority to help empower consumers with nutrition information they can use to make healthier choices more easily. FDA is focused on 1) creating a healthier food supply for all; 2) establishing a healthy start to set the foundation for a long, healthy life; and 3) empowering consumers through informative labeling and targeted education.

As part of its labeling efforts, FDA is conducting an experimental study to explore the establishment of a standardized, science-based front of package (FOP) nutrition labeling scheme that helps consumers, including those with lower nutrition literacy, quickly and easily identify how foods can be a part of a healthy diet.

Versar, Inc., an independent contractor, coordinated an external letter peer review of the experimental study. The peer review was conducted for FDA's Center for Food Safety and Applied Nutrition. For this peer review, three experts were selected by Versar, Inc. to evaluate and provide written comments on the format and content of the reports from the experiment, including the clarity of the documents, the presentation of the studies, and the scientific content of the study.

In Section II of this peer review response report, we list the charge questions given to the reviewers regarding the objective of the peer review and specific advice sought through the peer review. In Section III of this peer review response report, we provide a summary of the peer reviewers' comments followed by either a description of any changes made to the study report in response to peer reviewer comments or an explanation of our decision to not make suggested changes. For comments that came up in response to multiple charge questions, we have responded to that feedback in the most relevant charge question and indicate that the feedback was raised elsewhere. The individual peer reviewer comments are provided in tabular format in Appendix I.

Below are the names and affiliation of the peer reviewers:

Omni Cassidy, PhD
New York University Grossman School of Medicine/Langone Health

Jennifer L. Falbe, ScD, MPH
University of California, Davis

Joshua Petimar, ScD
Harvard Medical School/Harvard Pilgrim Health Care Institute

II. CHARGE TO REVIEWERS

Background for Reviewers

FDA is exploring the development of a front of package system to help consumers interpret the nutrient information on food products. Front of package (FOP) nutrition labeling is intended to complement the Nutrition Facts label by giving consumers a simple aid to provide additional context for making healthy food selections. As part of our food-labeling efforts, we are exploring the establishment of a standardized, science-based FOP scheme that helps consumers, particularly those with lower nutrition knowledge, quickly and easily identify foods that are part of a healthy dietary pattern.

A 15-minute online experiment was conducted in Fall of 2023 using members of a nonprobability-based consumer panel. Participants in the experiment were randomly assigned to (a) view and select the healthiest and least healthy of three nutrient profiles (least healthy, middle, most healthy) of a single FOP scheme and (b) respond to questions on product and scheme judgements for a single scheme on one of three food packages.

The document to be reviewed is the final report for the experiment.

It is important to note that the research is/was *not* intended to cover the following matters:

- comparing FOP schemes to other schemes found in the worldwide marketplace
- evaluating whether the schemes affect purchase intentions.

The peer review should provide input on clarity and, where appropriate, the soundness of the research design and analysis, and whether the conclusions reached are supported.

Reviewer Charge Questions

Clarity of the final report memo

1. Is the document logical and clear? If not, please provide suggestions that would make the document more logical and/or more clear.
2. Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results? If not, what specific additional information should we provide?

Scientific soundness of the methods used

3. Is the method appropriate for the purpose? Please provide overall impressions and thoughts about the method used.
4. Are the outcomes that are measured appropriate given the study's purpose? If not, are there other considerations?
5. Are the study participants that are included appropriate given the study's purpose? If not, please explain why.

Quality of the analysis/data

6. Is the analytic approach appropriate given the design and purpose of the study? Are there additional considerations for the approach FDA used?

Study conclusions

7. Are the conclusions drawn from the study supported by the data presented? If not, are there other considerations?
8. Please share any additional comments.

III. SUMMARY OF PEER REVIEWER COMMENTS AND FDA RESPONSE

1. Is the document logical and clear?

Summary of Comments:

Stated purpose of study needs revisions (e.g., replace "decisions" with "assessments" or "evaluations".) Reviewers' comments were mixed about whether the document was logical and clear, with one saying it was not and the others saying that it was. There were comments about inconsistent terminology used to talk about the nutrition knowledge grouping, requests for more detail about the schemes and nutrient profiles of the schemes, request for screenshots, requests to change a descriptive word, and a request for inclusion of strengths and limitations of the study.

FDA Response:

FDA appreciates the recommendations for ways to provide further clarity to the report. In response to the comments on clarity we made sure the terminology used to describe the nutrition knowledge groups was consistent throughout, we added information about where to find some of the information in the appendices and added task numbers to the headings to make clearer what part of the study we were discussing. We added screenshots of the study as seen from the point of view of the participant (for one of the study conditions) and added a discussion about the strengths and limitations of the study to the conclusion.

The stated purpose of the study was "to identify which FOP schemes enabled participants to make quicker and more accurate decisions about the healthfulness of a product without needing to consult additional nutrition information." FDA respectfully declines to replace "decisions" with "assessment" or "evaluations" in the stated purpose. Although we understand that assessments and evaluations may precede decisions, they sometimes do not, as noted in Daniel Kahneman's (2011) book, *Thinking Fast and Slow, (*Farrar, Straus and Giroux, publishers) in which decision-making can be thought of in terms of two systems: System 1 is fast, intuitive, and emotional; System 2 is slower, more deliberative, and more logical.

2. Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?

Summary of Comments:

Reviewers wanted more detail on the description of the study, including details about the cognitive interviews and pretest, justification for the study design and the models selected, and specific information about procedures such as why we excluded participants from the data analysis who completed the study in less than five minutes and why we disallowed the use of smart phones. Also included in the request for more information was a desire to have a table of the variables, the variable construction, and the associative question from the questionnaire, a discussion of the covariates, a request to see results from the no-scheme control, a desire for confidence intervals versus P values, insertion of results in the text where they were reported in the results section, and more information about statistical power. One reviewer requested that FDA follow the CONSORT reporting guidelines.

FDA Response:

FDA responded to the requests for more information by adding more detail about how the schemes were chosen including a more detailed description of the purpose of the study and more detail about the cognitive interviews and pretest. We provided an explanation for the covariate adjustments and added the scientific depiction of the statistical models.

References to the No-scheme control were removed from the report because, although the study included a no-scheme control, no comparisons were made with it because the questions about nutrient characterization were given only to those assigned to an FOP condition. When questions were specific to the FOP nutrition labeling scheme, those in the no-scheme control were skipped out. Moreover, the purpose of this study was to compare the three scheme categories and not assess the effects of the FOP nutrition labeling versus no FOP nutrition labeling, as those effects are already very well-established in the literature.

Tables of results had been provided and were referenced in the text. We believe it is not necessary to duplicate information that is in the tables, to mitigate redundancy and unnecessarily increase the length of the report.

We did not report results from the two High In schemes separately, like we did for the Nutrition Info Schemes, because they did not perform well in the analysis.

The researchers employed p-values rather than confidence intervals because p-value cutoffs allow for a clear decision-making standard which is expedient in a regulatory environment. Specifically, we did not employ confidence intervals to assess statistical significance because they have the potential to overlap, even with p-values smaller than .05. Furthermore, the false positive risk associated with the use of p-values was mitigated in this study using the Bonferroni adjustment, one of the more conservative approaches for multiplicity adjustments (Ref. 1,2, and 3).

Although we agree that the addition of information in various parts of the report could be helpful, we believe that using the CONSORT reporting guidelines will make the report difficult to follow. CONSORT guidelines, and many other popular guidelines for reporting study findings, are geared toward clinical research instead of social science research conducted to contribute to policy option discussions. There are major differences in purpose, design, and administration between clinical research and nonclinical research. For example, clinical studies generally have primary and secondary outcomes supported or refuted through hypotheses. The FDA FOP nutrition labeling experiment is exploratory to the extent that potential outcomes were unknown except as suggested by the scientific literature, and there were multiple dependent measures, each providing part of the picture that, when considered as a whole, would provide a snapshot of information for policy discussions.

3. <u>Is the method appropriate for the purpose?</u>

Summary of Comments:

One of the reviewers believed the study design was weak, wanting a completely different study with a different purpose and different measures; the other reviewers conveyed that the design was appropriate given the purpose. All three reviewers mentioned threats to internal validity because nutrient levels and other design elements were not completely crossed. Some concerns were expressed that the mock food labels were different from those found in the marketplace (less information and no examples of foods

that are clearly unhealthy). One reviewer was concerned that participants had to react to schemes without having any prior knowledge or education about them.

FDA Response:

The goal of the study was to test the ability of different schemes to provide simple, easy to access information. The study evaluated and compared the ability of various FOP nutrition labeling schemes to communicate information using a scenario that was as close to that of a real shopping experience, looking at schemes on a food package label. The scientific literature has shown that claims on the front of the food label have unintended effects, such as health halos (Ref. 4). To mitigate confounding explanations for the scheme effects, mock product labels were simplified to contain only enough information to convincingly represent a food product. In cognitive testing, no issues or problems arose in that regard.

By necessity, the nutrients included on the High In schemes needed to be "high in." The other schemes displayed nutrients that were low, med, and high in depending on the characterization of "least healthy", “middle healthfulness”, and "most healthy."

Moreover, the study did not test all design elements on all the scheme types. For example, the use of color, magnifying glass, interpretive language, and shape were not manipulated such that results could answer questions specific to these elements; Additionally, there were no GDA or Nutrition Info Schemes that displayed all high-in nutrients to limit. However, in a subsequent analysis including the least healthful nutrient profiles for the GDA and Nutrition Info black and white with no DV (two highs and a medium) and the healthiest nutrient profile for the High In scheme (one nutrient listed), results mirrored that of the full study; the Nutrition Info scheme outperformed the GDA and High In schemes. It is important to note that the main purpose of this study was to compare the different categories of schemes with each other and not to test components of the schemes. The literature on schemes displaying nutrient summaries, interpretive information, and warnings is clear; interpretive schemes do best for conveying an understanding of nutrient content.

Participants were asked to respond to schemes for which they had no previous experience or information. However, although education and experience may be helpful for participant understanding of the scheme, there is no guarantee in the real world that consumers will have seen instructional materials about the scheme or have any familiarity with it. It was important to assess the degree to which the schemes communicated to consumers without having prior knowledge or experience with them.

Not showing the schemes on food products for the first task, where participants were asked to identify the healthiest and least healthy nutrient profile, could be seen as a study limitation because the schemes were presented outside of the food label context, however the potential for out-of-context confusion was tested in cognitive interviews and participants conveyed that they had no difficulty understanding what was being asked of them.

4. <u>Are the outcomes that are measured appropriate given the study’s purpose?</u>

Summary of Comments:

One of the three reviewers believed the outcome measures were not appropriate and questioned their validity, recommending different outcome variables. The other reviewers believed the outcome

measures were appropriate but had some clarifying questions. One reviewer wondered how participants interpreted “overall nutrient profile” from the questions in Task 1 and was concerned that the attitude and perception measures had limited usefulness if participants did not have previous exposure to them or knowledge about them.

FDA Response:

Regarding outcome measures, the purpose of the FOP experiment was to assess which scheme type would be most helpful for providing information to consumers to help them be able to construct a healthy dietary pattern. This is stated clearly throughout the report. FDA developed many of the questions and thoroughly tested them in cognitive interviews to assess whether participants could correctly interpret them and easily select a response. Only when this was the case were the measures used. Cognitive interviews increased the validity of the measures.

The study was designed to assess multiple outcomes that, when considered together, would provide a holistic overview of consumer reaction to the tested schemes. Important outcomes included ability to use the scheme to 1) identify the most healthful nutrient profile (without referring to the Nutrition Facts label), 2) lessen time spent to complete the task, and 3) characterize the level of nutrient. Multiple measures on consumers’ attitudes and perceptions of the tested schemes were also assessed

The phrase "overall nutrient profile" was carefully tested in the cognitive interviews because researchers had concerns about what the phrase would mean to participants and whether the question’s instructions were clear. As part of the question, participants were instructed to use the information available on the scheme to determine “overall nutrient profile.” The cognitive interviews revealed that participants had no problem understanding the meaning of the question or the response options and understood what was being asked of them.

5. Are the study participants that are included appropriate given the study’s purpose?

Summary of Comments:

The reviewers generally agreed that the study participants were appropriate although one reviewer preferred geographical representation and the addition of non-English speakers to the sample. Another reviewer wanted details about the cooperation rate.

FDA Response:

We added information about geographical location of the participants to the table describing the participants demographics (Table 1 in the report). We did not have a non-English version of the study instrument because currently all required label statements must appear in English.

We added text about the AAPOR Cooperation Rate 1, which is the most conservative cooperation rate.

6. Is the analytic approach appropriate given the design and purpose of the study?

Summary of Comments:

Reviewers had some comments about the data analysis. Two of the reviewers questioned having covariates in the models and, although these were described in the report, wanted to see the specific

models that were tested. One of the reviewers wanted to see confidence intervals versus P values as indicators of statistical significance. One reviewer thought the interactions should have been tested one at a time instead of all in the same model. One reviewer questioned our employment of the Bonferroni adjustment, and another asked that the Y axes on the figures be adjusted to show 0% to 100%. One reviewer asked if schemes were compared both within and between subjects for the comparison task and asked why the means for the expanded Nutrition Info schemes from Table 5 were incongruent with the combined Nutrition Info schemes mean from Table 2. One reviewer was concerned about the sample size distribution when the schemes were collapsed into the three scheme categories. A reviewer was concerned that not all design elements were tested for all schemes.

FDA Response:

The question about the Bonferroni adjustment and confidence intervals was addressed in response to Charge Question 2 and the question about design elements was addressed in response to Charge Question 3.

In response to the comments, we provided a brief explanation to the report about why covariates were added to the analysis; we added "Covariate adjustments were included in the models to mitigate any potential covariate imbalances."

Additionally, randomization to U.S. (Census) population benchmarks ensures representativeness of the U.S. population but can also introduce covariate imbalance (e.g., 60.5% Non-Hispanic White vs. 11.6% Non-Hispanic Black). When covariate imbalance is present at baseline, as in race ethnicity imbalance in the U.S. population, including covariates in data analyses is even more important. As Moerbeek and Schie (2016, Ref. 5) state, “ignoring relevant covariates while analyzing the data may lead to severely biased estimates of the treatment effect and its standard error…All relevant covariates should be … included in the statistical model to avoid severe levels of parameter and standard error bias and insufficient power levels.”

The majority of covariate variables adjusted for in our analysis presented imbalances: e.g., rurality: 80% urban, 20% rural; age: 33.6% 30-49 yrs. Although the covariate variables gender and education did not present severe imbalance, they were included per FDA and scientific literature guidelines, some of which are stated below.

The authors adjusted for the covariate variables (rurality, age, gender, etc.) in accordance with the FDA guidance, updated in May 2023, (https://www.fda.gov/drugs/drug-safety-and-availability/fda-issues-final-guidance-adjusting-covariates-randomized-clinical-trials), which states that “adjusting for covariations in randomized studies can result in narrower confidence intervals and a greater statistical power to detect outcome effects.”

There is an abundance of recent peer-reviewed publications advocating for covariate adjustment.

Per Holmberg and Andersen (2022, Ref. 6), adjusting for characteristics in the analysis of randomized clinical trials “is advised by both the European Medicines Agency and the US Food and Drug Administration because it may improve statistical efficiency, enhancing the ability to draw a reliable conclusion from the available data.” They further state that “By accounting for factors influencing the outcome other than the randomly assigned intervention, adjustment leads to increased statistical power (i.e., the ability to detect a treatment effect when present).”

Kahen et al (2014, Ref. 7) state that “Adjustment for known prognostic covariates can lead to substantial increases in power, and should be routinely incorporated into the analysis of randomized trials.”

Per Nicholas et al (2016; Ref. 8), “failure to adjust for covariates that influence outcome in the analysis phase, regardless of prior adjustment at randomization, results in treatment estimates that are biased toward zero, with standard errors that are deflated.”

The rationale for simultaneous inclusion of the interaction terms in the model is two-fold: 1) Interaction effects can modulate each other. Finding significance for an interaction in a one-at-a-time approach may be spurious if that interaction term depends on the value of other interaction terms; i.e., model endogeneity. The significance for an interaction term in the one-at-a-time approach may be attenuated or disappear altogether when the terms are tested simultaneously. 2) To avoid biasing the model, the researchers found it prudent to examine all factors and their interactions in a global environment to avoid reporting biased results. As Garofalo et al (2022; Ref. 9) state “To provide a complete interpretation, it is essential to observe how values are modulated by all levels of the factors simultaneously.” Unless a high degree of multicollinearity is a concern, omitting simultaneous factors and their interaction effects may lead to biased results (Semadeni et al, 2013; Ref. 10). Although we recognize that a full model risks over-fitting the data and thereby inflating variances, the study was strongly powered to detect even small effect sizes and as such any loss of power accompanying model overfitting was not a major concern and a smaller concern than reporting biased results.

Regarding the sample size of the collapsed scheme categories, although we recognize that a full model risks over-fitting the data and thereby inflating variances, the study was strongly powered to detect even small effect sizes and as such any loss of power accompanying model overfitting was not a major concern and a smaller concern than reporting biased results.

Regarding within and between comparisons, between-subjects factors are a component of a repeated measures study. The subjects were measured in three conditions (three schemes) to comprise the within-subjects component but there were still distinct groups (label schemes) across all subjects which comprise the between-subjects component.

Regarding the Y axes labels on figures, we agree that it would be preferable for the axes to reflect 0% to 100%, but the statistical program we used (IBM SPSS) does not have a way to adjust the labeled Y axes; the axes are currently clearly labeled.

The scientific nomenclature for the specific models that were tested in the study were added to the report.

When investigating the seeming incongruity between the means for the expanded Nutrition Info schemes in Table 5 and the combined mean for the Nutrition Info schemes in Table 2, we discovered an error in our data analysis syntax for the expanded Nutrition Info analysis for Task 1. Participants in Task 1 of the study were randomized to three schemes from the full set of schemes in the study. For the subset analyses of only the Nutrition Info schemes, not each respondent will end up with three repeated schemes in this subset. Therefore, the design for the subset analysis can no longer be treated as repeated measures; we corrected the syntax such that the data was treated as fully independent across both schemes and respondents. This resulted in small changes to the means for the expanded Nutrition Info schemes – but no differences in overall trends. The tables in the report have been updated.

With respect to the reviewer's comment about the incongruity between the means in Table 2 and Table 5, the means for the expanded Nutrition Info scheme (Table 5) more closely align with the mean for combined Nutrition Info schemes from Table 2 after revisions, although not exactly. The reason for the difference is that these are estimated marginal means, meaning they reflect the covariate controls, the standard errors of which are affected by sample size. The sample sizes for Table 2 are different from Table 5, because Table 5 represents a subset of the data. We added an explanation about the sample size differences to the report and added sample size data to the tables.

7. Are the conclusions drawn from the study supported by the data presented?

Summary of Comments:

One of the reviewers requested that FDA specify primary and secondary outcomes, and two reviewers requested a discussion of study limitations. A reviewer wanted the addition of an outcome measure not already included in the study. Two of the reviewers wanted a discussion of statistical versus clinical significance. One reviewer wanted to see the findings from the no-scheme control. Two reviewers mentioned overgeneralizations and inaccuracies of the study findings in the abstract and conclusion section. One reviewer wanted study limitations added to the report. Another wanted more discussion about the differences between Exhibits 5 and 6 [there is no Exhibit 6 so it was assumed the reviewer meant Exhibits 4 and 5].

FDA Response:

The question about the no-scheme control was addressed in response to Charge Question 2 and primary outcomes was addressed in response to Charge Question 4.

Study strengths and limitations were added to the conclusion section of the report.

The study conclusions were supported by the data. The purpose of the study was to assess the ability of three types of schemes to provide information to consumers to help them be able to construct a healthy dietary pattern. However, some of the descriptive language in the report was clarified to reflect the findings more precisely: "Most" was added to a sentence implying the attitude question results for the Nutrition Infos were true for all schemes. "…Much more difficult…" was revised to "…more difficult…" in a sentence describing the scheme category results. Also, a sentence stating "…both the GDA and High In schemes performed poorly…" was revised to "…both the GDA and High In schemes did not perform as well…"

We disagree that a discussion of "clinical" significance versus "practical" significance is needed in the report. The purpose of the study was to help policymakers understand how well the three types of schemes communicate to consumers about the nutrients to limit, and all outcome measures are considered in policy discussions. There are notes beneath the tables of the subgroup analysis indicating the status of statistical significance.

The discussion section of the report goes into detail about Exhibits 4 and 5; therefore, we chose not to add more information on the differences between the Exhibits

References for FDA Response

1. Greenland, S., Senn, S.J., Rothman, K.J., et al. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." *European Journal of Epidemiology*. 2016;31(4):337-350. http://doi: 10.1007/s10654-016-0149-3.

2. Ioannidis, J.P.A. "The importance of predefined rules and prespecified statistical analyses: do not abandon significance." *Journal of the American Medical Association.* 2019;321(21):2067-2068. http://doi: 10.1001/jama.2019.4582.

3. Tan, S.H., Tan, S.B. "The correct interpretation of confidence intervals." *Proceedings of Singapore Healthcare.* 2010;19(3):276-278. http://doi:10.1177/201010581001900316.

4. Roe, B., Levy, A.S., Derby, B.M. "The Impact of health claims on consumer search and product evaluation outcomes: results from FDA experimental data." *Journal of Public Policy and Marketing.* 1999;18(1). https://doi.org/10.1177/0743915699018001.

5. Moerbeek, M., van Schie, S. "How large are the consequences of covariate imbalance in cluster randomized trials: a simulation study with a continuous outcome and a binary covariate at the cluster level." *BMC Medical Research Methodology*. 2016;16(79). https://doi.org/10.1186/s12874-016-0182-7.

6. Holmberg, M.J., Andersen, L.W. Adjustment for baseline characteristics in randomized clinical trials. *Journal of the American Medical Association.* 2022;328(21):2155-2156. http://doi: 10.1001/jama.2022.21506.

7. Kahan, B.C., Jairath, V., Doré, C.J., et al. "The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies." *Trials*. 2014;15(139). https://doi.org/10.1186/1745-6215-15-139.

8. Nicholas, K., Yeatts, S.D., Zha, W., et al. "The impact of covariate adjustment at randomization and analysis for binary outcomes: understanding differences between superiority and noninferiority trials." *Statistics in Medicine.* 2015;34(11):1834-40. http://doi: 10.1002/sim.6447.

9. Garofalo, S., Giovagnoli, S., Orsoni, M., et al. "Interaction effect: Are you doing the right thing?" *Public Library of Science One*. 2022;17(7). http://doi: 10.1371/journal.pone.0271668.

10. Semadeni, M., Withers, M.C., Certo, S.T. "The perils of endogeneity and instrumental variables in strategy research: understanding through simulations." *Strategic Management Journal*. 2014; 35(7):1070–1079. http://www.jstor.org/stable/24037334

APPENDIX I – INDIVIDUAL PEER REVIEWER COMMENTS

A. CHARGE QUESTIONS

1. Clarity of the report memo

CHARGE QUESTION 1. Is the document logical and clear? If not, please provide suggestions that would make the document more logical and/or more clear.

REVIEWER	COMMENT
Reviewer #1	The logic of the document is adequate, but the clarity is poor. There are several examples of confusing sentences that make it difficult to know what was done (see Specific Observations below). One of the main issues is that there is no prespecified primary outcome. As a result, it is difficult to know which results are the most important to consider when weighing all of the evidence. This creates problems for drawing conclusions (see my response to Q7 below). I strongly suggest that the authors include information about what the primary and secondary outcomes were and use those to inform conclusions about the study.
Reviewer #1	Also, different terminology is used to describe the same thing at different points of the document. For example, the Methods and Results sections refer to stratified analyses by "nutrition knowledge/motivation", but the corresponding figure (Exhibit 2) describes this as "High Nutrition Literacy." These are not the same thing, and it creates confusion to have a single concept described in these two different ways. Another example is in the last paragraph of the Abstract, where the Nutrition Info scheme is referred to as the "High/Medium/Low" scheme. I strongly suggest that the authors revise the document to use consistent language to describe all treatments, outcomes, subgroups, etc.
Reviewer #2	The overall flow and organization of the document is clear. However, there are an extraordinary number of groups and combinations, so careful attention to the labeling and description throughout would be helpful. For instance, it was a little difficult to follow the differences between the three *major* FOP scheme versus the eight FOP scheme subtypes. I suggest explicitly stating that while there are three major schemes, there are actually eight different schemes being tested. Although FDA eventually states this, they also mention throughout that they are only testing three different designs, which makes it a bit confusing. It would be helpful to include the illustration of the FOP Schemes Tested that the FDA presented during the kick-off meeting. It was easier to follow the eight experimental groups than Appendix A.
Reviewer #3	The document is generally logical, clear, and well-written. The document would benefit from including an Appendix with screenshots of all questions seen by participants, including clarifying text about what treatments were randomized.

REVIEWER	COMMENT
Reviewer #3	The stated purpose was "to identify which FOP schemes enabled participants to make quicker and more accurate decisions about the healthfulness of a product without needing to consult additional nutrition information." The word "decisions" implies purchasing, selection, or consumption decisions, which were not assessed. Based on the outcomes assessed, I recommend replacing "decisions" with "assessments" or "evaluations," or using the term "knowledge." The stated purpose should also include a statement that the study is testing FOP schemes *without* consumer education about how to use the schemes. This is important to add in the introduction and conclusions sections because making "accurate" and especially "quick" decisions without consulting the NFL hinges on whether consumers know how to use the FOP schemes, which in absence of education, is related to prior exposure to a scheme.
Reviewer #3	Clarity would benefit from explicitly stating which outcomes were primary and which were secondary or exploratory. Lastly, the study would benefit from a discussion of limitations in the conclusions.

CHARGE QUESTION 2. Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results? If not, what specific additional information should we provide?

REVIEWER	COMMENT
Reviewer #1	There is not sufficient information provided about several of these aspects of the study. The overall presentation of the Methods and Results is unsatisfactory. I have specific comments below, but in general the authors should consult the CONSORT guidelines to ensure that they are including sufficient information about the study.
Reviewer #1	One of the biggest issues is the lack of information about a primary outcome, which is necessary to determine whether the study objectives were met. Details are also missing on the construction of the outcomes. For example, there are no details on how "knowledge of nutrient content" was measured, even though this appears to be a main outcome. This should be provided up front in the main text, but it is not even provided in the Appendix. I had to look at the questionnaire to find the questions that corresponded to these outcomes, but it was not at all clear to me that I was looking at the right questions. This only became clear to me during the kickoff call with the FDA, where they explained that answers of 1-2 were low, 3-4 were medium, and 5-6 were high. This needs to be laid out explicitly. Details on the construction of all variables need to be written out explicitly. When describing the outcome construction, it would also be helpful to refer to the exact questions so that readers can find this information easily.

REVIEWER	COMMENT
Reviewer #1	The document is insufficiently transparent about how FDA chose which labeling schemes to test. It refers to completion of a literature review and focus groups, but the results of these activities are not presented here. Similarly, on page 7, the document refers to excluding participants from the main experiment if they participated in cognitive interviews or a pretest, but there are no other mentions of either of these throughout the document. These are potentially important details that the readers have not been given adequate information about, and they should be provided at least in the Appendix for transparency. Similarly, the authors should explicitly list all of the design features that were considered, such as colors, shapes, and other elements. Evidence from other studies (both laboratory and real-world) demonstrates that features like an octagon shape and the word "warning" are effective in improving customer understanding of the nutritional profile of unhealthy foods, yet these were not included among the FOP labeling schemes. There needs to be rationale for why the FDA chose certain labeling designs to be tested and why others with strong evidence for improving customer nutrition knowledge and behavior were omitted.
Reviewer #1	Justification is lacking for several decisions that were made throughout the study. For example, why did the authors decide to exclude those who completed the questionnaire in less than five minutes? Why did they exclude people who did the study on a smartphone? I cannot come up with a good reason why the authors would exclude those taking the study on a smartphone, and without sufficient justification, this decision seems arbitrary. If the authors have a valid scientific reason for this decision (e.g., it presented a threat to internal or external validity), they should state it. They should also present the number (%) of participants who were excluded from the main experiment for each exclusion criteria.
Reviewer #1	Similarly, the authors do not present rationale for many of their analytic decisions. It is not clear why they adjusted for so many variables in their main model (rurality, age, gender, etc.). Is this because there was imbalance between treatment arms? Typically, you would not need to adjust for all of these variables in a randomized trial, so this decision needs justification. It would be helpful to explicitly write out the statistical model that was used, including how each term was defined (e.g., was age included in the model as a continuous term? In categories?). Another example is when the authors state that they adjusted for whether a person had been assigned to a scheme that was viewed in the first part of the study; again, there is no reason that is given to justify this decision. Lastly, the authors used a modified Bonferroni adjustment, but more details are needed on how this was calculated and what the rationale was.
Reviewer #1	Many important details are missing from the Results section, and I found nearly the entire reporting of the results to be unscientific. First, the authors do not present a table to demonstrate whether or not randomization was successful (i.e., comparing the distribution of covariates across treatment arms). This is a basic principle of reporting results from a randomized trial (see the CONSORT guidelines).

REVIEWER	COMMENT
Reviewer #1	Second, in the Single Product Evaluation Task, it is unclear to me why the authors did not report results from the control group. It would be extremely useful to know whether these schemes led to more accurate characterization of nutrient levels than the no-label control, and if so, by how much. I strongly recommend that the authors include results from the no-label control and report differences in the outcomes between each of the schemes and the control group (in addition to differences between schemes).
Reviewer #1	Third, there is unsatisfactory reporting of the differences between the schemes or between subgroups. For example, in section C1a, the text states: *"Compared to those with higher nutrition knowledge/motivation, those with lower nutrition knowledge/motivation were slightly less likely to correctly answer the question if assigned to the GDA or the Nutrition Info schemes but appreciably less likely to correctly answer if assigned to the High In schemes."* It is unclear how the authors determined "slightly less likely" versus "appreciably less likely." Are these based on p-values from the interaction terms of the model? This language is not scientific, which makes it impossible to determine the importance of these differences. This also applies to the reporting of differences by rurality: *"Participants with rural residency were slightly more likely to correctly select the healthiest nutrient profile if assigned to the GDA scheme and to the Nutrition Info schemes but slightly less likely than nonrural residents to correctly answer if assigned to the High In schemes."* Another example of this is in C1b, where the document states: *"those who viewed the least healthy, versus the middle or healthiest, nutrient profile, were far less likely to correctly characterize the level of saturated fat when viewing the GDA scheme, more likely when viewing the Nutrition Info schemes, and less likely when viewing the High In schemes."* Again, the reader needs to see the actual numbers to know how different these were from each other. Saying "far less likely", "less likely", and "more likely" is not transparent and makes it impossible to interpret the results. It is also unclear what these are all in relation to (far less likely than what?). Overall, the authors need to revise the document to use purely scientific standards for reporting differences.
Reviewer #1	Fourth, there are inconsistencies in what results the authors have chosen to report and what they have chosen not to report. It is not clear to me why the authors chose not to compare the two High In schemes (with/without %DV) on ability to correctly identify the healthiest and least healthy nutrient profile, likelihood of clicking to see the NFL, and the total amount of time spent responding to the questions. They do this in C2a for the Nutrition Info schemes, so it is confusing as to why this was not done for the High In schemes. Similarly, Exhibits 4 and 5 present interactions between the scheme categories and nutrient profile for correctly characterizing levels of saturated fat and sodium, respectively, but they do not include results for added sugar. They should make this addition to be consistent and transparent. Further, Exhibits 2 and 3 display interactive effects between the scheme categories and rurality/nutrition knowledge on ability to select the healthiest profile. I recommend the authors do this for ability to select the unhealthiest profile and report those results.
Reviewer #1	Fifth, the text of the Results section does not provide any of the actual quantitative findings, but these should be added (as opposed to solely referring to the exhibits).

REVIEWER	COMMENT
Reviewer #1	Lastly, the authors report differences according to p-value cutoffs, but in my opinion, it would be better to present 95% confidence intervals in addition to p-values. 95% CIs give a much better sense of the likely effect estimate than p-values alone given that p-values are a function of sample size. Large studies can therefore report statistically significant p-values for clinically unimportant differences, whereas small studies can fail to find statistically significant differences even if there are large, clinically important differences between treatments. I would be particularly interested in knowing the 95% CI around the difference in outcomes between the schemes, as opposed to just highlighting that the difference between the outcomes is statistically significant.
Reviewer #2	In the introduction, the FDA distinguished between noninterpretative, interpretative, nutrient specific, and summary schemes, but do not extend this labeling to their FOP schemes. It would be helpful to see which of these labels the test FOP schemes fell into. Within the study design, although the FDA indicated they test the nutrition info with *and without* a magnifying glass, the NO magnifying glass scheme is just reflected in the fact the none of the other schemes have a magnifying glass. It would be clearer to simply state the scheme as "nutrition info with magnifying glass." In the analysis section, once authors start describing the variables, I suggest listing out all eight schemes, rather than only discussing three. (similar to comment in #1). Although authors chose to collapse all nutrition info schemes into one large Nutrition info category for ease of interpretation, the original groups should be clear throughout.
Reviewer #3	Stimuli: A table or figure is needed in the main text that describes the nutrient profiles used in each condition and their nutrient contents. Also, it is not stated whether the labeling schemes were all the same size (in square inches or cm).
Reviewer #3	Methods: When describing the outcomes assessed, it would help to include exact wording, response options, and coding of responses for primary outcome measures. Specifically, the coding of the 6-point response scale to the "correctly characterizing the level of nutrients" questions should be described. For outcome measures, evidence of psychometric testing, cognitive interviewing, or other rationale for inclusion should be reported.
Reviewer #3	Analysis: Regarding statistical power, the report provides a nice power analysis. It would help to also describe the extent to which there was adequate power to detect the modest but meaningful differences between the 6 Nutrition Info schemes (e.g., color, magnifying glass).
Reviewer #3	Results: The methods mention cognitive interviews and a pre-test, but results of these are not described. The report also mentions a no scheme control condition, but no results are provided for that condition. Additionally, the results should describe how the 2 High In schemes compared to one another. Lastly, given the likely limitations on study power (after accounting for multiple testing) to detect modest but meaningful differences between the 6 Nutrition Info schemes, the narrative results section should avoid only stating "There were no significant differences..." It should report on the magnitude of the nonsignificant differences.

2. Scientific soundness of the methods used

CHARGE QUESTION 3. Is the method appropriate for the purpose? Please provide overall impressions and thoughts about the method used.

REVIEWER	COMMENT
Reviewer #1	The methods are very limited and are not appropriate to determine which labeling scheme would best help consumers either identify or purchase healthier packaged items.
Reviewer #1	The overall study design is weak. Both the FOP Comparison Task and the Single Product Evaluation Task simply measure participants' abilities to classify the label (or products with the label) by overall level of nutrient content, which has limited utility and was possibly measured with an invalid instrument (see comments to Q4 below). It would have been much better to present participants with several different items to choose from (e.g., in an online store) and ask them which they intend to purchase (using the different labeling schemes). For example, if participants had been presented with, say, 20 options for frozen foods and asked to select up to 3 to purchase, then investigators could have compared the actual nutritional quality of the items purchased under different labeling schemes and the no-label control. The results of that kind of study would much more likely reflect real-world behavior than assessing how accurately participants can report whether a single product has low, medium, or high levels of nutrients of interest. The current study therefore provides no evidence on the effects of these labeling schemes on customer food selection or purchasing intentions. There is an assumption being made that if customers can identify what is healthiest, they will make healthier choices. It would have been better to test this directly rather than to rely on this very strong (and possibly false) assumption.
Reviewer #1	There are also major limitations with the choice of food products presented to the participants that reduce the generalizability of the findings. First, the authors used only 3 products – a cereal, a frozen entrée, and a canned soup. These products represent only 3 categories, and results from this study might not generalize to other categories. The fact that there was only a single option for each category is also highly limiting because it ignores variability in the types of products that are found in each category. Relatedly, all of these products look like healthy products, or at the very least are what most consumers think healthy products look like (a vegetable grain bowl, a can of bean soup, and a box of oat flakes). It would be much more useful to examine the effects of the different labeling schemes on both healthier and less healthy groups (e.g., chips, cookies, candy, sweetened beverages, etc.). Second, the products are all mock products and are not representative of what actual products look like. Many products have much more front-of-pack marketing than what is depicted on these examples (e.g., cartoon characters on cereal boxes), as well as more colors, images, pictures of celebrities, etc. As a result, it is possible that the results of this study will not generalize to real-world products where the labels will be competing for customers' attention with other FOP imagery and text. Many studies have used graphic design to alter real-world products to test labeling interventions, and that approach should have been taken here. Third, there are no children's products represented here, which again lends itself to poor external validity, especially given the amount of FOP marketing on those products in particular. Moreover, adults' reactions to labeling schemes may differ if they think they are purchasing products for themselves or their children. The fact that this was not tested is an unfortunate missed opportunity.

REVIEWER	COMMENT
Reviewer #1	The stimuli also have limitations. First, the High In labels with %DV have a different %DV than the other labels (e.g., for the healthiest condition, 15% added sugars for the Nutrition Info and GDA labels; 22% for the High In labels), which makes them not comparable. This means that results from the Single Product Evaluation Task could partially be driven by differences not only in the labeling scheme, but also by differences in nutrient composition. This could be particularly important for participants who reported monitoring added sugars, though my understanding is that the analysis did not take this into account.
Reviewer #1	Second, for all schemes, the %DV of added sugar is the same across all health categories and the %DV of sodium is increasing. Therefore, in this study, added sugar is not associated with overall product healthfulness while sodium is. This could create a problem in the FOP Comparison Task, where participants are asked to choose the healthiest item. The accuracy of a person's response to this question might therefore depend on whether they monitor any particular nutrient. If they especially value products with low added sugar, their responses to the healthfulness task might not be as accurate because added sugar is not associated with healthfulness for those schemes. If they value products with low sodium content, their answers might be much more accurate. Given the randomized nature of the study, this might not create bias, though it could add statistical noise to the results.
Reviewer #1	Third, there was no option for a High In label with color. Color is shown to be an important feature of labels that can influence behavior, and the fact that color was used in some of the Nutrition Info schemes but not the High In schemes limits inferences that can be made when comparing the labeling scheme types. If the results comparing GDA vs. Nutrition Info vs. High In are pooling across all of the different labeling types within each scheme, it makes it hard to know whether one labeling scheme is actually superior to another, or whether the results are due to that labeling scheme having alternate versions with different designs (such as color).
Reviewer #1	Fourth, it is very strange to me that the authors selected only one of the schemes (Nutrition Info BW no %DV) to test in the lower right corner of the package. This is insufficient to determine whether placement in the upper or lower corner is better because the relationship between label location and selection could differ by scheme, and it seems arbitrary that they chose only this scheme to do this test. It would have been better to do this across all of the schemes. Additionally, it is problematic that they used a single Bonferroni adjusted p-value cutoff to compare different labeling schemes to each other *as well as* to compare the Nutrition Info BW no %DV labels in different locations. These are different "families" of tests, and it is inappropriate to apply the same p-value cutoff across all of them. In my opinion, results from the "BW, No %DV in lower right of package" should be excluded from the tables with other labeling schemes (e.g., in Tables 6, 7, and 7a).
Reviewer #2	The single product condition likely would've benefited from also having products that did not appear healthy just as comparison. I also think it might've been helpful to remove the marketing statements (e.g., "100% whole oats"). The other comparison groups appeared appropriate.

REVIEWER	COMMENT
Reviewer #3	Several aspects of the design were appropriate for the purpose, including the use of a randomized experiment, a task in which different nutrient profiles were viewed side-by-side, and the use of three different food products (instead of just one) in the second task. It was an excellent decision to not test schemes that contained both nutrients to limit and nutrients to get enough of due to consumer confusion about such labels and because companies already promote positive attributes of their products.
Reviewer #3	There are limitations to external validity that should be acknowledge. The first is that no conditions were tested in which participants were told how to use and interpret the schemes, which may be especially important for the High In schemes that are less familiar in the U.S. Further, the information about the labels given to participants did not explain that all packages would be required to have such labels (vs. voluntary use), and it mentioned that the labels were intended to help "identify foods that are part of a healthy eating pattern," which may be especially confusing for participants viewing High In labels that indicate only unhealthy levels of nutrients. Future studies testing different schemes should include conditions in which participants are instructed on how to use the labeling schemes, which mirrors real-world education that would accompany mandatory labeling. Second, in the Comparison Task, participants viewed the label schemes in isolation (not on a product package). The ability of a label to grab attention is important for a label to have its intended effect in the real-world, which the Comparison Task could not assess by virtue of showing only the label close up.
Reviewer #3	I have 2 main concerns about internal validity, particularly regarding the ability to draw conclusions about the performance of High In schemes compared to GDA and Nutrition Info schemes. The first concern is that the High In schemes were confounded by having different nutrient contents for each of the 3 nutrient profiles shown. The GDA and Nutrition Info schemes had identical nutrient contents: • Healthiest: 0 nutrients were "high"; 4% DV sat fat, 4% DV sodium, 15% DV added sugars • Middle: 0 nutrients were "high"; 4% DV sat fat, 15% DV sodium, 15% DV added sugars • Least healthy: 2 nutrients (sat fat and sodium) were "high"; 25% DV sat fat, 25% DV sodium, and 15% DV added sugars) In contrast, all the High In nutrient profiles were objectively less healthy in terms of being high in these nutrients: • Healthiest: 1 nutrient (added sugar) was "high"; 22% DV added sugar • Middle: 2 nutrients (added sugar and sodium) were "high"; 22% and 21% DVs • Least healthy: All 3 nutrients were "high": 25%, 25%, 22% DVs Without identical nutrient profiles for each scheme, it is uncertain if the differences between High In and the other scheme categories were due to scheme design or the differences in nutrient profiles across conditions. For this reason, I would consider dropping the High In conditions from the primary analyses comparing scheme groups. Instead, the High In responses can be used to determine which High In label performed better. At a minimum, the use of different nutrient profiles for the High In schemes should be clearly disclosed with a rationale provided, and it should be mentioned as a limitation both when describing results and in the conclusion.

REVIEWER	COMMENT
Reviewer #3	My second concern about internal validity relates to how participants were randomized to view 3 of the 8 labeling schemes in the Comparison Task. The concern is that 5 out of 8 schemes were Nutrition Info Schemes. Thus, by virtue of participants having a higher likelihood of repeated exposure to the Nutrition Info schemes, participants are expected to score better on outcomes related to understanding and use of the Nutrition Info schemes vs. the GDA or High In schemes. To address this concern, the analysis comparing scheme categories from the Comparison Task (Table 2) could be restricted to the first scheme participants saw. However, results from all 3 labels each participant saw can be used to compare schemes within scheme category (e.g., Table 5).
Reviewer #3	An additional limitation is that only the Nutrition Info scheme was presented with the magnifying glass and color variations. From the perspective of statistical power, it makes sense to limit the number of schemes tested. However, the analysis comparing the scheme categories (e.g., Tables 2, 3, 4) should exclude the magnifying glass and color variations from the Nutrition Info scheme category. Otherwise, the comparison of scheme categories is confounded by these attributes not shared by the other labeling scheme categories.

CHARGE QUESTION 4. Are the outcomes that are measured appropriate given the study's purpose? If not, are there other considerations?

REVIEWER	COMMENT
Reviewer #1	Several of the outcomes are not appropriate and in some cases have questionable validity. As mentioned in Q3 above, it would have been better to directly test customer selection (e.g., purchase intentions) of packaged products, rather than the much weaker outcome of perceptions of healthfulness/recall of nutrient content levels. Item selection would more closely approximate real-world conditions and generate much more convincing evidence in favor of a particular scheme. It is very possible that customers in the real world can recognize the healthfulness of products under different kinds of labeling schemes, but that some would influence purchases to a greater extent than others.
Reviewer #1	In the Single Product Evaluation Task, the outcome of "knowledge of nutrient content" was measured using a scale with questionable validity. The scale goes from 1 to 6 and does not contain labels (except 1 as Low and 6 as High). My understanding from the kickoff call (i.e., not listed in the text) is that 1-2 were considered unhealthy, 3-4 were moderately healthy, and 5-6 were healthy. Why did the investigators choose to use this scale and grouping? Has it been validated in previous studies? If the answer is yes, this needs to be written out explicitly with appropriate citations. If not, then it is unclear whether the results are valid, or how they should be interpreted. To this point, the investigators found that participants were much less likely to correctly identify products with a "middle nutrient profile," but it is unclear whether this is a true phenomenon or whether it is due to using an invalid and incorrect measure of assessing nutrient profiles. It is not even clear to me whether it is useful for customers to be able to distinguish products by these 3 categories (low/medium/high); it likely would have been more useful to understand whether customers are more likely to select the healthiest option when presented with multiple options the way they are in actual retail settings.

REVIEWER	COMMENT
Reviewer # 1	The attitude and perception questions overall seem reasonable to me, but it is unfortunate that the investigators did not ask questions about intentions, which may be better predictors of behavior. Again, the authors need to include citations to support the validity of the questions they used, or else it is unclear whether the findings are valid.
Reviewer # 2	Outcomes appear to be appropriate.
Reviewer #3	The outcome related to "correctly characterizing the level of nutrient" in the Single Product Task (measured by "How low or high is this product in the following nutrients?") is an appropriate outcome given the purpose. It directly and unambiguously assessed whether a consumer viewing a product with the label could evaluate how high a product was in each of the 3 nutrients on the FOP labeling schemes. This should be the primary outcome given the study's purpose.
Reviewer #3	My main questions pertain to the outcome measures in the Comparison Task, "Which one of the three Front of Package nutrition labels shows the (healthiest/least healthy) overall nutrient profile?" Was there evidence from psychometric testing or cognitive interviews about how participants interpreted "overall nutrient profile"? How did FDA intend for participants to interpret "overall nutrient profile"? Specifically, did participants (and did the FDA want participants to) think of only sodium, saturated fat, and added sugars, or did participants also think about nutrients to encourage like fiber, vitamins, and minerals? Did this perception differ based on exposure to the GDA, Nutrition Info, or High In schemes? And what information did participants think they needed to evaluate an "overall nutrient profile"? Did they believe they needed quantitative information? The different results for how the GDA and High In schemes performed on "overall nutrient profile" outcomes vs. the "correctly characterizing the level of nutrient" outcomes suggests that consumers may not have considered only these 3 nutrients when thinking about "overall nutrient profile." Specifically, with the former outcomes (overall nutrient profile), the High In schemes performed moderately lower than GDA and Nutrition Info schemes, but with the latter outcomes (correctly characterizing nutrient content), High In schemes performed as well if not better than the Nutrition Info schemes while GDA performed substantially worse than both. Unless there is clear evidence that consumers interpreted "overall nutrient profile" in the manner intended, the "correctly characterizing the level of nutrient" outcomes should be primary, and the "overall nutrient profile" outcomes should be secondary with limitations stated.
Reviewer #3	Given the lack of education participations were given about how to use the labeling schemes, especially the less familiar High In scheme, the other attitude and perceptions outcomes have limited usefulness. These should be classified as secondary or exploratory outcomes. Relatedly, I am curious what the rationale was for including some of the specific perception and attitude items. There is limited evidence that measures of trustworthiness, for example, predict real-world responses to labels. Also, had the items been cognitively tested for use with these schemes? For example, for the item, "Can Easily Find Nutrition Information on this Label," participants may just be looking for whatever format looks most similar to the NFL and contains the most numeric information regardless of how well understood that numeric information was.

CHARGE QUESTION 5. Are the study participants that are included appropriate given the study's purpose? If not, please explain why.

REVIEWER	COMMENT
Reviewer #1	The study participants overall seem to be appropriate, though it would have been preferable to ensure that the sample was geographically representative in addition to representative on other characteristics. The authors should consider adding geographic region to Table 1. A limitation is that participants had to be English-speaking, as it would be interesting to see whether the results differed for those who did not speak English. It is possible that non-English speakers will respond differently to different labels, and that these effects might be distinct from English speakers (e.g., color may be a more important indicator of healthfulness than text for non-English speakers compared to English speakers).
Reviewer #2	The FDA shared their cooperation rate, but I suggest a note about how it compared to the field or their expectation. Otherwise, the description of the sample procedures was sufficient and well-done.
Reviewer #3	Yes, the study participants included were very appropriate for the study's purpose, and the use of sampling quotas to match key demographics of the U.S. population was wise.

3. Quality of the analysis/data

CHARGE QUESTION 6. Is the analytic approach appropriate given the design and purpose of the study? Are there additional considerations for the approach FDA used?

REVIEWER	COMMENT
Reviewer #1	The analytic approach has several issues. First, as mentioned in my response to Q2, there is missing rationale for many of the analytic decisions that the authors made, including adjustment for several covariates in the model and use of the Bonferroni adjustment (i.e., over other methods for accounting for multiple testing). The model needs to be written out explicitly and the terms that were included in the model need to be defined. It is difficult to fully determine whether the analytic approach was appropriate without these details.
Reviewer #1	The model seems to have included several interaction terms, and though it is not stated why, I assume this was to test for potential effect modification by product type, nutrient profile, rurality, and nutrition knowledge. If this is true, then I think these interaction terms should have been examined one at a time to determine if there was effect modification, rather than running a model with all of them simultaneously. Again, more rationale is necessary.

REVIEWER	COMMENT
Reviewer #1	The analyses for the Single Product Evaluation Task need to account for the fact that %DV is different across different labeling schemes at the same level of healthfulness. For example, the middle category has 15% DV sodium for the GDA and Nutrition Info schemes, whereas it is 21% for the High In scheme. It is possible that participants were more easily able to correctly classify these products as having amounts of sodium if they saw the High In label versus the other schemes. This may be why Exhibit 5 shows the High In scheme perform so much better than the other schemes for the "middle healthy" products. One way to explore this would be to rerun the analysis and exclude the people who saw a label with a %DV. Then you would be comparing just the schemes themselves without incorporating bias due to different level of sodium across the different schemes.
Reviewer #1	As mentioned in Q3, the analyses comparing the Black and White with no %DV scheme in different locations is insufficient to answer this question given the potential for effect modification across different labeling schemes. The Bonferroni adjustment should not include the 8 schemes tested and this "change in location" test as one family of tests when adjusting the p-values. It is also not clear to me whether a Bonferroni correction is appropriate here, especially given that many of the tests are very similar (i.e., it is not convincing that these are all independent tests considering that many of the treatments are very similar to one another). Allowing that a modified approach could be less strict, it might have been an appropriate approach, but details on the Bonferroni correction need to be specified. It is possible that a different approach (controlling for the false discovery rate) would be more appropriate.
Reviewer #1	As noted above, I think it would be better to report 95% confidence intervals rather than just provide significance tests when reporting results.
Reviewer #1	Lastly, as noted above, I strongly recommend that the authors report not only the findings for the control group for the Single Product Evaluation Task, but also differences between the schemes and the control group, which would provide understanding of how much better (if at all) the schemes perform versus the status quo.
Reviewer #2	My only question about the analysis is what happened to the sample group size distribution of the Nutrition Info scheme after all six schemes were collapsed into one for analysis? I suggest making a note of this in the analysis section.
Reviewer #3	The statistical models used were generally appropriate, and the use of a modified Bonferroni procedure to account for multiple testing was a nice feature.
Reviewer #3	Because this study design used unconditional randomization, the models used to determine the marginal effects of the labeling schemes should not adjust for any covariates (e.g., rural residency, age, gender). See CONSORT guidelines. Instead, the analysis could weight the data to represent the U.S. distribution of all of these variables.
Reviewer #3	The extent to which repeated measures (across the 3 label schemes viewed) in the Comparison Task were accounted for in the statistical models was not adequately described. Did the model for the Comparison task compare schemes both within and between subjects?

REVIEWER	COMMENT
Reviewer #3	Related to my comment about how the magnifying glass and color variations were not tested in all schemes, these variations should be dropped when analyses look at scheme category as the primary treatment variable for both tasks. Alternatively, instead of grouping by scheme category, the analysis could compare all individual schemes against one another and show them in the same table.
Reviewer #3	Lastly, figures showing percentages would benefit from the Y axis including 0% to 100%.

4. Study conclusions

CHARGE QUESTION 7. Are the conclusions drawn from the study supported by the data presented? If not, are there other considerations?

REVIEWER	COMMENT
Reviewer #1	Not all of the document's stated conclusions are supported by the data presented, and there is insufficient discussion of the clinical importance of the findings. The authors have also not considered the study's limitations when interpreting the results.
Reviewer #1	First, as mentioned above in my response to Q1, the lack of a prespecified primary outcome makes it difficult to synthesize all of the results without knowing which analyses were considered the most important to answer the question at hand. The authors need to specify primary and secondary outcomes and use this designation to guide their conclusion.
Reviewer #1	Second, there are many limitations in this study, as outlined above. These should be incorporated into the conclusion, and there should be discussion of how the results might be different under more ideal circumstances. For example, the fact that there was a very limited array of products tested suggests that these results might not generalize to most products sold in stores. Similarly, the fact that they did not use purchase outcomes (or even intention to purchase) limits inferences that can be drawn from this study about a future FOP labeling policy. These are just two examples, but there are many other limitations to this study that need to be considered when determining how to interpret the results. Not discussing them explicitly reduces transparency and is misleading.
Reviewer #1	Third, there is no discussion of statistical vs. clinical significance (i.e., how and whether any statistical differences between labeling schemes would translate to meaningful differences in the real world). One major issue is that statistical significance is not even presented for many of the findings (e.g., subgroup analyses), yet language is used to imply important differences between groups ("slightly less likely" or "appreciably less likely;" see my comments to Q2 above). Statistical differences between groups need to be adequately presented in the Results, and then interpreted in the Abstract and/or Conclusion section to help the reader determine whether these would yield clinically meaningful differences in the real world. If the authors do not feel comfortable making this kind of statement given the limitations of their study, they should say so. Any statements about the clinical significance of the findings should be supported with references to the scientific literature. Relatedly, the authors imply in the Background and Purpose section that FOP labels could "encourage industry innovation," but this study only considered consumer reactions. It would be helpful to mention this (and possibly that any real-world effects of FOP labels could be partially driven industry reformulation).

REVIEWER	COMMENT
Reviewer #1	Fourth, and related to the above comment on clinical significance, the authors tested a no-label control group, but did not present any results on this group, or differences between the labeling schemes and the control group. This restricts the reader's ability to understand the importance of these labeling schemes for improving customers' nutrition knowledge of packaged products compared to the status quo. The authors should strongly consider adding in these results and synthesizing the overall findings along with the results from the labeling schemes.
Reviewer #1	Fifth, there are several inaccuracies and over-generalizations in the Abstract and Conclusion sections. In the Abstract, the biggest problem is with the last sentence of the second-to-last paragraph, which states *"…and the versions that were black and white with %DV performed best in several instances*." This is not accurate. In Table 6, the Black and White with %DV performed statistically significantly worse than several of the other schemes for correctly answering questions about level of added sugars. The point estimates were also on the lower side of all the schemes for saturated fat and sodium. The Black and White with %DV performed better than some of the schemes for some of the attitude and perception questions, but that is true of several other schemes. This sentence is therefore misleading as written.
Reviewer #1	Another major issue is the second line of the last paragraph of the Abstract, which states that "*High In schemes performed the worse among the schemes tested*." This is too much of a generalization. The High In schemes performed better than the GDA scheme for every nutrient in the single product evaluation task, and better than the Nutrition Info schemes for sodium. The next line states "*Consumers reacted positively to the GDA concept but were less likely to use GDA to correctly identify product healthfulness*." This needs to be stated in relation to the Nutrition Info scheme. It would be more accurate to say something like "*Consumers' attitudes and perceptions were equally positive between the GDA and Nutrition Info schemes, but they were more likely to correctly identify product healthfulness using the Nutrition Info scheme.*"
Reviewer #1	In the Conclusion, the first line of the first paragraph is far too much of a generalization ("*Overall, both the GDA and High In schemes performed poorly on tasks associated with an understanding of the nutrient content displayed on the schemes*"), especially because results were not compared to that of a control group to provide a comparison to the status quo. The High In scheme performed objectively moderate or well on many of the tests, though it is hard to know given all of the issues surrounding transparency and clarity of the Methods. It would be better to frame the main findings in relative terms (e.g., "*Overall, the Nutrition Info schemes appeared to perform better than the other schemes on most of the tasks associated with an understanding of the nutrient content displayed on the schemes*"), and, again to compare the results to a control group.
Reviewer #1	The last line of the first paragraph of the Conclusions section is not entirely accurate ("*Results for the Nutrition Info schemes show that they did not produce incorrect answers or low scores at rates similar to those of the GDA and High In schemes*"). For example, in Table 3, 60% of participants in the Nutrition Info group answered correctly about sodium level, compared to 70% in the High In group. Additionally, GDA did as well as Nutrition Info on several measures (see results in Tables 2, 4, and 4a). This sentence should be modified to reflect these nuances.

REVIEWER	COMMENT
Reviewer #1	Lastly, the last line of the Conclusions section is not supported by the evidence from this study ("*Some consumer education about the middle nutrient profile might be helpful if a front-of-package nutrition labeling scheme is adopted*"). There is no evidence cited to support this claim and it was not tested in the study. I suggest removing it.
Reviewer #2	Perhaps it is not customary for the FDA to go too far in their interpretation of the results. That being said, a few more comments about the differences observed in Exhibits 5 and 6 and how those results may speak to real-world consumer behavior would have added to the conclusion. They also do not share specific strengths or weaknesses of the study, which would be important in terms of deciding next steps.
Reviewer #3	The conclusion that the Nutrition Info labels performed well is supported by the data, but some other parts of the conclusion need addressing. In particular, the conclusion would benefit from describing and interpreting the absolute values of and the magnitudes of differences in label performance on measures related to identifying the nutrient contents of products, rather than just reporting on statistical significance.
Reviewer #3	The following sentences are not supported by the data: "Overall, both the GDA and High In schemes performed poorly on tasks associated with an understanding of the nutrient content displayed on the schemes. Results for the Nutrition Info schemes show that they did not produce incorrect answers or low scores at rates similar to those of the GDA and High In schemes." The best measures of understanding nutrient content were the "correctly characterizing the level of nutrient" questions. For these outcomes, the GDA performed poorly in an absolute sense (only 25.5% correctly answered sodium question) as well as relative to the Nutrition Info and High In schemes. In contrast, both the Nutrition Info and High In schemes performed well on these measures (60-86% correct for Nutrition Info and 70-87% correct for High In), with High In outperforming the Nutrition Info for level of sodium. Additionally, although there are potential problems with the outcomes related to "overall nutrition profile," none of the labels performed "poorly" for these outcomes in an absolute sense. Across scheme category, the majority of participants could correctly identify the healthiest (70-95%) and least healthy profile (88-93%) from the labels, with the least healthy profile being the most relevant of these two outcomes given the nutrients on the label. The High In did not perform as well as the other labels on the "overall nutrient profile" outcomes but did not perform poorly. Considering the results for these two groups of outcomes (including not only significant differences but also the absolute % of participants answering questions correctly), I would conclude that when participants viewed the Nutrition Info and High In labels, there was a desirable understanding of nutrient contents regardless of outcome (the majority answered correctly); whereas the GDA did not perform acceptably, especially on "correctly characterizing the level of nutrient" questions. But I would also recommend acknowledging that, 1) the use of different nutrient profiles for the High In schemes and 2) the lack of education on how to use the labels, especially for the less familiar High In scheme, limits this study's ability to directly assess how the High In schemes would perform relative to other schemes in a real-world setting. In particular, the lack of education provided suggests that the less-familiar High In schemes are likely to perform better than what was observed in this study.

REVIEWER	COMMENT
Reviewer #3	Another sentence only partially supported by the data is the second sentence that follows: "The level of saturated fat on the healthiest and the middle nutrient profiles was low and for the least healthy, the level was high. This proved much more difficult for participants to discern when viewing the GDA and the High In schemes than when viewing the Nutrition Info schemes." This sentence is true for the GDA (only about 30% answered correctly), but less so for the High In (nearly 70% answered correctly) compared to the Nutrition Info (about 90% answered correctly).
Reviewer #3	The conclusions should also summarize how the BW with %DV Nutrition Info scheme did not perform as well on the "correctly characterizing the level of nutrient" outcomes. Based on Table 6, the BW with %DV performed significantly lower than the following 3 designs for correctly answering level of added sugar: BW No %DV in lower Right, Magnifying glass, and Color No %DV. The BW No %DV also performed meaningfully lower than these designs (but not significantly so) for correctly answering level of saturated fat (81% vs. 86-88%) and sodium (56% vs. 62-63%). The lack of statistical significance of these differences likely resulted from inadequate power given the high number of schemes, comparisons, and modified Bonferroni corrections. Also, the conclusions' summary of the attitude and perception outcomes for the Nutrition Info schemes gives a different impression that the BW with %DV performed relatively well. Given the questionable predictive value of the attitude and perception outcomes in understanding consumer knowledge and behavior in response to labels, I would limit discussion of these outcomes in the conclusion or, at a minimum, include a caveat about the uncertain value of these outcomes.

CHARGE QUESTION 8. Please share any additional comments.

REVIEWER	COMMENT
Reviewer #1	No comments provided.
Reviewer #2	Overall, FDA has done an incredible job executing a study of this size to better understand how consumers engage with FOP and how effectiveness they may be in making healthy food choices.
Reviewer #3	This report is researching a topic of tremendous public health importance, and the FDA's efforts on developing an FOP scheme are commendable.
Reviewer #3	In the event that future studies are conducted by the FDA on FOP schemes, additional icons that could aid in noticeability and understanding, such as an exclamation mark in a shape, should be tested. Primary outcomes should include the objective healthfulness of foods selected for purchase or consumption, including their contents of added sugar, sodium, and saturated fat, and whether the food meets the FDA's definition of healthy. Also, instead of asking participants their perception of whether they think they would quickly notice the label, any future studies should objectively assess noticeability of the label's contents after exposing participants to a label in the context of a package of food. These, in addition to addressing limitations noted above, would improve future experimental work on FOP schemes.

REVIEWER	COMMENT
Reviewer #3	Although more research can always be done, the study herein already provides evidence that both the Nutrition Info and High In labeling schemes performed objectively well on questions assessing consumer understanding of the nutrient contents of the products viewed, with both scheme categories resulting in the majority of participants correctly classifying levels of saturated fat, sodium, and added sugars. This study also suggests that adding color and a magnifying glass, and not including %DV, may strengthen the performance of a Nutrition Info scheme in helping consumers identify products high in added sugars and potentially saturated fat and sodium as well.

B. SPECIFIC OBSERVATIONS

REVIEWER	Page	Paragraph/ Line	Comment
Reviewer #1	2	Third paragraph of the Abstract	The text states: "Moreover, ratings on the attitude and perception questions were significantly lower for the High In schemes than they were for the GDA and Nutrition Info schemes." This is true for all except for the Simple to Complex question. It would be better to say: "Moreover, **most** of the ratings on the attitude and perception questions…"
Reviewer #1	3	Last line of last paragraph of the Abstract	The text states: "Trends were the same across demographic groups." First, trends were not reported, so I think it would be better to describe them as "results". But perhaps more important, there is a lack of transparency surrounding this finding which makes it hard to know if the statement is true (see my comments for Q2). This should be confirmed and reported explicitly in the document before it is included in the Abstract.
Reviewer #1	5	Last line of the first paragraph of FOP Comparison Task	The text states: "The presentation order for FOP scheme and nutrient profile were randomized such that participants viewed **any combination of these two variables**, resulting in 336 experimental conditions (see Appendix A for FOP schemes and nutrient profiles)." This was confusing to me. The "any combination of these two variables" part made it sound like a participant could be shown the same nutrient profile across 3 different schemes, which would make it impossible to rank in terms of healthfulness. I think they actually were shown 3 schemes, each with 3 nutrient profiles. And that the order of the schemes and the order of the profiles (which were shown first/last and left/right) were random. This should be confirmed and reworded to make it clearer.

REVIEWER	Page	Paragraph/ Line	Comment
Reviewer #1	5	Last paragraph, second line	The text states: "The objective of this task was to determine which FOP schemes are perceived more accurately, perceived more favorably, and facilitated greater understanding of nutrient content." I am confused by what "perceived more accurately" means. This makes it sound like they were testing which schemes participants perceived to be accurate (i.e., trustworthy), but I think it's better to say "which they more accurately classified" or something similar.
Reviewer #1	5	Last paragraph, last line	The text states: "This task included three independent variables: FOP scheme, nutrient profile, and product type." It is confusing to refer to these as "independent variables" when there are other actual independent variables in the model (e.g., race, gender, age). I think it's better to refer to these as three labeling factors, or something similar.
Reviewer #1	9	Data Analysis	The text states that the model included a variable for "whether the participant was paying attention to their intake of sodium, saturated fat, and added sugars." It is not clear whether this was a single variable (i.e., paying attention to any of them) or if it was 3 separate variables for each. I recommend the latter approach given issues with some nutrients being more correlated with product healthfulness than others (see my response to Q3 above).
Reviewer #1	11	First line of paragraph for C2b results	The text states: There were no significant differences between the six Nutrition Info scheme conditions… on correctly characterizing the level of saturated fat, sodium, and added sugars (see Table 6)." This is not true. There was a significant difference between scheme 3 with schemes 1a, 2, and 4.
Reviewer #1	11	Second line of the first Conclusions paragraph	The text states: "Moreover, the High In schemes performed worse than both the GDA and Nutrition Info schemes on the attitude and perception measures." Again, it should say this is true for **most** of the attitude and perception measures (because this statement is not true for the Simple to Complex measure).

REVIEWER	Page	Paragraph/ Line	Comment
Reviewer #1	11	First line of the second Conclusions paragraph	The text states: "The interactions between the scheme categories and both nutrition knowledge/motivation and rural residency are minor…" However, the quantitative results are not presented in the main text of the document, making it hard to assess the veracity of this statement.
Reviewer #1	12	First line	The text states: "…with those of lower nutrition knowledge/motivation correctly selecting the healthiest nutrient profile at a lower incidence than those with higher nutrition knowledge/motivation…" The word "incidence" is being used in a confusing way. Perhaps the authors could change to "less frequently"?
Reviewer #1	14	Exhibit 4	I think the figure would be easier to understand if the health categories were on the x-axis and you can track how each scheme did across different levels of healthfulness (i.e., if it were flipped).
Reviewer #2	4	71	The authors note that "Eight different FOP schemes were used in both tasks." However, earlier they note that participants were only viewed three randomly selected tasks. It may be clearer to say that *selections were made from the same eight FOP schemes in both tasks*, or something
Reviewer #2	5	91, 110	Suggest adding "Task 1" and "Task 2" to titles (here and throughout) to make it clearer given that titles weren't used initially
Reviewer #2	7	Footnote	Indicated that power analysis suggested 9,000, but p. 43 notes 10,000
Reviewer #2	16	Table 4a	Typo on the "Can sometimes eat [this] product…" column (An embedded number appears; when I try a copy and paste, it's a different number, so if it doesn't show up, may be a tech issue on my end)
Reviewer #3	11	C2b	The following statement is not consistent with results in Table 6: "There were no significant differences between the six Nutrition Info scheme conditions (five total schemes plus one scheme tested in the lower right corner of the label) on correctly characterizing the level of saturated fat, sodium, and added sugars (see Table 6)." In Table 6, the BW with %DV scored significantly lower in correctly answering level of added sugar than 1a, 2, and 4 (lower right, color, and magnifying class).

REVIEWER	Page	Paragraph/ Line	Comment
Reviewer #3	12	Table 2	The result for "correctly identified healthiest nutrition profile" in Table 2 for Nutrition Info category (95%) does not appear consistent with the data on the specific Nutrition Info schemes in Table 5 (range: 91-94%). Should the 95% in Table 2 be more like 92% or 93% based on Table 5?
Reviewer #3	6	Paragraph 1	It would help to state that the same schemes shown in Appendix A were used for this part of the experiment.
Reviewer #3		Exhibit 1	Higher resolution photos and at least one example of the High In should be shown.
Reviewer #3	18		The definition of "cooperation rate" should be provided, including clarifying whether the denominator includes only those who passed the screener or all those who attempted the screener.
Reviewer #3		Table 7A	Does the FDA consider a higher or lower score on the following item indicative of better label performance, "Can Sometimes Eat This Product Even if Limiting Sat Fat, Sodium, or Added Sugar"? It should depend on how healthy the item is. Technically, everything can be consumed in moderation, even if limiting a nutrient, but I would guess that higher scores (indicating higher agreement) in response to a product high in any of the 3 nutrients is predictive of overconsumption of such items. So I would interpret a lower score to indicate better performance in response to products high in one of these nutrients.
Reviewer #3	9	Analysis	Which "modified Bonferroni" approach was used to adjust p-values? A citation would help.
Reviewer #3	N/A		Appendix with question screen shots: The reviewers requested screenshots of the questions, which were not included in the original report. In the screenshot of the question that states "The Food and Drug Administration (FDA) is exploring the idea of developing nutrition labels..." the accompanying image shows the GDA label as an example. Did all participants see the GDA before being randomized to their first label scheme, or did they see the first label they were randomized to? If the former, this is a study limitation that should be acknowledged.