

FDA's Response to External Peer Review of Quantitative Consumer Research on Cigarette Health Warnings Required by the Family Smoking Prevention and Tobacco Control Act (Experimental Study on Warning Statements for Cigarette Graphic Health Warnings and Experimental Study of Cigarette Warnings), February 4, 2020

I. INTRODUCTION

To fulfill its statutory obligation under Section 201 of the Tobacco Control Act (TCA) (Pub. L. 111-31), Food and Drug Administration (FDA) has developed, refined, and tested new cigarette health warnings that depict the negative health consequences of cigarette smoking. Pursuant to Section 202(b) of the TCA, the Secretary may, through a rulemaking, adjust the text of the label requirements if doing so would “promote greater public understanding of the risks associated with the use of tobacco products.” As part of the cigarette health warning development process, FDA developed new textual warning statements that were tested in a quantitative consumer research study (Study 1). Based on the results of Study 1, FDA selected warning statements that were then paired with concordant photorealistic images that depicted the negative health consequences of cigarette smoking to form cigarette health warnings; those warnings were tested in a second study (Study 2).

Versar, Inc., an independent contractor, coordinated an external letter peer review of these two quantitative consumer research studies on cigarette health warnings. The peer review was conducted for FDA's Center for Tobacco Products. For this peer review, six experts were selected by Versar, Inc. to evaluate and provide written comments on the format and content of the reports from these two studies, including both the clarity of the documents, the presentation of the studies, and the scientific content of the studies.

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 reports in response to peer reviewer comments or an explanation of our decision to not make suggested changes. For comments¹ 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 in elsewhere. The individual peer reviewer comments are provided in tabular format in Appendix I.

¹ At times, peer reviewers provided comments related to a specific charge question in response to a different charge question, meaning that similar content was included in response to multiple charge questions.

Below are the names and affiliation of the peer reviewers:

Joseph N. Cappella, Ph.D.

University of Pennsylvania, Annenberg School for Communication

Jennifer C. Morgan, Ph.D.

University of Pennsylvania, Tobacco Center of Regulatory Science

John P. Pierce, Ph.D.

University of California, San Diego

Kurt M. Ribisl, Ph.D.

University of North Carolina, Gillings School of Global Public Health

William G. Shadel, Ph.D.

RAND Corporation

James F. Thrasher, Ph.D.

University of South Carolina, Arnold School of Public Health

II. CHARGE TO REVIEWERS

FDA has completed two quantitative consumer research studies, one (Study 1) testing textual warning statements concerning the negative health consequences of cigarette smoking, and the second (Study 2) testing combinations of textual warning statements paired with concordant photorealistic images depicting the negative health consequences of cigarette smoking (i.e., cigarette health warnings). The purposes of the studies were (1) to assess whether new FDA-developed textual warning statements increased understanding of the negative health consequences of cigarette smoking relative to the warning statements provided in the Tobacco Control Act (Study 1); and (2) to assess whether FDA-developed cigarette health warnings increased understanding of the negative health consequences of cigarette smoking relative to the Surgeon General's warnings currently used on cigarette packs and advertisements (Study 2). Both studies were provided to the peer reviewers as reports containing background, methods, results and conclusions from the study (Experimental Study on Warning Statements for Cigarette Graphic Health Warnings: Study 1 Report ("Study 1 Report"); and Experimental Studies of Cigarette Warning Labels: Study 2 Report ("Study 2 Report")) FDA requested that the peer reviewers provide input on clarity of the documents describing those studies and the soundness of the design and analysis of the studies including: (a) the clarity of the description of the study designs, analyses, and results as presented in the documents; (b) the scientific soundness of the methodology used; (c) the quality of the analysis/data; and (d) whether the conclusions reached are supported. As noted in Section I of this peer review response, the study reports were updated where appropriate and subsequently finalized. The final versions of the study reports can be found at <https://www.fda.gov/science-research/peer-review-scientific-information-and-assessments/completed-peer-reviews>.

Charge Questions (answered separately for each study report)

Clarity of the documentation of the studies

1. Is the document logical and clear?
2. Does the Executive Summary accurately reflect the content of the overall document?
3. Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?

Scientific soundness of the methodology used

4. Is the methodology used appropriate to address the study's purpose?
5. Are the stimuli used appropriate given the study's purpose?
6. Are the outcomes measured appropriate given the study's purpose?
7. Are the study participants included appropriate given the study's purpose?

Quality of the analysis/data

8. Is the analytic approach appropriate given the design and purpose of the study?
9. Are results presented consistent with the analytic approach?

10. Are there any concerns with the results presented?

Study conclusions

11. Are potential limitations of the study appropriately identified?

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

III. SUMMARY OF PEER REVIEWER COMMENTS AND FDA RESPONSE

Study 1: Experimental Study on Warning Statements for Cigarette Graphic Health Warnings

Summary of General Impressions:

Overall, the reviewers agreed that Study 1 was well-conceived and implemented using high-quality methods and analyses. They also agreed that the study was appropriate to address the stated aim of the study and that the conclusions closely follow from the data. The reviewers also generally agreed that the reports were comprehensive and provided clear details about the study. In response to this request for general impression of the study, reviewers highlighted several suggestions for improvements in the clarity and flow of some sections of the report, and some requested additional information or clarification to be added to the report about the methods, stimuli (i.e., warning statements), and analyses that were also raised in response to other charge questions. Reviewers also provided some overarching comments about how to better contextualize this study in terms of the regulatory purpose of the study and a theoretical framework for the outcomes used.

FDA Response:

We appreciate the reviewers' support of the study and their suggestions to further strengthen the study report. We have responded to the specific comments included in the general impressions in the responses to the relevant questions below. We have streamlined the Study 1 report as noted in response to Question 1 below and added additional details about the theories that informed the development and selection of study questions as noted in response to Questions 4, 6, and 10 below.

1. Is the document logical and clear?

Summary of Comments:

Reviewers stated that the report was logical and clear. Reviewers suggested that the clarity of the document could be further improved by removing redundant information that was a result of the document including multiple previously stand-alone study documents that had been combined. They also suggested that a figure on the flow of the study would aid clarity to help the reader understand the series of steps used in the study.

FDA Response:

In response to these comments, the report has been streamlined, better integrating the separate sections of the report by eliminating additional cover pages and combining and eliminating redundant text. Additionally, the report is now continuously paginated. In response to the request for a figure to better explain the study procedure, a new figure has been added as Figure 2 on page 6 of the Study 1 Report.

2. Does the Executive Summary accurately reflect the content of the overall document?

Summary of Comments:

Overall, reviewers found the Executive Summary of the Study 1 Report helpful and clear and generally

reflecting the contents of the overall document. To improve the Executive Summary, reviewers suggested some clarifications including about the stimuli (i.e., warning statements), the scope of outcomes measured, and the overall methodology. One comment suggested that we link the outcomes to cessation.

FDA Response:

In response to these comments, the Executive Summary of the Study 1 Report has been edited to provide additional details about the stimuli, methods, and outcomes included. We have also added details about all findings from Study 1 to the summary. However, we decline to make the suggested change related to cessation, as behavioral outcomes are neither relevant for this study nor for the warnings being studied as the purpose of the study was to assess which, if any, of the revised warning statements promote greater public understanding of the risks associated with cigarette smoking as compared to the TCA statements across a range of outcomes..

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

Summary of Comments:

Reviewers noted that the Study 1 Report presents detailed information about the study design, stimuli, sample, methods, analysis, and results and provided suggestions to further improve their descriptions in the report. For example, reviewers suggested adding a figure to depict the study flow, providing additional information about the development of the stimuli (i.e., warning statements) used in the study, including more information on the source of the questions used in the study, and adding more details on the planned versus actual assignment of participants to conditions.

FDA Response:

In response to these comments, a new figure has been added as Figure 2 on page 6 of the Study 1 Report that visually depicts the study flow. Throughout the document, additional details about the study procedure have been added and simplified to reduce redundancy between sections as most peer reviewers noted the repetition of information was unnecessary. Information about assignment of participants was already included in the report (see Table 14). The source of the questions asked in the study is now noted both in Section 2.3 (Instrument Development), which discusses survey instrument design and content, and Section 4.1 (Measures and Coding), which discusses the selection of study variables used to inform warning message responses by consumers. Additional details have been added to Section 1 describing the development of the stimuli (i.e., revised warning statements); however, a detailed description of the multiple studies and other activities that were included in the development process of the textual warning statements used in this study is beyond the scope of this study report.

4. Is the methodology used appropriate to address the study's purpose?

Summary of Comments:

Most reviewers agreed that the study methods were generally appropriate. Reviewers provided only minor comments on the methodology of Study 1, and noted the steps already taken by FDA to address limitations common to online studies. Reviewers suggested adding additional details to the report: on the decision to present only one revised warning statement in Phase 1 of the study; adding additional

details about the power calculation; and adding additional support for the use of the mock cigarette packs by citing similar studies found in the peer reviewed literature that demonstrate the consistency of tested warnings prior to implementation to effects once implemented . An additional comment noted that use of different numbers of health beliefs items per health condition may be a drawback to the design. Another comment noted that comparison of the revised warning statements to the TCA statements seems to be a weak comparison.

FDA Response:

In response to these comments, additional details have been added about the use of only one revised warning statement in Section 2.1 (Experimental Design) of the Study 1 Report, and additional details that effect sizes used in the power calculation were drawn from previous similar studies, as is standard practice, have been added to Section 4.2 (Power Analyses). We have added additional details and a citation justifying the use of an online study to Section 6.2 (Limitations). We acknowledge the challenge of measuring health beliefs using different numbers of questions and interpreting results; however, we note that we have addressed that issue analytically, as is already explained in the Section 4 (Analysis) of the report. With respect to the robustness of the comparison of revised warning statements to TCA statements, we note that the comparison was made to inform a very specific regulatory purpose per Section 202(b) of the TCA (as noted in the Background section of the Study 1 Report), thus using a different comparison, such as a no-exposure control, would not provide the necessary data to inform FDA's further development of the warnings. For this reason, the Study 1 report was not updated in response to this comment as that explanation is already included.

5. Are the stimuli used appropriate given the study's purpose?

Summary of Comments:

Overall, reviewers noted that the stimuli (i.e., warning statements) were appropriate given the goals of the Study 1, and reviewers also noted that these types of stimuli are standard to other similar studies. Reviewers suggested adding more information to the Study 1 Report about the development of the stimuli and suggested adding an example of the stimuli as it was presented to participants. One comment noted that using statements to teach something new may be viewed with skepticism by the public; however, this comment was not specific to the warning statements being tested.

FDA Response:

An example of the stimuli (i.e., warning statements) has been added as Figure 1 of the Study 1 Report. Additional details describing the development of the stimuli have been added to Section 1 of the Study 1 Report. However, a detailed description of the iterative development process of the textual warning statements used in this study, which included multiple studies, is beyond the scope of this study report. We acknowledge that new information may be viewed with skepticism and note that such a finding is common in pre-implementation studies that test warnings for health effects for which there are low levels of consumer awareness (e.g., Hammond Fong 2006; Kennedy 2012; Hammond 2011). Importantly, such findings in pre-implementation studies do not suggest that the warnings will not have the intended effects. This study was designed to meet a specific regulatory purpose (as noted in the Background section of the Study 1 Report), and note that the warning statements tested in this study were paired with an image and tested in an additional study (Study 2), which is now completed; this is noted in the Executive Summary of the Study 1 Report to better contextualize this study.

6. Are the outcomes measured appropriate given the study's purpose?

Summary of Comments:

Reviewers noted that the measures used are appropriate; some provided suggestions for other measures that could have also been included in Study 1; and one asked if two outcomes were collinear (i.e., essentially measuring the same underlying construct) and, if so, suggested that one outcome should be dropped. Some reviewers noted that additional details on a theoretical framework or model to better understand how the measures relate to the study's purpose would be helpful, and requested that one be included. One reviewer suggested that such a model was needed to link this study to behavioral outcomes. Reviewers also asked that additional details about the selection and validity of the measures used be added to the report and requested clarification of whether all participants were asked the same set of health belief questions.

FDA Response:

Details about the design of the study, including an explanation of the theory that informed the selection of health belief items evaluated has been added to Section 2.3 of the Study 1 Report. The sources of the health belief items used in the study are now included in Section 2.3 (Instrument Development) and Section 4.1 (Measures and Coding). As one reviewer correctly noted in their comments, behavioral outcomes are outside the scope of the study and therefore are not included in the study. Therefore, we decline to include behavioral outcomes in response to another reviewer's suggestion to the Study 1 Report. Although we appreciate the suggestion of additional items that could have been used, many of the suggested items are not consistent with the narrowly focused purpose of this study or are already addressed by other questions already included in the study. Furthermore, Study 1 is complete, and we are unable to include new measures in this study. Importantly, these suggestions for additional measures to be included do not diminish the usefulness of the measures that were used in the study. With respect to the two potentially overlapping outcomes, since both are included in the Statistical Analysis Plan, we do not support dropping either from the study report for two reasons: 1) for the sake of transparency in the analysis; and, 2) we do not agree that these are the same construct, although we might expect these outcomes to be correlated.

7. Are the study participants included appropriate given the study's purpose?

Summary of Comments:

Overall, the reviewers noted that the sample and sub-groups included in the study were appropriate. Some reviewers asked for more clarification about whether the data were weighted, whereas others recognized that the sample, albeit large and diverse, was a convenience sample. Some reviewers asked for more details on: the smoking status of participants to further analyze the results by sub-group; the variable used to screen participants (i.e., susceptibility); an explanation for excluding adult non-smokers or adult "ever smokers who are susceptible."

FDA Response:

As stated in the Study 1 Report, the sample is a large and heterogeneous convenience sample. Weighting was not performed, and, as explained in the report, the inclusion of different age and smoking-status subgroups was not designed to make the study nationally representative or with the intent to perform sub-group analysis. Rather, as now explained in more detail in Section 2.2 (Sampling Frame and Sampling Methodology) the intent was to obtain a large and heterogeneous sample to test

outcomes across a range of individuals to strengthen the study. We have updated the Study 1 Report with additional text to clarify that the data were not weighted. The study was not designed or powered to test for differences between age and smoking-status groups, thus the approach to sampling was appropriate. The measure to define susceptibility was included in Section 2.1 of the Study 1 Report, and we note that the measure of susceptibility is only validated among never smokers, thus there is no potential categorization of participants as “ever smokers who are susceptible.”

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

Summary of Comments:

Overall, reviewers stated that the analytic approach was appropriate and logical. Some reviewers suggested additional ways to analyze the data (e.g., dichotomizing certain responses) or additional ways to examine some of the comparisons (e.g., compare each statement to the grand mean of all other statements rather than to one randomly selected statement). Reviewers also suggested some clarifications that could be made with respect to the hypotheses, and suggested analyses to evaluate the results stratified by age and tobacco use status sub-groups. One reviewer also asked for clarification about the use of Cronbach's alpha for scales with two items.

FDA Response:

Although we appreciate the feedback suggesting potential additional analyses, we note the suggested analyses are not contained in the Statistical Analysis Plan that was developed prior to data collection. The Statistical Analysis Plan was implemented as designed. Many of the analyses suggested, such as analysis by sub-group, are not consistent with the design or aims of Study 1, nor was the study powered to examine those differences as is explained in the Study 1 Report. Importantly, the reviewers noted that the analyses performed were appropriate – and reviewer suggestions were for additional analyses, not questioning the appropriateness of the analyses planned and conducted. We have updated the phrasing of the hypotheses in Section 4.3 (Analyses) per one reviewer's suggestion to better align with the two-tailed nature of the analyses that were performed. With respect to Cronbach's alpha, although it might erroneously result in a lower reliability in some cases for two-item scales (Eisinga et al., 2013), it still has useful and has meaning for how we use it in our study. If anything, such a lower reliability would suggest that this is a more conservative approach.

9. Are results presented consistent with the analytic approach?

Summary of Comments:

The reviewers agreed that the results were consistent with the analytic approach and noted that the results section (Section 5) is strong and provided a great deal of information. Reviewers suggested the addition of a summary figure to highlight the results. One reviewer also asked for clarification about the methods used in Phase 2 of the study to better understand the results for outcomes measured in that phase, and another asked for clarification about the hypotheses.

FDA Response:

Because we already include a table (Table 29) in the Study 1 Report that summarizes the results across all outcomes and all participants, we decline to add a similar figure as it would be redundant and could cause confusion. As discussed in earlier responses to charge questions 1 and 3, we have also added to

the report a new figure (Figure 2) that illustrates the study flow from condition assignment through Phases 1 and 2 to give better context for the Phase 2 outcomes, and show how they are different from those assessed in Phase 1. We have also updated the phrasing of the hypotheses in Section 4.3 (Analyses) per one reviewer's suggestion to better align the hypotheses with the nature of the of analyses (i.e., two-tailed analyses, to examine results in either direction) performed.

10. Are there any concerns with the results presented?

Summary of Comments:

Overall, reviewers noted that the results closely followed from the analysis plan and did not note major concerns. Some reviewers suggested additional explanation for how the results of the study were related to the study aims, and some reviewers also noted the lack of theoretical framework (as has been raised in response to other charge questions). Other reviewers suggested adding figures, making edits to the tables to note which warning statements were randomly selected for comparison of the revised warning statements to the original TCA warning statements, omitting discussion of findings that are not statistically significant, and other minor changes for clarification.

FDA Response:

We have added detail to the tables (e.g., Table 22) to note which statements were randomly selected for comparison to clarify which revised warning statements were compared to which TCA statements. In the interest of transparency, and in line with the Statistical Analysis Plan developed prior to data collection, we decline to omit findings that are not statistically significant from the study report. Similarly, we decline to perform additional sub-group analyses as one reviewer suggested, as the study was neither designed nor powered for sub-group comparisons (as has also been explained in response to Question 7 above). As explained above in response to Question 9 above, we include summary tables for the results for all participants and across study outcomes but decline to add additional figures. Our response and explanation of the changes to the Study 1 Report resulting from peer reviewer comments related to suggestions on the theoretical framework is presented in response to Question 6 above.

11. Are potential limitations of the study appropriately identified?

Summary of Comments:

The reviewers noted that the Study 1 Report includes a comprehensive list of important limitations, many of which are common to other similar online and pre-implementation studies of warnings and other topics. Reviewers suggested additions to the limitations section to further bolster the report, such as noting that the results will not necessarily translate into behavior, identifying limitations of the composition of the sample if sub-groups analyses are performed, citing similar findings with to other similar studies (i.e., convergent validity) to better demonstrate the strength of these findings, and considering weighting the findings. One comment noted a minor technical issue with study allocation that was already discussed in the Study 1 Report. Other comments focused on the distinctions between understanding the content of the warnings and fully integrating the information into core beliefs.

FDA Response:

As explained in response to previous questions above, we decline to include additional limitations that are not consistent with the aims of the study (e.g., those that focus on behavior). Similarly, as sub-group

analyses were not planned or conducted, there is no need to acknowledge a limitation of the sample composition for those analyses. We decline to weight the findings; weighting a convenience sample that was not designed to be weighted does not necessarily make the data representative in a way that would be useful for additional interpretation. Furthermore, weighting the findings is not consistent with the Statistical Analysis plan and, as discussed in response to Question 7, we have clarified this in the Study 1 Report. We have added some information about how studies of warnings before they are implemented, like the current study, relate to effects of warnings once they have been implemented in response to a reviewer comment. However, this study was designed with very specific aims to support regulatory action for the specific warnings that were developed and tested in this study. Thus, although other studies that we now include in the Study 1 Report in response to a reviewer comment suggest that pre-implementation studies, such as this one, are a very good predictor of eventual effects when warnings are implemented in a country, our focus in this study was more specific to the warning statements being tested. Similarly, the integration of the findings, or how the findings would be used, including the distinction between understanding the content and further integrations into beliefs, are beyond the scope of this study report.

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

Summary of Comments:

Overall, the reviewers noted that the conclusions flow very well from the findings presented in the Study 1 Report. Reviewers also noted that the report appropriately does not go beyond the data, nor does it make broader conclusions about what may result from this study or other research questions not included in the study aims. Some reviewers questioned how some outcomes were prioritized over others in the interpretation of results, and others suggested minor edits to improve clarity in this section.

FDA Response:

The purpose of the study report was to discuss the study, not to make regulatory conclusions. Therefore, the reviewers correctly noted that the conclusions of this study do not go beyond the data in the Study 1 Report and that the report does not discuss the regulatory decisions to be made that are based, in part, on this study, and that the report does not include outcomes (i.e., behavior) that are beyond the scope of this single study and not relevant to the study aims.

Summary of Specific Observations about Study 1:

Five of the reviewers did not provide any specific observations about the study, but provided feedback based on FDA's charge questions as discussed above. One of the reviewers provided several specific observations about Study 1, many of which were requests for additional clarification of minor details, suggesting revisions to complex sections to aid in comprehension, or noting a small number of typos. For example, the reviewer inquired about the "actual amount of statistical power," potentially dropping the outcome of "Factuality," and requested confirmation of one specific result (a TCA statement and a revised statement for which results showed a different number of participants reporting that the statement was new information but identical means on the item assessing "Thinking about Risks").

FDA Response:

We have corrected the typographical errors noted and confirmed that the results of the two statements

inquired about were accurate. As noted elsewhere in our responses, we have added a new figure (Figure 2) to illustrate the study flow and have noted which statements were randomly selected for comparison in tables (e.g., Table 22). With respect to statistical power, that information is already included in Table 13. We have concerns about including post-hoc power calculations (i.e., “actual amount of statistical power”), for reasons summarized in an article focused on this very topic (see Zhang et al., 2019, that states “when [power analysis is] used to indicate power for outcomes already observed, it is not only conceptually flawed but also analytically misleading.”). Because both the outcomes of “Believability” and “Factuality” were included in the Statistical Analysis Plan, we do not agree with dropping either from the Study 1 Report. Further, although we might expect these two outcomes to be correlated, we do not think they are the same construct, and the pattern of results supports this distinction.

Study 2: Experimental Study of Cigarette Warnings

Summary of General Impressions:

Overall, the reviewers agreed that Study 2 was well-conceived and implemented using high-quality methods and analyses. They also agreed that the study was appropriate to address the stated aim of the study, and that the conclusions closely follow from the data. The reviewers also generally agreed that the report is comprehensive and provides clear details about the study. Reviewers highlighted several suggestions for improvements in the clarity and flow of some sections of the report, including requests for information to be added to the report about the measures used and power analysis. Reviewers also provided some overarching comments about a theoretical framework for the outcomes used.

FDA Response:

We appreciate the reviewers' support of the study and their suggestions to further strengthen Study 2 Report. We have responded to the specific comments included in the general impressions in the responses to the relevant questions below. We have streamlined the report as noted in response to Question 1 below and added additional details about the theories that informed the development and selection of study questions as noted in response to Questions 6 and 10 below.

1. Is the document logical and clear?

Summary of Comments:

Reviewers stated that the report was logical and clear. Reviewers suggested that the clarity of the document could be further improved by removing redundant information that was a result of the document including multiple previously stand-alone study documents that had been combined. One reviewer also suggested that this study be better linked to the previous study conducted (Study 1). One comment questioned whether the outcome of interest was in fact behavior change and suggested that more details be added to better explain how the measures were related to the study purpose.

FDA Response:

In response to these comments, Study 2 Report has been streamlined, better integrating the separate sections of the report by eliminating additional cover pages and combining and eliminating redundant text. Additionally, the report is now continuously paginated. In response to the comment questioning whether the outcome of interest was in fact behavior change, we note that behavioral outcomes are neither relevant for this study nor for the warnings being studied, thus they are not discussed in the reports or included in the Background and Purpose sections; therefore, we decline to update to the report. Details of the development of the stimuli (i.e., cigarette health warnings), including an explanation that they were informed by the previous study (Study 1), are included in Section 2.1 (Experimental Design). Additional details about the questions used in the study have been added to Section 2.3 (Instrument Development) and Section 4.1 (Analyses).

2. Does the Executive Summary accurately reflect the content of the overall document?

Summary of Comments:

Overall, reviewers noted that the Executive Summary clearly and logically summarized Study 2. Two reviewers suggested additions to the summary, such as including each of the cigarette health warnings

tested, or again suggesting that the outcomes be linked to behavior change.

FDA Response:

One typo in the sample size for young adult current smokers was corrected (2,071 rather than 2,701). Rather than include all cigarette health warnings tested in the summary, we have referred the reader to Table 1 in the Study 2 Report immediately following the Executive Summary, the location of this information. Additional details about the measures included have now been added to Section 2.3 (Instrument Development) and Section 4.1 (Analyses). Behavioral outcomes are neither relevant for this study nor for the warnings being studied, thus they are not discussed in the Executive Summary or elsewhere in the study report.

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

Summary of Comments:

Reviewers noted that the details provided about the study design, stimuli (i.e., Cigarette Health Warnings & Surgeon General's warnings), sample (i.e., participants included in the study), methods, analysis, and results were thorough and clearly described. Reviewers provided suggestions to add details better linking this study to the previously conducted Study 1, to provide additional details about the development of the stimuli (i.e., Cigarette Health Warnings), to describe attrition (i.e., drop-out of participants in the study) over the three sessions of the study, and to better describe the motivation for the study beyond to support the rulemaking.

FDA Response:

In response to these comments, we have added details about the source of the measures used, as noted in response to Question 2 above. Similarly, we have added details about the development of the stimuli (i.e., cigarette health warnings). To address the comment about sample attrition, we have added details to Section 3.2 (Disposition of Sample) explaining that the balance of sociodemographic and tobacco use characteristics was examined by condition and across the three study sessions. The balance of those sociodemographic and tobacco use characteristics did not vary significantly ($p < 0.05$) in any of these analyses. This analysis is now described in the study report and demonstrates that although there was participant drop-out from the study, as can be expected with any study over multiple sessions, that differential attrition (i.e., participants with specific demographic or tobacco use status characteristics did not drop out of the study more than those with other characteristics that could influence the study results at Sessions 2 and 3) is not a concern in this study. Finally, the Study 2 Report accurately describes the motivation for the study specific to the regulatory action it will inform, and we disagree that additional justification for this study is needed, consequently none has been added to the study report.

4. Is the methodology used appropriate to address the study's purpose?

Summary of Comments:

Overall, the reviewers noted that the methodology is appropriate and strong, especially in the context of the study's purpose and the common limitations of such studies. Reviewers suggested adding additional information about the decision for the follow-up timing, the additional exposure to the stimuli (i.e., the Cigarette Health Warnings and Surgeon General's Warnings) at Session 2, justification for the effect sizes

used in the power analysis, and attrition. Additionally, two reviewers commented on the use of repeated measures for assessment of the health beliefs and noted concerns that it could induce bias for that outcome.

FDA Response:

In response to these comments, we have added information about the second exposure to the stimuli as well as a justification for the timing of the follow-up session. We have also added information explaining that the effect sizes used in the power calculation were drawn from prior similar studies as is standard practice. As explained in the response to Question 3 above, we have addressed attrition in the report by adding details in Section 3.2 (Disposition of Sample). Those new details discuss the balance of sociodemographic and tobacco use characteristics by conditions and across sessions. Given the additional analyses showing a lack of difference in attrition by demographic and tobacco use characteristics, differential attrition is not a concern with this study. With respect to the concern about statistical power, as with Study 1, we have concerns about including post-hoc power calculations, for reasons summarized in an article focused on this very topic in Zhang et al., 2019, that states “when [power analysis is] used to indicate power for outcomes already observed, it is not only conceptually flawed but also analytically misleading.” Accordingly, we decline to add this information to the Study 1 Report.

As suggested by a reviewer, we have added an explanation in Section 2.1 (Experimental Design) that the multiple exposures to the warnings was done to better approximate real-world conditions, thus improving the external validity of the study. As for potential response bias that could have been induced through the repeated measurement of health beliefs whereby the questions prime the responses, we acknowledge that there are alternate ways to design the study but that each study has a set of limitations including the feasibility of the design. Due to the long latency between Sessions 2 and 3, the issue of response bias is minimized. Additionally, because this approach of multiple exposure to the warnings and the timing between the sessions was consistent across all conditions, any potential effect would have been observed in all conditions. Therefore, the effect of the repeated measurements would be accounted for with the “difference in difference” analytic approach used for the health belief items such that the difference between the treatment and control would still reflect the effect and the results would not be biased.

5. Are the stimuli used appropriate given the study's purpose?

Summary of Comments:

Overall, reviewers noted that the stimuli (i.e., cigarette health warnings) were appropriate and replicate how they would appear in the real world. Reviewers requested additional details about how the warning images were developed, how the mock advertisement was developed, how the results from Study 1 informed the selection of the warnings to be tested in this study, and also requested that an example of the pack be included in addition to the “flattened illustration” that was included in the previous version of the Study 1 Report.

FDA Response:

In response to these comments, we have added details about the development of the images used in the cigarette health warnings in the Executive Summary as well as in Section 2.1 (Experimental Design). Similarly, we have added details about the development of the mock advertisement to Section 2.1

(Experimental Design) and added three examples of the warnings displayed on the mock pack in Appendix A (Study Stimuli). As for the decisions made for inclusion of specific warnings in this study, we note that those decisions were informed in part by the results of Study 1, which we now explain in the Executive Summary.

6. Are the outcomes measured appropriate given the study's purpose?

Summary of Comments:

Three reviewers noted that the measures used were appropriate, and some also provided suggestions for other measures that could have also been included in Study 2. Some reviewers noted that a theoretical framework or model to better understand how the measures relate to the study's purpose was not included in the report and requested that a framework/model be included. Reviewers also asked that more details about the selection and validity of the measures used be added to the report. Additionally, some reviewers requested clarification about participants' interpretation of item B1 ("Before today, had you heard about the specific smoking-related health effect described in the warning"), suggested that additional analyses should have been conducted to rule out a general halo effect (the effect of exposure to a warnings on one topic having an effect on outcomes for a similar, but unrelated topic), and asked questions about how the relative lower levels of perceived factuality may translate into acceptance of the warnings once implemented.

FDA Response:

Details about the design of Study 2, including the theories that informed the selection of questions used, has been added to Section 2.3 (Instrument Development). The source of the items used is now noted both in Section 2.3 (Instrument Development) and Section 4.1 (Measures and Coding). The item "Before today, had you heard about the specific smoking-related health effect described in the warning" (B1) was asked once for each cigarette health warning; the intent of this item, as the reviewer noted, was to assess if the information about any of the health effects included in each warning was new information to participants, as now further explained in Section 4.1 (Measures and Coding). Participants in this study were only exposed to one warning (with a single warning statement) and outcomes focused on the content of that warning. Thus, although the study included multiple treatment conditions, with participants in each exposed to a different warning, the main focus of the analysis was on the effect of each individual warning on outcomes related to that warning. Even if there were a halo effect on beliefs other than those directly related to the warning participants were exposed to, such a finding would not diminish the findings of the study for that warning. We note that the inverse association between the "novelty" of a health warning and its believability/perceived factuality is a common finding in pre-implementation studies (e.g., Hammond Fong 2006; Kennedy 2012; Hammond 2011). Importantly, such findings in pre-implementation studies do not suggest that the warnings will not have the intended effects. We also note, as described in the Study 2 Report, that the absolute level of perceived factuality was very high for all cigarette health warnings tested, even though many of the health conditions included in them were on novel topics.

7. Are the study participants included appropriate given the study's purpose?

Summary of Comments:

Overall, reviewers commented that the study participants were appropriate to address the research

questions and purpose of the study, and some noted limitations common to the use of a convenience sample. Reviewers requested additional details about the recruitment of participants, additional clarifications about the criteria for determining certain tobacco use status groups in recruitment, and justification for the inclusion of certain groups (e.g., non-smoking adults).

FDA Response:

In response to these comments, we have added clarification to the criteria for inclusion in study subgroups in Table 3 of the Study 2 Report better describing inclusion criteria by tobacco use status. Also, in Section 2.2 (Sampling Frame and Sampling Methodology), we have added details about the recruitment process, including what potential participants were told when originally contacted. We also have added an explanation for inclusion of the various subgroups in an effort to obtain a broad and heterogenous study sample, but note that behavior is not an outcome informative for this study. For these reasons, all the groups (including adult non-smokers) included in the study contribute to FDA's assessment of understanding of the negative health consequences of smoking in response to exposure to the cigarette health warnings tested.

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

Summary of Comments:

Overall, most reviewers noted that the analytic approach was solid and appropriate. Reviewers noted the potential for the multiple exposure to the health beliefs questions priming respondents (as in Question 4 above), some reviewers asked for more clarification on the analytic approach for health beliefs, and one reviewer suggested we test each cigarette health warning against each of the four individual Surgeon General's warnings rather than an average of the Surgeon General's warnings.

FDA Response:

As described in our response to Question 4 above, we do not believe that the repeated measurement of health beliefs would differentially affect any one condition, and thus the "difference in difference" approach still yields valid results. As was already explained in Section 4.3.3 (Phase 2 Hypotheses and Analyses), although we measured many health beliefs, those used in each analysis for each condition (i.e., cigarette health warnings with a specific warning statement and paired concordant image) are specific to the treatment condition—the belief questions are all congruent to the content of the warning. Finally, we decline to conduct additional analyses to compare each individual cigarette health warning against each Surgeon General's warning. The decision to use a combined control group represents a more conservative approach. Unlike Study 1, where slight changes in warning statements were compared, this study's goal was to approximate the potential effect of these cigarette health warnings on understanding when changing from the current *status quo* (i.e., Surgeon General's warnings) to the proposed warnings (i.e., cigarette health warnings). Therefore, a one-to-one comparison was not necessary, nor would such a comparison of one cigarette health warning against one Surgeon General's warning provide information that would be useful for FDA in assessing the effect of the cigarette health warnings on understanding of the negative health consequences of smoking in support of this effort.

9. Are results presented consistent with the analytic approach?

Summary of Comments:

The reviewers noted that the results are consistent with the analytic approach and that the results clearly flow from the analytic approach. One reviewer noted the strength of the findings even after two weeks between sessions suggests that the warnings were particularly strong as such a finding of an effect after that length of delay with no additional intervention would normally not be expected in other similar studies. Another reviewer suggested that the results be reordered to reflect the relative strength of the findings across outcomes.

FDA Response:

We decline to reorder the results within the tables based on relative strength of findings; these tables follow from the Statistical Analysis Plan and are ordered by condition (i.e., the assignment of what warning participants were exposed to in the study) across outcomes for consistency. Additionally, the goal of this study was not to compare (or rank order) each cigarette health warning against the other cigarette health warnings tested, but rather to compare each cigarette health warning to the Surgeon General's warnings. We agree with the reviewer that the finding of supportive results – even after a two-week delay without additional exposure to the cigarette health warning between Sessions 2 and 3 – suggests that even with the relatively brief exposure to cigarette health warnings in this study that these warnings are effective in communicating information to improve understanding about the negative health consequences of smoking.

10. Are there any concerns with the results presented?

Summary of Comments:

Reviewers noted questions about whether randomization was successful, differential attrition, and the lack of statistical power to test for sub-group differences. One reviewer suggested that this report be combined with the Study 1 Report, and a request to include a theoretical framework to contextualize the findings.

FDA Response:

To address the comment about randomization and attrition, we have added details to Section 3.2 (Disposition of Sample) that discusses the balance of sociodemographic and tobacco use characteristics by conditions and across sessions and note that they did not vary significantly ($p < 0.05$) in any of these analyses, eliminating potential concerns about randomization and further supporting the results obtained from this study. As noted in the Study 2 Report, the study was neither designed nor powered to detect within-group differences, and the recruitment by subgroup was to ensure a broad and heterogeneous study sample which was achieved. As noted in response to previous questions, additional details about the measures with respect to how they relate to the study purpose has now been added to Section 2.3 (Instrument Development) and Section 4.1 (Measures and Coding), including the theories that were considered when designing the study and selecting the study measures. We decline to combine this study report with that of Study 1, but we have added information to both study reports to better link the studies as described in response to questions for both Study 1 and Study 2 above.

11. Are potential limitations of the study appropriately identified?

Summary of Comments:

Overall, reviewers noted that the discussion of limitations was appropriate. As in previous questions, reviewers asked about differential attrition and the potential for priming health beliefs through repeated assessment. Additionally, reviewers suggested that additional information about convergent validity be added along with a comparison of Study 1's convenience sample to more representative samples.

FDA Response:

As noted previously, we have added details about the balance of sociodemographic and tobacco use characteristics by conditions and across sessions and note that they did not vary significantly ($p < 0.05$) in any of these analyses, reducing potential concerns about differential attrition and subsequent bias in the results. We decline to conduct additional analyses to examine attrition as a function of Session 1 beliefs – with a successful randomization (as noted in the responses to Question 4 and Question 10 above in this section), the distribution of beliefs about the health consequences of smoking prior to Session 1 should be evenly distributed between the different conditions. Additionally, the "difference in difference" analytic approach accounts for Session 1 health beliefs, thus minimizing any potential effect of differences in health beliefs by condition at Session 1 or differential attrition across study sessions. As noted in response to previous questions, we do not believe the repeated measurement of health beliefs would differentially affect any one condition and thus the "difference in difference" approach still yields valid results.

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

Summary of Comments:

Overall, reviewers noted that the conclusions in the study report are well supported by the results and data presented. A few reviewers suggested that the non-significant results not be included, that the warnings be ranked by strength of results or directly linked to regulatory decisions, and that an overarching theoretical framework would be helpful to interpret the results. Some reviewers also offered their own additional interpretation of some of the findings or suggested that additional exploratory analyses be included.

FDA Response:

In the interest of transparency, we decline to exclude the findings that were not statistically significant from the conclusions. As for the comment about ranking the results, as explained in response to Question 8 and Question 9 above, the goal of this study was not to compare (or rank order) each cigarette health warning against the other cigarette health warnings, but rather to compare each cigarette health warning to the Surgeon General's warnings. Thus, we decline to include such information in the study report.

Summary of Specific Observations about Study 2:

Five of the reviewers did not provide any specific observations about the study, but provided feedback based on FDA's charge questions as discussed above. One reviewer provided some specific suggestions, including suggestions to add details to a table, to add details about the duration of exposure to the

stimuli, clarifying about the use of mobile devices for data collection, simplification of some sections, and to provide additional details about the number of participants rejected by the online panel for quality control purposes. That reviewer also requested clarification about why the average score for the control groups on some analyses were different from each other, when the average score of the Surgeon General's warnings was used as the control comparison, as well as other issues (e.g., post-hoc power calculations) already discussed in previous questions.

FDA Response:

In response to the reviewer's specific comments, we have added the condition number to Table 1 and added the median duration of stimuli exposure toward the end of Section 2.1 (Experimental Design). We have also added clarification that participants were eligible only if they completed the study on a laptop or desktop computer in Section 2.2 (Sampling Frame and Sampling Methodology). In response to this reviewer's specific comments and other comments in other questions, numerous sections have been simplified for clarity. Because we do not have the information on the number of participants rejected by the panel for quality control purposes, we have not added any details to the report beyond the extensive description of that process already included which reviewers noted appeared very strong. Finally, to answer the reviewer's question about the different means in the control condition: it is because the outcomes are different for each analysis on those specific health beliefs questions. Although the control condition is the pooled average across participants assigned to view the Surgeon General's warnings, the mean represents their scores only on those health belief items included in that specific analysis. This is explained in more detail in Section 4.3.3 (Phase 2 Hypotheses and Analyses).

References for FDA Response

- Eisinga, R., M. T. Grotenhuis, and B. Pelzer. 2013. "The Reliability of a Two-Item Scale: Pearson, Cronbach, or Spearman-Brown?" *International Journal of Public Health* 58 (4): 1–6.
- Hammond, D., G. T. Fong, A. McNeill, et al. "Effectiveness of Cigarette Warning Labels in Informing Smokers About the Risks of Smoking: Findings from the International Tobacco Control (ITC) Four Country Survey." *Tobacco Control*, 15(Suppl. 3):iii19-iii25, 2006.
- Hammond, D. "Health Warning Messages on Tobacco Products: A Review." *Tobacco Control*, 20(5):327-337, 2011.
- Kennedy, R. D., M. M. Spafford, I. Behm, et al. "Positive Impact of Australian 'Blindness' Tobacco Warning Labels: Findings from the ITC Four Country Survey." *Clinical and Experimental Optometry*, 95(6):590-598, 2012.
- Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. "Post hoc power analysis: is it an informative and meaningful analysis?" *General psychiatry*, 32(4): e100069, 2019.

APPENDIX I - INDIVIDUAL PEER REVIEWER COMMENTS

STUDY 1

General Impressions	
REVIEWER	COMMENT
Reviewer #1	<p>Overall, this represents high quality statistical design to address key questions about a revised set of warning labels compared to the set of nine that were outlined in the TCA. The statistical detail presented is of outstanding quality and thoroughly documented. However, the lack of an appropriate theoretical framework to the document means that it is not clear how the nature of the different outcomes being addressed relate to each other.</p> <p>The report outlines the purpose of the study as the identification of whether the proposed revised warning labels will likely lead to higher levels of public understandings of the risks associated with tobacco use than is achieved with the TCA warnings. The report notes that there are two questions needed to meet this goal: do the revised statements lead to new knowledge (question A-1) and do they lead to new learning (question A-2). Quite appropriately, these two questions are primary aims. However, the reader needs a theoretical framework to understand the rationale for including the other two primary aims and the four secondary aims. Indeed, is the overall purpose just to increase the public understanding or is it to increase the public understanding in a way that will motivate more behavior change (reduced uptake and increased quitting)?</p> <p>Indeed, from the executive summary, there is no indication that there are eight outcomes being investigated in this study, let alone how each might relate to the purpose of the study. The case for including outcome #3 (thinking about the risks) might be expected to be that receiving new knowledge that relates to new learning should be most important when these two translate into cognitions on risks.</p> <p>It is particularly important that the reader understands the importance of inducing a change in health beliefs. Indeed, health beliefs did not change with different textual warnings in Study 1 but were responsive to graphic warnings in Study 2. This is a critical finding from these two studies as it shows how graphic warnings have an added effect to text warnings. Indeed, this indicates that the two studies should be presented as a single report. This could be accompanied by a combined methodology report as an appendix. Currently, there is a lot of replication in the methodology reports.</p> <p>Without an appropriate theoretical framework and expanded study purpose, this study should be limited to addressing only three of the study outcomes. The remaining outcomes take up considerable space in the report (with appropriate analyses and results), but not even addressed in the executive summary.</p>

General Impressions	
REVIEWER	COMMENT
	This reviewer does not think that the report should be limited to the three obvious outcomes from the current specification of the purpose of the study. Rather the outcomes as outlined are appropriate to the real issues at stake. It is the study purpose that needs some expanding as suggested above. In comments on the second report, a suggested theoretical framework is outlined that contains each of the eight aims investigated in the two studies.
Reviewer #2	My primary comments are indicated below, but a few stand out: 1. The need for some overarching conceptual framework to bring coherence to the outcomes assessed and interpretation of results; 2. stronger justification for the measures used, including information on the validity of measures, especially novel ones; 3. consideration of prior research to determine meaningful effect sizes and power; 4. stronger justification for the phase 2 belief assessment; 5. stronger justification for the analytic approaches and inclusion of sensitivity analyses to evaluate the consistency of results under different specifications (e.g., approach to randomly selecting comparative TCA warning).

General Impressions	
REVIEWER	COMMENT
Reviewer #3	<p>The comments that I have made in response to the 12 charge questions include various elements that would fall under “general impressions.” I draw them out here in a separate answer.</p> <p>Both studies are very well done in terms of design and data analysis. The designs selected provide for an appropriate control groups which are the current standards for warnings (Surgeon General in Study 2) or immediate past selections (initial FDA warning labels in Study 1). The data analysis plan is strong and straight forward. It is careful on statistical treatment of the quality of data (e.g. continuous vs. rank order) and especially strong on correcting for experiment-wise error and power considerations. Both documents do a good job in communicating their procedures and the results except as noted regarding Study 1’s complex and difficult design that requires some supplementation with a clear and effective graphical description. If one accepts that the operational measures of learning and novelty are valid measures of their underlying constructs, then there is a clear picture that the new warning texts and warning labels are effective relative to their comparison.</p> <p>However, I am concerned that the measures deployed – perceived novelty and awareness – are not convincing measures of the underlying constructs that the research is targeting. In Study 1, the researchers do employ a measure of believability and of facticity (opinion vs. fact) finding that the new labels are less believable. In Study 2 the believability measure is not present even though it was diagnostic in Study 1. Both the believability and facticity measures underscore the fact that the new warnings may not be accepted by the target audience. The authors are well aware of this and comment on it in both studies. But coupled with self-report measures of learning and novelty (awareness), the lower levels of acceptance of the labels reduce the overall impact of the results. The implicit rejoinder in the data to the argument that the results are not so convincing is that the warnings affect the acceptance of negative health consequences (i.e. beliefs) in the warning label conditions versus the SG warnings condition and do so over time. But as I note in my comments on Study 2, two concerns arise about these findings. The first is that asking beliefs at baseline before message exposure taints the message processing by focusing respondents’ attention to the messaging in ways that privilege the beliefs being targeted. Second, the beliefs are measured three times reinforcing the warning labels’ content. Third, it is not clear why all beliefs would be affected by a specific warning label as opposed to a more targeted set of outcomes wherein warning label X affects beliefs related to warning label X but not beliefs Y and Z.</p> <p>In the end, these are both very carefully done studies that adhere closely to the data that has been gathered. This reviewer is raising interpretive considerations that essentially claim that the overall set of results are less convincing than they might be had the same constructs been operationalized differently and slight changes in the design in Study 2 been implemented.</p>

General Impressions	
REVIEWER	COMMENT
Reviewer #4	<p>This extremely clearly written report documents an online experiment that was conducted to evaluate the efficacy of new smoking warning statements (versus older statements) for improving knowledge of the health risks of smoking. The information is accurately reported, and a very good amount of detail is provided to support transparency in reporting the methods and results. The methods mostly follow logically from the stated goals of the research, and the complex experimental design is suited to the task at hand. The report carefully describes power analysis, rationale for taking specific analytic approaches, and drawing conclusions based on the results (essentially a restatement of those results). The report was lacking in several areas and would benefit from: providing more detail and rationale on the need for warning labels and conceptually, the link between warnings and behavior or behavior change; being clearer on the origin of the assessments and their validity; and providing more information on vendor and sampling issues. Presentation of the results would be improved with more frequent summaries and inclusion of graphic presentation of key results (e.g., for primary outcomes). It is unfortunate that the study was not powered to detect differences in different subgroups or by age; warning statements are not received the same by some segments of the population and to the extent that the warning labels need to cover a wide population base, it would be important to know what worked and for whom. The key limitation, for this reviewer, is in the sampling frame and sampling design. The report acknowledges the limitations with employing a convenience sample (albeit one that is very diverse), but the robustness of the conclusions depends, in part, on the extent to which they generalize to the population as a whole; and this is just not as easily possible with a convenience sample. Additional information on vendor choice, features of their panel, and decisions to not utilize weighting or some other sampling scheme would help to contextualize the results.</p>
Reviewer #5	None provided.
Reviewer #6	<p>This study was well designed to assess whether the revised warning statements would meet the mandate to increase understanding of the risks associated with smoking. It is clear why the main outcomes of new knowledge and learning were included and prioritized because of the statutory obligation in the TCA, and the writers mention that novelty is important for drawing attention. However, there is little discussion of theory in the selection of the other measures that have been included and analyzed. For example, the cognitive elaboration measure (“think about risks”) and the potential importance it may have in retention of information included on the labels, is not included in the report.</p> <p>Each of the individual choices for the stimuli and comparisons make sense for each individual label, and for each phase of the study. The choices are clearly explained, and the logic is sound. But because the labels are not parallel in wording (some with multiple health effects, others with only one), and because some of the labels have clear controls in the statutory labels, while others don’t, it makes the results section quite dense. It hampers an easy comparison between the categories of labels. In the summary section, it may help to expand on the tie between novel information and believability. It is common for people to view new information with skepticism, this should not be seen as a potential drawback. Table 4-1 was particularly helpful in bringing a large amount of information together in an easily digestible</p>

General Impressions	
REVIEWER	COMMENT
	way, but the summary section does not help the reader synthesize the findings or leave the reader with an overall take-home message.

CHARGE QUESTION 1: <i>Is the document logical and clear?</i>	
REVIEWER	COMMENT
Reviewer #1	The document is written in clear English and for the most part the presentation is logical and concise. The problem is that it is not clear why there are eight different aims addressed in this study. It is well into the document before the reader learns that there are eight aims and, for many of them, the relevance to the study purpose is far from clear. It needs a theoretical framework from which the relevance of each aim is presented.
Reviewer #2	Yes. However, to make it even clearer, I recommend that the authors integrate a figure to illustrate the protocol steps, including when stimuli were shown and the timing of specific measures. I recommend eliminating redundancy between the methodology report and the results report, which includes a LOT of the same information.
Reviewer #3	Yes, the document is logical and clear in many respects, most in fact. However, the design of Study 1 is very complicated and unusual, and it took this reviewer multiple readings before it became clear what was taking place and what the exact nature of the protocol was. I would strongly recommend a visual presentation of the protocol to help readers understand what the sequence was and to understand the kinds of questions being asked at various stages along with the warnings to which respondents were exposed at the different stages and phases of the protocol.
Reviewer #4	The report of Study 1 is very well-written, and clearly described. Some additional details on study rationale, features of the design, and analyses would have improved the depth and quality of the report, however (these are described in subsequent sections of these review comments).
Reviewer #5	The document is pretty logical, but I believe that clarity could be improved. The study designs are complex given the research questions. Although it would increase length by a few pages, I think some extra information would be helpful. The report reads like it was written by someone who is very succinct, but who knows this study really well and forgets that others do not know it as well. Some additional phrases and explanation, even if redundant, will help the reader navigate this complex and important study. I also believe that the 3 'reports' in this Study could be condensed and one could remove all of the cover pages and do one larger Table of Contents and you remove some of the redundant recap summaries and the Table of Contents that are not necessary if this is one large report instead of 3 reports.

CHARGE QUESTION 1: <i>Is the document logical and clear?</i>	
REVIEWER	COMMENT
Reviewer #6	Yes. There is some redundancy between sections, but it does allow for each section to be read and digested independently. I understand the reasoning behind using linear and ordinal regression for the health beliefs, but it makes the findings more burdensome to digest and compare.

CHARGE QUESTION 2: <i>Does the executive summary accurately reflect the content of the overall document?</i>	
REVIEWER	COMMENT
Reviewer #1	Similar to the above statement, the executive summary does present how the findings of the study relate to the purpose of the study. This is written in clear and understandable English. However, this means that the numerous additional aims that are addressed in this study are not mentioned at all. The authors can’t have it both ways. Either make the case for including the additional aims or delete the analyses relating to them from the report.
Reviewer #2	<p>Clarify that the stimuli are just text (i.e., no imagery/pictures). Otherwise, it is a bit confusing since the title of the report indicates “graphic” warnings.</p> <p>Add a justification for the three primary outcomes selected, including an indication of the constructs that they are supposed to measure. I would also mention the “secondary” outcomes since they are generally important constructs for evaluating message effects in a brief experiment like this (see comments in the outcome section for concerns about designating some measures as secondary without any theoretical or empirical justification).</p> <p>For phase 2, clarify this statement so that the reader understands what it means without going to the methods section: “Participants assigned to the treatment conditions viewed one of several different combinations of 9 revised warning statements.”</p> <p>Not clear what this means: “After viewing the 9 warning statements, all participants answered questions about their beliefs about the link between smoking and each of the health consequences presented in the warning statements.” Did everyone answer the same questions, some of which included health effects that were on the warnings that they evaluated and some of which were not? Or did people just evaluate outcomes that were on the warnings they evaluated. What was done has implications for the analysis and its interpretation (see below).</p> <p>For phase 2, did the control group get the same health belief questions asked as in phase 1? Clarify.</p> <p>Include descriptive information for the 4 of 15 revised statements that were higher than the standard warnings on thinking about risks given its importance in predicting cessation.</p> <p>Clarify what is being compared in the health beliefs summary at the end of the results section.</p>

CHARGE QUESTION 2: Does the executive summary accurately reflect the content of the overall document?	
REVIEWER	COMMENT
Reviewer #3	Yes, I thought the executive summary was a good representation of the positive findings and procedures discussed in much greater detail in the subsections of the ensuing documents. The only respect in which I thought the executive summary was a little bit misleading -- and this may reflect more of my views about the research than others who read the document differently -- is the presentation of findings about believability. These can be construed as negative and/or problematic for the research and they should be a part of the executive summary. I’m pretty sure that the authors who prepared the work do not agree. They have offered some commentary in Study 1 and definitely in Study 2 about concerns that might be raised about believability and factualness. More about this below for Studies 1 and 2.
Reviewer #4	Yes. The executive summary is quite detailed and provides sufficient information that accurately reflects the report as a whole.
Reviewer #5	The summary does a pretty good job of summarizing the overall document.
Reviewer #6	Yes. The executive summary provides a good overview of the purpose, methods and findings. Given the level of detail provided in the full report, it could likely be simplified further to assist an audience not well versed in scientific nuance and methodology.

CHARGE QUESTION 3: Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?	
REVIEWER	COMMENT
Reviewer #1	This reader was very impressed with the presentation of the overall study design, the choice of sample and experimental methods used for this study. As above, the problem was the additional aims that were not well justified and, indeed, the study did not have the power to address all of them. This puts at least one of these aims into the ‘exploratory aim’ category. While it may meet some internal needs of the FDA to obtain this information at the same time as conducting this study, it is not clear why the final aim should be included in this report.
Reviewer #2	The information provided is mostly sufficient. I would add the following material to enhance the clarity of the information presented: <ul style="list-style-type: none"> • Integrate a figure to illustrate the protocol steps, including when stimuli were shown and when specific measures were used. • Include a figure with the actual stimuli as shown to participants. • Provide more information on how revised warning statements were selected Include more information about measurement (see comments in the next section)

CHARGE QUESTION 3: <i>Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?</i>	
REVIEWER	COMMENT
Reviewer #3	Yes, I thought that there was plenty of detailed information about the design, stimuli, sampling methods, and analysis both in the two documents that made up Study 1 plus the supplementary materials. If I was looking for any additional information it would have been about the measures of susceptibility to smoking among adolescents which I couldn’t initially find although it does turn up later in the Study 1 and in Study 2. Second, as I mentioned above, I think that the design of the study is complex and difficult to understand and could profit from a careful visual presentation of the protocol. It’s not that the information is not present; it is. However, it’s just difficult to fully comprehend what. My final point concerns the stimulus materials that are generated for testing. There is some brief discussion of how these materials were generated and while it may not be of value or of interest to understand the sifting and winnowing process here, this reviewer was a bit perplexed about the sources and topics that generated the new warnings.
Reviewer #4	<p>The study design, analysis, and results are described and are sufficiently detailed. Aspects of the sample, stimuli, and methods are less detailed. Some of the information needed includes (these are described in more detail in subsequent sections of this review): more conceptual background and motivation for studying warning labels; the value of warning labels; and need for new labels (beyond statutory requirements); formative work that went into development of the new warning labels; decision rules for segmenting the sample (e.g., using days smoked in past month; excluding committed never-smokers) and for choice of vendor (given the important limitations inherent in the non-representative sample); and validity of key measures used.</p> <p>The study is framed in a very practical manner, i.e., TCA developed text warnings, this study is designed to improve on those warnings by providing additional detail. As important as this feature is for ease of communication, it is lacking in conceptual and empirical motivation; the result is that the rationale for the study comes across as thin. An enhanced background section (it does not necessarily require pages of dense theory and analysis of previous research) – one that reviews relevant literature on health warnings, particularly the efficacy of such warnings (and need to change them periodically/frequently); importance of warnings for improving knowledge; importance of knowledge to motivating behavior change, etc. – would really improve the scholarship of the report and motivation for the study. Such information would also help to motivate the study hypotheses (which appear later).</p>
Reviewer #5	The detail is nearly sufficient. I had to read some sections 3 times to understand them and that is not ideal. At the most basic level, any study report or publication should provide enough detail to fully replicate this study. In this case, I think that I could fully replicate that study by reading this report. My primary critique is to help the reader with short reminders and rationales about the study design or key findings, especially given the many nuances of these studies. See specific suggestions in the table below.
Reviewer #6	Yes. The methods section was thorough and detailed. I do think that a quick summary of the planned vs. actual assignment to condition error would be useful in the results section. The explanation of the effects of the error were clear.

CHARGE QUESTION 4: <i>Is the methodology used appropriate to address the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #1	The study is well designed and is appropriate to address the study’s purpose.
Reviewer #2	<p>In section 2.1, experimental design: The rationale for the approach used for the treatment conditions should be spelled out, as it is not clear. In particular, justify showing only one revised statement and eight of the original warning statements.</p> <p>Provide a justification for evaluating beliefs in two separate moments with different measurement approaches.</p> <p>In the methods section or the limitations, the authors could do a better job of citing literature indicating the consistency of results from online studies of warning responses and those that use either physical packs in brief experiments or that compare responses to warnings in online studies with those smokers have after policies are rolled out (e.g., Hammond D, Thrasher JF, Reid JL, Driezen P, Boudreau C, Arillo-Santillán E. Perceived effectiveness of pictorial health warnings among Mexican youth and adults: A population-level intervention with potential to reduce tobacco-related inequities. <i>Cancer Causes and Control</i>. 23 (Supp1): 57-67. 2012; Huang L, Thrasher JF, Reid J, Hammond D. Predictive and external validity of a pre-market study to determine the most effective pictorial health warning label content for cigarette packages. <i>Nicotine & Tobacco Research</i>. 18(5):1376-1381. 2016; Thrasher JF, Carpenter M, Andrews JO, Gray KM, Alberg AJ, Navarro A, Friedman DB, Cummings KM. Cigarette warning label policy alternatives and smoking-related health disparities. <i>American Journal of Preventive Medicine</i>. 43(6):590–600. 2012; Hammond D, Reid JL, Driezen P, Boudreau C. Pictorial health warnings on cigarette packs in the United States: an experimental evaluation of the proposed FDA warnings. <i>Nicotine Tob Res</i>. 2013 Jan;15(1):93-102).</p> <p>Power calculations that adjust for false discovery rates (Benjamini & Hochberg, 1995) are important given the great number of comparisons made. However, the authors would ideally provide citations and empirical justification for the anticipated effect size (difference of 0.5 and standard deviation of 1) given the substantial body of research in this area. Otherwise, it is hard to determine if the study is over or underpowered. This information will also be useful when considering the unanticipated equal allocation of sample to treatment and control groups, especially given the significantly lower power found for the equal allocation scenario relative to the optimized allocation with a larger control group. The authors may be able to address this concern by using the literature to show the effect size, including meta-analyses.</p>

CHARGE QUESTION 4: <i>Is the methodology used appropriate to address the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #3	<p>Sampling procedures seem to be reasonably well presented and there appears to be sufficient care in monitoring the attentiveness and legitimacy of individual respondents whether adult or adolescent. The sample’s vendor appears to be especially careful about this. Kudos here.</p> <p>The key analysis appears to be a comparison between responses to the new statements in the 16 experimental conditions to the old statements in the control condition on criteria such as newness, perceived learning, and links between beliefs and outcome. Although this sounds like it makes sense, it’s actually a weak criterion because the old statements have been rejected precisely because they are already well understood, and the new ones selected precisely because they are not so well known. I think I would’ve been more impressed if there was a no-exposure control and/or an inaccurate belief control to show that the new information is better believed than both the old information and the information that is a part of inaccurate claims.</p> <p>The analysis plan for the knowledge learning and thinking about outcomes seems to make a fair amount of sense in that there is a comparison between items that are new and old but roughly matched on content between the treatment and control groups. This allows one to infer that the new version is effective on these three measures versus similar content in comparison to the old version. Where there is none whose content is comparable to that of the new than a random comparison is made between the new and the old which could easily overstate the effectiveness of the new. The authors recognize this, but it nevertheless does run the possibility of overstating the result.</p>
Reviewer #4	<p>The study design is complex but given some of the built-in limitations of internet survey panels that challenge internal validity of experiments, the design was well-appropriated for purposes of this study. In other words, the randomization scheme and viewing allocations and random ordering of study stimuli in the control and experimental conditions (as well as the data security checks) improved internal validity and helped to overcome the limitations with internet panels.</p>
Reviewer #5	<p>I believe that the randomized design and other methods appropriate to answer the study purpose.</p>
Reviewer #6	<p>Yes. One drawback to the method is that some of the labels had multiple health beliefs and others only one. Because these are treated differently in analysis (some linear, some ordinal), it makes comparisons about label features more difficult.</p>

CHARGE QUESTION 5: <i>Are the stimuli used appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #1	<p>The experimental design is excellent and allows the assessment of the study aims related to the study purpose. The rationale for presentation of study stimuli is well presented.</p>

CHARGE QUESTION 5: <i>Are the stimuli used appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #2	<p>Yes. The stimuli seem pretty standard for online studies of warnings; however, it would be clearer if the report showed example stimuli to illustrate what the stimuli looked like to participants (which they do not currently do).</p> <p>Topics for the new warnings generally capture outcomes about which there is likely to be lower awareness in the general population. How these topics were selected should be clarified, as there is no information about this issue in the report.</p>
Reviewer #3	<p>The FDA has chosen to study lesser-known health consequences in the new warning statements. The argument here is that more well-established and better known health consequences are already well-known and only need reinforcement not creation or conversion. The problem with teaching people something that is new is that “new information or claims” run the risk of being unpersuasive, and indeed raise skepticism about the new information given its novelty. We have run across this problem in several different contexts where what is new is less likely to be believed. This is a major issue here in Study 1 and in Study 2.</p>
Reviewer #4	<p>The study has a very focused purpose, and to the extent that the experimental stimuli reflect warnings that will actually be used and (potentially) implemented, these stimuli are appropriate to the task at hand. It makes logical sense to compare these newer warnings to the TCA warnings provided to evaluate whether those newer warnings improve knowledge and retention over and above the prior ones.</p> <p>Information on the development of the revised statements is lacking but is needed to evaluate their conceptual adequacy and literacy level, among other issues. More description of the formative work that drove the development of these new warnings is needed.</p>
Reviewer #5	<p>The stimuli, warning statements, were appropriate and captured the key health risks of using cigarettes that smokers and at-risk youth should be aware of.</p>
Reviewer #6	<p>Yes. The stimuli are appropriate. Using the statutory and revised statements allow for a direct comparison and allows for conclusions about which ones are better suited to increase knowledge about the harms of tobacco.</p>

CHARGE QUESTION 6: <i>Are the outcomes measured appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #1	<p>No. Only a subset of the aims are appropriate for the study’s purpose, as it is currently framed. This reviewer suggests that the wording of the study purpose be qualified so that the new understandings relate to a potential change in smoking behavior. The study lacks a theoretical framework section that demonstrates why each of the outcomes measured is relevant to the study’s purpose.</p>
Reviewer #2	<p>The conceptual model(s) that orients this study is underdetermined and never clearly defined. At the end of a single sentence, the authors cite a bunch of studies to support their measurement strategy. The</p>

CHARGE QUESTION 6: Are the outcomes measured appropriate given the study’s purpose?	
REVIEWER	COMMENT
	<p>report would be stronger if it provided citations separately for each measure used and an indication of how the construct it measures fits within a framework for message effects and/or the conceptualization of “understanding” (given the FDA mandate to increase public understanding). The primary outcomes of “new information” and “self-reported learning” have some face validity as potential indicators of understanding as knowledge accumulation. Still, the report would ideally cite studies with more convincing validity for these indicators. The other primary outcome, “thinking about risks,” has substantial predictive validity and relevant studies should be cited (e.g., many studies of adult smokers have shown that this response to warnings is associated with downstream cessation attempts). Some researchers consider this measure as indicative of message engagement or elaboration, which is a more standard term in communication research.</p> <p>The authors should provide some justification for selecting some indicators as primary and others as “secondary,” ideally based on the conceptual model that orients the study. For example, “informativeness” and “factuality” appear to overlap with the conceptualization of understanding. Why are they secondary? Why is credibility not primary, especially given that the messages are mostly about less well-known smoking-related outcomes? Decisions to treat these as secondary appears even more arbitrary after reading Study 2, where all measures are treated equally (i.e., no distinctions are made between primary and secondary measures).</p> <p>Phase 1 health belief questions use 5-point Likert type response options, which, in my opinion, does not really “fit” with the idea of a belief. In section 4.3, the report states: “Conceptually, the response categories for a Likert response scale represent an underlying belief continuum”. The report would ideally provide some justification for this approach to measuring beliefs (as opposed to attitudes, frequencies or other constructs for which a continuum makes sense conceptually and is more standard).</p> <p>The authors should include citations for each specific measure, as most are not standard (e.g., learning, new knowledge, informativeness, factuality).</p> <p>For phase 1, it is not clear to me if all participants answered the same belief questions or, as was implied in the executive summary, that this list included only the beliefs associated with the warnings that they evaluated. If the latter, it is not clear how alphas were calculated (due to incomplete data). If the former, provide an explanation for why overall beliefs were evaluated.</p> <p>Better justify how asking health beliefs in phase 2 ads meaningful information.</p> <p>Phase 2 health belief questions appears to be a series of 22 questions with check boxes. Better justify the creation of summative measures across all knowledge outcomes if we are interested in determining sensitivity to content that is included in warnings to which they are exposed.</p> <p>The belief questions may be more about memory and test taking skills than “understanding”.</p> <p>The authors should consider this as a potential limitation, especially since they show the stimuli multiple times and evaluate beliefs at two distinct moments.</p>

CHARGE QUESTION 6: Are the outcomes measured appropriate given the study’s purpose?	
REVIEWER	COMMENT
Reviewer #3	<p>Public understanding which is clearly the goal here. It is being equated to self-reported learning and the newness of the information. So, if a warning statement is identified as new and is perceived as teaching someone something they didn’t know before, then presumably this is a warning statement that is understood. This is an odd use of the word understanding which, in common parlance, is identified with comprehension and linkage with established knowledge. If I created an exam for students in my class and I asked them “is this information new to you and I asked them to report whether they learned something from the information” would I then conclude that they understood it? I think the answer is obviously no. Understanding is generally a concept that refers to the ability to use information successfully in one’s life and to integrate the information with an established pattern of beliefs which is already accepted. So, I for one would find it difficult to equate these operational procedures with the ordinary concept of understanding or with the cognitive concept of understanding as used in the scientific literature.</p> <p>It’s clear that the revised warning statements outperformed the original TCA statements handily on criteria that really do not tap into understanding, acceptance, or knowledge other than as measured by self-reported learning. And as I argued above, these two measures do not tap into understanding in any sense of what the word understanding ordinarily means conceptually, in ordinary discourse, or in scientific measures of comprehension. I am also not a fan of self-reported learning nor of novelty - that is awareness - as a criterion.</p> <p>For this reviewer, the primary outcomes seem a lot less interesting than the secondary outcomes seem to be. My argument is that the knowledge, learning, and thinking about kinds of questions are transparent and in some ways don’t really get at what their labels say they are getting at. For example, the abbreviated wording called new knowledge is really an awareness question. Learning is really reported learning, not recall or understanding. The belief items are about the extent to which people agree or disagree with a claim which is actually a rewording of one of the warnings. Later questions assess whether people accept the rewording as factual or not and as causal or not. All the questions prior to this set of items are not really about acceptance and while this study is supposedly not about persuasion in reality it’s crucially important for people to accept the warnings and not simply say they are new or say it leads them to say that they learned something.</p>
Reviewer #4	<p>The measures provided were seemingly drawn from extant assessments (as generally asserted in Section 2.3) but there are no links between these references and the actual measures used. Providing these links (as well as the rationale for such choices) is important for determining validity of the items and whether they were used appropriately in this study. For example, does improving knowledge lead to behavior changes? Providing a link to actual behavior is, of course, outside of the scope of the goals of this project. However, to the extent that knowledge is a proxy (or alternatively, mediates behavior change from warning exposure to behavior), it is important to consider and discuss.</p>

CHARGE QUESTION 6: Are the outcomes measured appropriate given the study’s purpose?	
REVIEWER	COMMENT
	<p>The single item assessments (e.g., informativeness; believability; fact vs. opinion) would necessarily be lacking in reliability (versus a longer scale). Disposition of these items and any validity evidence would help to underscore their appropriateness.</p> <p>Other measures could have been employed, even in the context of this internet panel. For example, memory of risks, true-false items, etc. The stated measures are likely adequate (especially if evidence can be provided on their psychometric soundness and use in previous studies) but other assessments could have been used to further meet the study goals.</p> <p>The decision to parse outcomes into primary and secondary was not clearly stated; or put another way, the reasons that some outcomes are considered primary versus secondary were not clearly stated. For example, given the stated goals of the study, “number of health conditions” seems more like a primary versus secondary outcome. Other secondary outcomes (believability, factuality, informativeness) seems tangential to the primary goal of improving “understanding of the risks”. Additional rationale is needed on these points.</p>
Reviewer #5	Overall, the outcomes are appropriate. However, I have a concern about the collinearity of two of the secondary outcomes. I believe that they may be highly correlated and could either be combined or one could be dropped if they are highly correlated. I did not have access to a correlation table, but that is my suspicion.
Reviewer #6	Yes. It would be helpful for there to be some theoretical expansion on some of the constructs measured. The rationale for the new knowledge, learning and health beliefs is clear, but I would have liked to have seen some more rationale for including thinking about the risks, believability, and factuality. Why are these particular measures important to the purpose?

CHARGE QUESTION 7: Are the study participants included appropriate given the study’s purpose?	
REVIEWER	COMMENT
Reviewer #1	Yes, the choice of the study population and efforts taken to recruit them are appropriate. This is not a representative sample of the population, but it doesn’t need to be. The goal is to test how a diverse population respond to study messages and the allocation between study groups is unbiased.
Reviewer #2	Yes: The focus on adolescents who smoke and are susceptible to smoke is standard for this kind of study, as is the inclusion of established adult smokers. However, it is not clear why ever-smokers who are susceptible were excluded (and only never smokers susceptible were included). Quotas for young adult smokers and older adult smokers is also appropriate given differential effects of warnings found for these populations and concerns about trying to influence young adults before they become too addicted.

CHARGE QUESTION 7: *Are the study participants included appropriate given the study’s purpose?*

REVIEWER	COMMENT
Reviewer #3	<p>One question about the weighting of the sample is how well the three age brackets reflect the distribution of smokers in the society. The selection process was one third under 18, one third 18 to 24, and one third 25 and older. The justification for this distribution is not clear but at a minimum should be compared to nationally representative samples.</p> <p>In the sample demographics, I was surprised to see some significant asymmetries in male-female distribution by adolescent and young adult groups. Females significantly outnumbered males among adolescents and the opposite was true of the young adult sample. It’s not clear why such sharp differences are present in the sub-samples or whether these differences might affect the results differentially for adolescents and young adults. I was also surprised to see in the sampling section that there was no discussion of how adolescents that were susceptible were defined. The definition is available later but should be there with the first introduction of the subgroup.</p>
Reviewer #4	<p>The sample employed for this study is one of convenience, which brings with it a host of potential biases and limits to generalizability versus employing a representative sample. The report describes limitations with this approach to sampling versus other options (Section 4.2) and describes cost as being a core decision-making factor in choosing convenience instead of representative sampling. There are differences in vendors in terms of sample quality – so one wonders about the reasons that Lightspeed was used versus some other vendor. Were there substantive reasons to select this vendor over others beyond cost and some of the data security features described (Section 2.2)?</p> <p>It would be useful to know what potential participants were told during recruitment. Were they given full details about the study or were smoking, warning labels, FDA, etc. mentioned to potential study participants during recruitment? Was a cover story used? This point is important because study framing could have an impact on those that agree and refuse; and also help to understand more about potential sampling biases (e.g., were those that agreed more likely to do so because they already held knowledge about smoking or needed knowledge?).</p> <p>One also wonders whether any methods (e.g., weighting) were or could have been employed to increase representativeness and generalizability.</p> <p>That said, the sample is diverse, so this offsets lack of representativeness and potential biases related to sampling (and match to population characteristics) somewhat. Still, lack of representativeness is probably the most serious limitation with this research.</p> <p>Is it possible to conduct analyses to test for differences between those that completed the study versus those that were eligible and quit or that did not start the study (i.e., by using the vendor’s existing data base)? Such analyses would further help to contextualize the results in terms of generalizability.</p> <p>There is some ambiguity for the criteria used to group participants into smoking or smoking risk groups. For example, what does “smoked on some days” translate into? In addition, there is significant heterogeneity in intermittent smoking (that includes not just days smoked, but amount smoked on each</p>

CHARGE QUESTION 7: <i>Are the study participants included appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
	occasion); some rationale is needed for these grouping decisions. Relatedly, the decision to exclude non-susceptible adolescents or committed never smokers is not clear; more rationale is needed on this point.
Reviewer #5	Yes, I think it was very appropriate to assess youth susceptible to smoking, youth smokers, and adult smokers. In other words, I agree with not including youth who are non-susceptible nonsmokers and adults who are nonsmokers.
Reviewer #6	Yes. You may wish to include a brief rationale for not including adult non-smokers, I believe this is a strength of the study, as these statements are designed to appear on packs and therefore the intended audience is smokers.

CHARGE QUESTION 8: <i>Is the analytic approach appropriate given the design and purpose of the study?</i>	
REVIEWER	COMMENT
Reviewer #1	This reviewer was impressed with the study analytic approach which was dictated by a very good experimental design.
Reviewer #2	<p>See above comments on measurement of outcomes that are relevant to analyses.</p> <p>The authors could create quasi-control groups with the treatment groups that are not shown a revised warning with the health outcome of interest. This would increase power for phase 1 measures, which would be particularly beneficial for evaluating statistical significance within the 3 key subgroups.</p> <p>As mentioned above, the authors would ideally provide citations and empirical justification for the anticipated effect size given the substantial body of research in this area. Contrasts like the one that I mention in the prior comment would allow for greater power.</p> <p>Table 4-2 shows alpha for a number of different subscales of beliefs. It is unusual to use alpha for assessing the internal consistency for two questions. The methods literature with which I am familiar indicates the need for a minimum of three questions for alpha to be meaningful.</p> <p>Throughout the section on hypothesis statements (starting at 4-11), the hypotheses are presented as directional (e.g., “...treatment condition > ...control condition”), but the statistical tests are indicated to be “two-sided”. Either make the hypotheses NOT EQUAL or change the language around the statistical tests to indicate “one-sided”.</p> <p>4-13 is where the analysis approach that involves randomly selecting a TCA statement to serve as the control for the revised statements without clear parallel content. This results in random selection of all responses to just one TCA statement for each comparison. Given the very limited pool of TCA statements, this approach to random selection risks integrating a systematic bias around consumer responses to the particular statement that is selected. For example, the TCA statement on “addiction”</p>

CHARGE QUESTION 8: <i>Is the analytic approach appropriate given the design and purpose of the study?</i>	
REVIEWER	COMMENT
	<p>addresses a concept that is notoriously difficult to communicate and is often evaluated as less effective than well-known disease outcomes. Nevertheless, it was randomly selected as the comparison for the statement about “macular degeneration”. A more robust comparison would involve a random selection from all TSA statements or even the grand mean of responses to all TSA statements, which would help iron out systematic idiosyncrasies around the specific topic that is randomly selected from a pool of only 9 possible TSA statements. This could be done when comparing treatment and control, as well as within person comparisons.</p> <p>4.3.3, Phase 1, part 2: Hypotheses and Analyses - The hypothesis phrase “average or level health belief score” is not clear, particularly the meaning of the term “level”.</p> <p>The analyses that involve evaluating whether the score was significantly higher than “not at all” (or 0 on a 0 to 7 scale) for learning, believability, and informativeness is unorthodox and should be better justified. I do not see this analysis as adding anything meaningful to what is already done with means and linear regression.</p> <p>Given the study design, provide a clearer justification for creating and evaluating a summative measure of smoking-related health consequences, as well as an overall health consequences measure. This justification should speak to issues around how we would expect beliefs to be higher for participants who are exposed to warning statements that address that specific belief (compared to participants who are not exposed to statements with that content). The summative measures used seem to reflect a broad conceptualization of risk perception (that goes beyond the content of the warnings) and should be justified. The more specific domains around secondhand smoke and pregnancy consequences do a better job of mapping onto specific warning content and therefore do not raise this issue.</p> <p>Appendix A indicates that a “control” belief was evaluated for each of the three domains of health consequences. To control for social desirability and acquiescence biases, it would be standard to include this as an adjustment variable in analyses that involve these beliefs. Was that done?</p> <p>[The Prior comments are from the methods report. Unless otherwise indicated, what follows is on the results report, although issues I raise above are pertinent to the background and methods sections of the results.]</p>

CHARGE QUESTION 8: <i>Is the analytic approach appropriate given the design and purpose of the study?</i>	
REVIEWER	COMMENT
Reviewer #3	<p>Table 3.4 offers pretty strong results for learning although clearly the results are primarily the result of adolescents and young adults in lesser so for older adults. Table 3.5 presents a strong evidence that new knowledge is enhanced for the new topical areas but no real advancement in terms of thinking about risks for these new topical domains except in a few cases (five to be precise).</p> <p>Table 3.6 makes my point that the new statements are often seen as less believable even though they are newer and quotes informative and elevated an awareness of new information. But simply put, new is not necessarily acceptable; new is often less believable and that’s borne out here. Similar findings are obtained with regard to facticity versus opinion.</p>
Reviewer #4	<p>The analyses follow a logical procedure and process; analyses are appropriate given the study questions (use of the powerful Benjamini-Hochberg procedure to account for multiple tests is a nice feature of the approach). Describing, in detail, each hypothesis, its link to a specific assessment, and the analytic approach used strongly underscores transparency of the work and approach (i.e., that no “short cuts” were employed).</p> <p>It wasn’t clear from the write-up whether there were baseline differences within each age group between conditions. There were no differences in the overall sample, as reported in 4.3.1, but it’s not clear whether that holds for specific age groups.</p>
Reviewer #5	<p>Although I am not a statistician, I believe that the statistics were appropriate. The authors conducted a large number of statistical tests and performed analyses to avoid making Type 1 errors. I am aware of Bonferroni corrections, but not the FDR method; however, it seemed appropriate from the description.</p>
Reviewer #6	<p>Yes. While the reasoning behind using ordinal and linear regression based on whether the items could be scaled is scientifically sound, it does make the interpretation of the results more difficult. The sample seems large enough to consider dichotomizing the responses as agree/disagree. If the pattern of findings holds, it would be easier to interpret, and would help the reader draw conclusions.</p>

CHARGE QUESTION 9: <i>Are results presented consistent with the analytic approach?</i>	
REVIEWER	COMMENT
Reviewer #1	<p>Yes, the results to all the study aims are thoughtfully and clearly presented.</p>
Reviewer #2	<p>Yes, the results are presented in a way that is consistent with the analytic approach.</p> <p>Tables use a $p < 0.05$ for unadjusted results. This is consistent with a one-tailed test, not a two-tailed test. As I mention my comment about the analytic approach, the wording of the hypothesis is directional but stating that two-tailed tests were used suggests non-directional hypotheses.</p>

CHARGE QUESTION 9: Are results presented consistent with the analytic approach?	
REVIEWER	COMMENT
	3.6 Phase 2 results - The description mentions respondents who “saw only revised statements,” but even these saw TCA statements in phase 1. After reading the results, I am still not clear how phase 2 results add anything meaningful. This should be clarified.
Reviewer #3	I appreciated the care with which the analytic plan was laid out and with the detailed attention to correcting for multiple comparisons and the presentation of both standard and corrected levels of statistical significance. Overall I think the analytic approach is not just solid but strong.
Reviewer #4	Yes, there are no concerns – the analyses are presented exactly as intended – no “fishing” expeditions were conducted.
Reviewer #5	Yes.
Reviewer #6	Yes. The results section provides a great deal of information on the individual statements and warnings, and it is consistent with the proposed analytic strategy. However, it would benefit from a few more summary figures or tables that help highlight the results of the slightly bigger picture question being asked. What do these results tell us about the statutory labels vs. the revised labels as a whole?

CHARGE QUESTION 10: Are there any concerns with the results presented?	
REVIEWER	COMMENT
Reviewer #1	Yes, some of the aims appear to be unrelated to the study’s purpose as currently written. The report tries to make the case for this as it presents the results. However, the danger is that this looks a little like post-hoc rationalization. There was no theoretical framework section outlining the relevance of each of these additional measures, so the reader has no knowledge of whether the hypothesis for each aim is met or not. Nor what a negative finding might mean to the overall purpose of the study. Indeed, it is not until the second study that the point of some of the aims becomes clear (some change with graphic warning labels but not with text warning labels). A combined theoretical framework is very much needed for these studies.
Reviewer #2	I am not clear why results are shown for tests that are not adjusted for multiple testing. The inclusion of these more problematic assessments does not seem to add any information of import and weakens the presentation by raising the question “Why have this information?”. For all tables that involve the comparison with a randomly selected TCA statement (e.g., Table 3-4), include a footnote or indication of the topic of the randomly selected warning in the table (if you do not choose to follow my recommendations above to do a different comparison).

CHARGE QUESTION 10: Are there any concerns with the results presented?	
REVIEWER	COMMENT
Reviewer #3	In Table 3.4 where the regression coefficients plus there are 95 percent confidence intervals are presented the authors use one of the kinds of presentations that drive this reviewer crazy. That is in presenting the confidence intervals they confuse dashes with minus signs whereas a simple modification could make it clear that something is a negative number versus something is a positive number by simply separating them with a comma; it’s a trivial thing but it makes for clarity in presentation or at least the absence of confusion.
Reviewer #4	Graphical presentation of key findings would improve readability (for those that prefer figures or graphs) and periodic summaries of the findings would assist in interpretation (lots of results are presented). It seems inappropriate to discuss non-significant findings, even to simply comment descriptively on mean differences. It is recommended that the results only comment on those results that are significant as defined in this report (i.e., with all corrections employed). The report asserts that the study was underpowered to conduct subgroup analyses (e.g., between smoking groups; ages). This limitation is unfortunate because some important conclusions about the widespread applicability and utility of the new statements could be made.
Reviewer #5	No.
Reviewer #6	No. If there is a way to summarize the results of Tables 3-6 and 3-7 in an additional table, that would be helpful. Consider making 3-8 into a figure or providing a figure that highlights some of the significant findings.

CHARGE QUESTION 11: Are potential limitations of the study appropriately identified?	
REVIEWER	COMMENT
Reviewer #1	Yes, the limitations section in this report is appropriate.
Reviewer #2	When raising the issue of the ecological validity of responses to online images of warning statements, the authors could cite studies that indicate convergent validity (see comment above). Concerns about generalizability could be addressed with studies showing that patterns of response in experiments are generally consistent across population sources. Also, as a sensitivity analysis, the authors could consider weighting observations so that they are more similar to the profile of smokers and nonsmokers in the general population. The limitations should do a better job of describing potential measurement error, citing the validity (or lack thereof) for the measurement approaches used. This could include considerations of content validity around approaches for measuring “understanding.” More broadly, there may be alternative

CHARGE QUESTION 11: Are potential limitations of the study appropriately identified?	
REVIEWER	COMMENT
	conceptualizations of “understanding” that would encompass embodied/experiential understanding. This kind of understanding may be stronger for smoking-related diseases associated with sensory perceptions from smoking (e.g., lung, throat, mouth, heart). Some evidence suggests that smokers perceive warning labels for these well-known outcomes as more effective than warnings for less-well known outcomes. Warnings may serve as reminders about this embodied understanding.
Reviewer #3	<p>I have no doubt that these new warnings will work versus the non-existent warnings we now have. But that said, acquisition of knowledge and recall of the new warnings is not the same as accepting the information as true. So while the purpose of the warnings is not to persuade, it is nevertheless to have people learn in the sense of accepting information, to understand in the sense of being able to use the information and to integrate it into a complex of information that is a part of people’s core understanding of the consequences of smoking combustible tobaccos. Simply being aware and saying that one has learned is not equivalent to having accepted the information. This is true in these data as well when the awareness levels (told me something new) are contrasted to the believability and facticity judgments.</p> <p>It’s very unfortunate that the allocation plan for the control group was not properly carried out. The reason of course is that the number of treatment conditions are so much greater than the control condition with equal allocation to all conditions without weighting. I had this problem in a study that I did a few years ago and rued the day when the treatment conditions were far out of proportion to the control condition. This could be a flaw in the study depending upon the kinds of analyses to be carried out in comparing some portions of the control group to various combinations of the treatment conditions.</p> <p>There is no commentary in the sample description regarding the potential confound of gender with adolescent and young adult samples. The problem this obviously creates is that comparisons between these groups -- if any (none so far) -- will be confounded with gender. I suspect some weighting will happen as needed but the asymmetries are pretty strong in the sample.</p>
Reviewer #4	Limits to sample representativeness are described in Section 4.2; other limitations are also described (e.g., endemic to online studies). Although the study was focused on assessing knowledge as an outcome, it seems important to state as a limitation that the results do not imply that these knowledge changes translate into less tobacco product use or initiation of use.
Reviewer #5	The study sample as a convenience sample is a limitation but that was clearly identified. The study also relied totally on self-report and that could have been highlighted a little more clearly.
Reviewer #6	Yes. Other potential limitations are in my comments in the methods section.

CHARGE QUESTION 12: <i>Are the conclusions drawn from the study well supported by the data presented?</i>	
REVIEWER	COMMENT
Reviewer #1	Yes, the conclusions outlined in the executive summary are well supported by the analyses and clearly relate to the study’s purpose as it is currently stated. However, there is a lack of consideration of the meaning of the additional aims that are not directly related to the current (limited) specification of the study purpose. See issues above on the need for a detailed theoretical framework for the study.
Reviewer #2	<p>The summary of findings should report on the significance of comparisons after adjustment for multiple tests. For example, I believe that the last sentence of the first paragraph in this section discusses 8 of 16 comparisons as higher for revised TCA statements, when I think it is only four after adjustment.</p> <p>The summary would ideally discuss patterns of findings across indicators, rather than treating them one at a time, which loses the broader patterns.</p> <p>The primary concern that I have is around the apparent prioritization of “primary outcomes” on “learning” and “new knowledge”. I am unfamiliar with prior research showing the validity and meaningfulness of the outcomes used to measure understanding (i.e., meeting FDA’s mandate). I am more familiar with indicators like the one used for “thinking about risks,” for which there is substantial evidence of predictive validity for cessation attempts across a variety of warning label policies and sociocultural contexts. Looking at the data for the revised statement on erectile dysfunction, for example, it generates more knowledge but lower thinking about risks and lower believability – which would recommend against its use. There are many other examples of inconsistent results, as well. These concerns about the validity of measures, primary vs secondary outcomes (voiced in the section above on outcomes), and the consistency of patterns across indicators of effect become particularly important when interpreting the results to inform Study 2. The documents would ideally be better linked, so that conclusions from Study 1 clearly inform the selection of stimuli for use in Study 2.</p>

CHARGE QUESTION 12: Are the conclusions drawn from the study well supported by the data presented?	
REVIEWER	COMMENT
Reviewer #3	<p>While it is certainly difficult and to some degree unfair to compare the revised warning statements to one another relative to the original TCA warning statements, at some point a decision has to be made as to which of these 15 should be prioritized if all are potentially eligible.</p> <p>The following statement is crucial in the summary because it makes clear that believability and facticity are called into question for information that is new and, as a consequence, the idea that people are becoming aware of a warning (i.e. a consequence) that they don’t actually accept as true undermines the quality of the results. Here’s the quote:</p> <p>“Though the revised statements were often considered to provide new information or improve understanding of the health effects of smoking compared to the TCA statements based on the primary outcomes, some statements were reported to be less believable or factual than TCA statements based on secondary outcomes. This pattern could be because a statement that provides new information that the respondent has not heard before might be viewed with some skepticism.”</p> <p>The report suggests that even though believability and facticity of the revised statements may be called into question in some cases the results are desirable or favorable because beliefs for the revised statement exposure were elevated as reported causes of negative consequences. But these negative consequences were themselves restatements of the warnings to which folks were exposed.</p>
Reviewer #4	The conclusions section is a summary of the results; it accurately represents those results. To the extent that broader conclusions need to be drawn (e.g., about specific statements and their suitability), those statements are absent.
Reviewer #5	Yes, the conclusions generally flow well from the data presented. However, the conclusions are very conservative. The report does not really ‘take a stand’ on which warnings might be the best according to the data. I prefer findings that are a little more prescriptive based on the evidence and that synthesizes rather than merely regurgitates the findings. That said, I do not think the report goes well beyond the data.
Reviewer #6	To the extent that there are conclusions presented, yes. The study presents a summary of findings, which is a better descriptor than conclusions. The report does not conclude which statements are the best, or which of the results presented are used to determine the utility of the statement for selection. Table 4-1 summarizes the findings nicely, but the summary does not provide any real “take-home” points in any easily digestible fashion. This section could be improved if it not only summarized the findings but helped place them in the context of the purpose of the findings.

Specific observations			
REVIEWER	PAGE	PARAGRAPH/ LINE	COMMENT
Reviewer #1			None provided.
Reviewer #2			None provided.
Reviewer #3			None provided.
Reviewer #4			None provided.
Reviewer #5	4	Whole page	The design is a little confusing and the verbal description would be improved by a study design figure similar to the one in Study 2 (Figure 2-1, p.15) that showed the 3 sessions and the outcomes. Please add a study design figure that shows the two phases and the outcomes and stimuli by condition.
Reviewer #5	17	Tables 2-1, 2-2	The ‘matching’ that was done between original TCA labels and the revised ones is quite confusing. I would consider making a new table in landscape that shows the original statement and their matches. I know that some don’t have a match and that could be denoted. This new table of matches could replace these tables or be an additional one. I think replacing might be better. This would also make it clearer that conditions 12-16 don’t have a true match.
Reviewer #5	18	Para “The Phase 2...”	This paragraph is dense and takes 2-3 readings to grasp. A picture in the form of that revised study design or table would help.
Reviewer #5	20	Middle	(added word) However, because respondents were recruited using non-probability, convenience sampling methods, results from this study are not <u>necessarily</u> representative of the populations from which the sample was drawn.
Reviewer #5	24	Middle	The final secondary outcome looks at secondhand smoke. I do not think the report pointed out that none of the labels focused on secondhand smoke, and I would point that out. It would be a type of halo effect to impact that outcome.
Reviewer #5	27	B8_1 to 4	If not mentioned, I would make it clear that those items were all mentioned together in a warning, but the specific elements are split apart for measurement.
Reviewer #5	30	Middle	Can you state the actual amount of statistical power for the Tx vs Control comparison rather than say it was lower?
Reviewer #5	32	Para 1	I believe these are analyses 12-16 – I would say that to aid the reader.

Specific observations														
REVIEWER	PAGE	PARAGRAPH/ LINE	COMMENT											
Reviewer #5	36	Section 1	For dichotomous outcomes, I think the last one should be Ha and not Ho.											
Reviewer #5	37	Top of page	<p>Same as above.</p> <p>For dichotomous outcomes (i.e., new knowledge, thinking about risks, factuality):–</p> <p>H0: the proportion (%) responding in a manner indicative of being better informed about the health risks of smoking (e.g., reporting that the statement provided new knowledge) for those in the treatment group = 0.</p> <p>–H0: the proportion (%) responding in a manner indicative of being better informed about the health risks of smoking (e.g., reporting that the statement provided new knowledge) for those in the treatment group > 0.</p>											
Reviewer #5	60		It might be easier for the reader to show the Phase 1 methods and results and then describe Phase 2 vs the method 1 & 2 and then results 1 & 2.											
Reviewer #5	63	Sect 3.2	I don’t recall seeing the S stood for Statutory so I might say that earlier. Maybe I missed it.											
Reviewer #5	68	Comparison2	<p>The pattern of findings seemed odd and I might recheck them. The two warnings had very different % of new knowledge yet had identical rates of Thinking about Risks.</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td rowspan="2" style="vertical-align: middle;">2</td> <td>Unspecified cancer (S4)</td> <td style="text-align: center;">12.2</td> <td style="text-align: center;">REF</td> <td style="border-left: 1px solid black; text-align: center;">68.9</td> <td style="text-align: center;">REF</td> </tr> <tr> <td>Head and neck cancer (R1B)</td> <td style="text-align: center;">64.2</td> <td style="text-align: center;">13.26 (7.20 - 24.4)^{a,b}</td> <td style="border-left: 1px solid black; text-align: center;">68.9</td> <td style="text-align: center;">1.00 (0.61 - 1.64)</td> </tr> </table>	2	Unspecified cancer (S4)	12.2	REF	68.9	REF	Head and neck cancer (R1B)	64.2	13.26 (7.20 - 24.4) ^{a,b}	68.9	1.00 (0.61 - 1.64)
2	Unspecified cancer (S4)	12.2	REF		68.9	REF								
	Head and neck cancer (R1B)	64.2	13.26 (7.20 - 24.4) ^{a,b}	68.9	1.00 (0.61 - 1.64)									
Reviewer #5	73	Sect 3.3.6	Believability and Factuality seem like the same construct. I also think the pattern of findings may be identical. If they are highly correlated, I would drop Factuality.											
Reviewer #6			None provided.											

STUDY 2

General Impressions	
REVIEWER	COMMENT
Reviewer #1	<p>There is a lot of overlap in the methods section for Study 2 with that of Study 1. I suggest that there be a re-positioning of these two studies so that you can combine the two methodology sections into a single presentation.</p> <p>The two studies could also be combined as they are somewhat hierarchical. The first study investigates the response to new and improved text messages. The second study investigates how graphic warning labels can enhance the response to these text messages. A very important conclusion from these two studies is that text messaging alone does not achieve changes in health beliefs, but, when combined with graphic imaging, health beliefs are influenced. It is not possible to do this when the material is presented as two separate, and apparently independent, studies.</p> <p>This report needs a much better section on the theory of health communications that incorporates why changing health beliefs is important and a step above the communication-persuasion achieved with the text only communication. I suggest a version of McGuire’s communication-persuasion matrix that incorporates emotive processing in the behavior change model. The linear nature of this model is too simplistic, however, there is a hierarchy that is important.</p> <p>For this application, exposure is controlled by the excellent experimental design that enables identification of associations with the different processes of communication-persuasion. First of all, the participant needs to identify that the message has new information and that it is understandable. When this new, understandable information is considered factual, it can lead to new cognitions (thinking about risks). However, it is important to behavior change that this new knowledge leads to a modification of health beliefs that are retrievable at the time of performance of the addictive behavior. When the exposure generates an emotive response associated with the risks of use, then it is likely that there will be greater change in health beliefs. The goal of putting graphic warning labels on cigarette packaging is that the image will continue to generate an emotive response which will be a cue to retrieve these health beliefs every time the person reaches for a cigarette.</p> <p>Both studies address how messages can assist people to make changes to their behavior and I would incorporate them into two sections of the same report.</p>
Reviewer #2	<p>I have provided my comments below, emphasizing those that are most important. A few stand out: (1) the need for some overarching conceptual framework to bring coherence to the outcomes assessed and interpretation of results; (2) stronger justification for the warnings selected, ideally based on Study 1 results and prior research; (3) stronger justification for the measures used, including information on the validity of measures, especially novel ones; (4) consideration of prior research to determine meaningful effect sizes and power; (5) stronger justification for the analytic approaches (e.g., combining four SG warning control groups for comparison instead of creating more comparable comparison groups) and</p>

General Impressions	
REVIEWER	COMMENT
	inclusion of sensitivity analyses to evaluate the consistency of results under different specifications (e.g., population weights; adjustment for variables that account for differential attrition).
Reviewer #3	<p>The comments that I have made in response to the 12 charge questions include various elements that would fall under “general impressions.” I draw them out here in a separate answer.</p> <p>Both studies are very well done in terms of design and data analysis. The designs selected provide for an appropriate control groups which are the current standards for warnings (Surgeon General in Study 2) or immediate past selections (initial FDA warning labels in Study 1). The data analysis plan is strong and straight forward. It is careful on statistical treatment of the quality of data (e.g. continuous vs. rank order) and especially strong on correcting for experiment-wise error and power considerations. Both documents do a good job in communicating their procedures and the results except as noted regarding Study 1’s complex and difficult design that requires some supplementation with a clear and effective graphical description. If one accepts that the operational measures of learning and novelty are valid measures of their underlying constructs, then there is a clear picture that the new warning texts and warning labels are effective relative to their comparison.</p> <p>However, I am concerned that the measures deployed – perceived novelty and awareness – are not convincing measures of the underlying constructs that the research is targeting. In Study 1, the researchers do employ a measure of believability and of facticity (opinion vs. fact) finding that the new labels are less believable. In Study 2 the believability measure is not present even though it was diagnostic in Study 1. Both the believability and facticity measures underscore the fact that the new warnings may not be accepted by the target audience. The authors are well aware of this and comment on it in both studies. But coupled with self-report measures of learning and novelty (awareness), the lower levels of acceptance of the labels reduce the overall impact of the results. The implicit rejoinder in the data to the argument that the results are not so convincing is that the warnings affect the acceptance of negative health consequences (i.e. beliefs) in the warning label conditions versus the SG warnings condition and do so over time. But as I note in my comments on Study 2, two concerns arise about these findings. The first is that asking beliefs at baseline before message exposure taints the message processing by focusing respondents’ attention to the messaging in ways that privilege the beliefs being targeted. Second, the beliefs are measured three times reinforcing the warning labels’ content. Third, it is not clear why all beliefs would be affected by a specific warning label as opposed to a more targeted set of outcomes wherein warning label X affects beliefs related to warning label X but not beliefs Y and Z.</p> <p>In the end, these are both very carefully done studies that adhere closely to the data that has been gathered. This reviewer is raising interpretive considerations that essentially claim that the overall set of results are less convincing than they might be had the same constructs been operationalized differently and slight changes in the design in Study 2 been implemented.</p>
Reviewer #4	This very clearly written report describes a study that compared new graphic health warning labels for cigarettes to those that are already in circulation in the U.S. on various measures of health education. The methods and results are very clearly described in a way that lends the study to great transparency in

General Impressions	
REVIEWER	COMMENT
	its approach. The conclusions and reporting of results are accurate. The randomized design is well described and appropriately conducted to ensure internal validity (especially important with an internet panel experiment). Concerns that detracted from the overall quality of the study and report include the following. First, the study needs greater levels of conceptual and empirical motivation – even as a practical matter, it is important to understand the reasons that graphic health warnings are important and why improving knowledge of the health risks of smoking could lead to changes in behavior. Second, the measures, while clearly described, are not linked to any specific study or research program, making it difficult to evaluate their utility and validity (beyond face validity). Third, the sampling scheme is a major limitation as the convenience sample limits generalizability of the study findings and the attrition rate over the three study sessions (over almost three weeks) is very problematic. Finally, the stimuli are necessarily artificial in experiments such as this one, but some design decisions (use of a blue package; model in the mocked-up design) could limit the overall generalizability of the study findings (i.e., for non-blue packages; and ads that do not feature a musical theme or male model).
Reviewer #5	None provided.
Reviewer #6	<p>This study was well designed to assess whether the revised warning statements would meet the mandate to increase understanding of the risks associated with smoking. It is clear why the main outcomes of new knowledge and learning were included and prioritized because of the statutory obligation in the TCA. However, there is little discussion of theory in the selection of the other measures that have been included and analyzed. For example, the cognitive elaboration measure (“think about risks”) and the potential importance it may have in retention of information included on the labels, is not included in the report.</p> <p>The longitudinal nature of the study, as well as a naturalistic placement of the messages on packs and in advertisements make these data compelling, as does the use of the SGW as the control condition.</p> <p>In the summary section, Table 4-1 was particularly helpful in bringing a large amount of information together in an easily digestible way, but the summary section does not help the reader synthesize the findings or leave the reader with an overall take-home message.</p>

CHARGE QUESTION 1: <i>Is the document logical and clear?</i>	
REVIEWER	COMMENT
Reviewer #1	To this reviewer the weakness in the report is the lack of an explicit theory that addresses why the different measures are undertaken and what each construct is expected to achieve in terms of the final outcome. Is this behavior change, or is it just the first step towards this – the identification of new knowledge?

CHARGE QUESTION 1: <i>Is the document logical and clear?</i>	
REVIEWER	COMMENT
Reviewer #2	Yes. However, there is significant redundancy between the methods and results reports that would ideally be deleted.
Reviewer #3	Yes, the document is logical and clear. The study design is easy to understand especially in contrast to Study 1 which is so much more difficult to figure out. The description of the design, the measures, the analysis techniques as well as the results are readily interpretable and readily comprehensible.
Reviewer #4	The document is exceedingly clear in its purpose and in its style of communicating. It follows a logical course (with some redundancy between sections adding to the readability of this dense, comprehensive report). Level of detail, particularly as it relates to the analytic plan and results, are crystal clear – supporting transparency of this endeavor.
Reviewer #5	Overall, the document is logically organized and is pretty clear. I have one comment on the general organization of the document. As with Study 1, I think it is inefficient and lengthier to have 3 cover pages for Executive Summary, Methods, and Results rather than having them all part of one document with a clear Table of Contents.
Reviewer #6	Yes. There is some redundancy between sections, but it does allow for each section to be read and digested independently. Many of the outcome measures are not justified in the background or methods section.

CHARGE QUESTION 2: <i>Does the executive summary accurately reflect the content of the overall document?</i>	
REVIEWER	COMMENT
Reviewer #1	The executive summary presents the findings clearly, however, it lacks the theoretical model that will help the reader understand the importance of the findings to the overall goal. Is it just new knowledge or new knowledge that will assist the smoker in the behavior change process? This reviewer thinks that the second study is built on the first study and that they should be presented as one. The second study shows how graphic images lead to more advanced processing than text only messages that results in a change in health beliefs. Thus, these warnings will be more associated with increased probability of behavior change. At the present, there is no attempt to address why graphic warning labels are a step-above the textual warnings, although the findings can directly address this.
Reviewer #2	Yes. However, it would be useful to indicate which of the 15 warning statements are revised and which are original statements from the TCA. Also, I would put the actual stimuli in the table that shows the warnings. Rather than listing the outcomes, the summary would be more compelling if these outcomes were somehow linked to the overarching conceptualization of “understanding” or to established frameworks on message effects. My comments about specific outcome measures from Study 1 apply to this Study 2,

CHARGE QUESTION 2: Does the executive summary accurately reflect the content of the overall document?	
REVIEWER	COMMENT
	as well, since many of the same outcomes are used. There are also some new outcomes assessed in Study 2 that should be justified (e.g., understandability, helpfulness) by linking to this conceptual framework. For the executive summary, this could be done briefly, with more detailed description and justification in the background and measurement sections.
Reviewer #3	Yes, I thought the executive summary was very good and clearly highlighted the overall findings as well as giving a sense of how the data were gathered and the empirical procedures carried out. So, from the point of view of communication and presentation of findings I think the study too does a very good job. In spite of its length, it really does a nice job of presentation and communication.
Reviewer #4	The detailed executive summary accurately reflects the content of the full report.
Reviewer #5	The Summary does a nice job of reflecting the overall content of the full report.
Reviewer #6	Yes. The summary gives a succinct overview of the study purpose, methods, and results. It is clear and should be easy to follow for interested members of the public.

CHARGE QUESTION 3: Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?	
REVIEWER	COMMENT
Reviewer #1	The study design is enlightened and allows for a rigorous testing of the multiple applications of warning labels. The stimuli are well chosen. The sample is an appropriate size. Allocation to the respective groups is by a minimization method first introduced in the biostatistical literature in the 1970s and when grouped with quotas for key sampling units ensures that there is comparability across study groups with appropriate representation of the key population components. This allows this research to be undertaken using the online panels. The downside is that it is not a randomization procedure, however, the assumptions of the statistical testing are robust enough that this procedure does not introduce significant bias. The analyses are appropriate and the results meaningful.
Reviewer #2	The two studies should be better linked, so that results from Study 1 clearly inform the selection of stimuli for use in Study 2. This is not currently done, and there is no justification for the selection of stimuli for Study 2.

CHARGE QUESTION 3: <i>Is sufficient information provided about the study design, stimuli, sample, methods, analysis, and results?</i>	
REVIEWER	COMMENT
Reviewer #3	<p>The level of information provided is thorough in every segment of the work and so I have only minor quibbles about some additional pieces of information that might be provided. The stimulus materials that were presented were clear enough, but it was not clear how they were derived especially in the context of Study 1. Study 1 aimed to develop some new warning texts and the connection between Study 1 and Study 2 should have been obvious in some ways but was not.</p> <p>Table 3-6 refers to completed surveys but needs to be more forthcoming by describing if this means completing all three sessions and what happens to those dropping out after the first session etc. Attrition is an issue and should be addressed in the sample completion section (somewhere for sure but this would be a good place).</p>
Reviewer #4	<p>The design and analyses all are very clearly described. The sample is clearly described, but some additional details would help to overcome structural limitations with the sampling frame (i.e., a convenience sample is used vs. a population-representative sample). The specific procedures that went into development of the new text warnings and images were not described in detail; more information on formative work would improve the presentation of the report. Additional information that links measures to their origin in the literature and provides some evidence of validity is needed. Results would be improved by providing graphical representation of the results. Overall, the study requires more conceptual motivation other than to build upon the practical motivation of testing labels as required by statute. All of these issues are described in detail below.</p>
Reviewer #5	<p>Overall, the document provides most of the key details. However, there are some additional explanations or ancillary analyses that I would recommend. These are listed below and in the table at the end of this document.</p>
Reviewer #6	<p>Yes. It would be helpful to provide a brief summary of the interesting and relevant findings from the subgroup analyses that are included in the appendices embedded in the results section of the report.</p>

CHARGE QUESTION 4: <i>Is the methodology used appropriate to address the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #1	<p>Yes. The methodology demonstrates careful consideration of the design principles adapted for use with online panels. The only thing that is lacking is a justification for the choice of the two-week follow-up. This is an easy addition as this time-point is a trade-off between the need to go beyond short-term memory recall while keeping the timing close-enough to minimize loss to follow-up.</p>
Reviewer #2	<p>The methods are generally appropriate.</p>

CHARGE QUESTION 4: <i>Is the methodology used appropriate to address the study’s purpose?</i>	
REVIEWER	COMMENT
	<p>There should be a stated rationale for the session 2 exposure, which I assume is to simulate naturalistic, repeated exposure to warnings that would happen in standard implementation periods.</p> <p>Power calculations that adjust for false discovery rates (Benjamini & Hochberg, 1995) are important given the great number of comparisons made. However, the authors would ideally provide citations and empirical justification for the anticipated effect sizes given the substantial body of research comparing pictorial and text only warnings. Otherwise, it is hard to determine if the study is over or underpowered. This information will also be useful when considering the recommendation to conduct additional analyses that compare specific SG warnings with comparable GHWs, as well as power for evaluating subgroup analyses.</p>
Reviewer #3	<p>Since the design is longitudinal over three points in time there clearly will be attrition and indeed there was. Some careful discussion of the process of participant loss is called for and I make some suggestions about what to take a look at here.</p> <p>Lightspeed’s quality control mechanisms seem strong to me. As a researcher who uses online samples, I would be interested in using a panel with such quality control.</p> <p>The items in the surveys are described in Table 4-1 and seem to imply that the set of 16 belief items were asked three times. This is fine except that asking these items at baseline distorts the way people process the information given in the labels cuing them into the content to be processed. In our message work we never ask the key outcome measures at baseline BEFORE the messages to be processed as we believe that that distorts how content is handled – priming, focusing, differential attention, etc.</p> <p>Every design can be criticized for failing to do something. In the current design my greatest concern is that the belief items are asked three times. I’m not primarily concerned about test-retest sensitization which occurs in both the control and the treatment conditions but rather the interaction effect between the belief assessments and the follow-up warning labels. The problem is that the belief statements have content which is consistent with and primes the respondents to focus on the warning label in a unique way. This creates the possibility -- indeed the likelihood-- that respondents are reacting to the content of the warning label in ways that they would not in the absence of pretest measures of negative health consequences. So, this can confound the results in ways that are different in the experimental context than they would be in the real world context where the beliefs are not primed systematically prior to exposure to the stimulus materials. This of course could have been designed out at substantial additional cost in terms of resources by having a post only condition to compare to the pre-post-conditions of the current design. But as I said every design can be criticized for failing to do something and jeopardizing internal validity.</p>
Reviewer #4	<p>The randomized design is straightforward and enables a mostly internally valid test of the study hypotheses (about the primacy of graphic health warnings over text-only warnings).</p>
Reviewer #5	<p>The methods of an online panel to test consumer reactions to new warning labels is generally appropriate. An in-person study with a sample this large and multiple follow-ups would be very cost and time prohibitive. I like that the design of this study involved two viewings of the stimuli. This mimics the</p>

CHARGE QUESTION 4: <i>Is the methodology used appropriate to address the study’s purpose?</i>	
REVIEWER	COMMENT
	real world exposure whereby people will see the warnings on their pack each time they pull it out to smoke. As such, showing the warnings twice are a likely underestimate of the impact of these warnings. In other words, the fact that this study detected significant effects bodes well for their real world impact given the more frequent exposure that will occur.
Reviewer #6	Yes. The naturalistic approach and longitudinal nature of the study are assets of the study and help address the study’s purpose. Comparing the proposed warnings to the current warnings is appropriate to the study’s purpose.

CHARGE QUESTION 5: <i>Are the stimuli used appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #1	Yes. Each graphic image was appropriately matched with a text message used in the first study. In the instances in which there was not an appropriate match, the investigators used random assignment. This is an optimal approach.
Reviewer #2	Stimuli seem appropriate, but the justification for selecting warning statements and associated pictorial imagery should be spelled out. Ideally, this would build on Study 1 findings and/or the scientific literature. When describing the stimuli in the text, the report should describe differences in the size and placement of the control vs. treatment statements.
Reviewer #3	What led to the choices of the 16 given the results of the prior study? Why not stay with the original set? Why drop back to some of the previous warnings in the tested set? How do the texts developed from Study 1 play into the selections for Study 2? What did I miss? Exposure to stimuli is not masked in any way. Respondents are exposed at time 1 and at time 2. Health beliefs are assessed at three points in time including prior to exposure at time 1. Asking all the beliefs together three separate times could have the effect of creating a clustering of beliefs such that any effects on one from a warning would transfer to the others even though that would be an unnatural result of the design and the set of items.
Reviewer #4	From a practical perspective within the context of the study goals, the stimuli are appropriate. The present study is linked to the former study (OMB# 0910-0848) but it does not appear as though the results of that former study were used to inform the stimuli choice in the present study. In other words, the most effective stimuli (text warnings) from the former study were not used in the current study; all of the text warnings were used. This is not a limitation or criticism per se, but it does suggest that the relationship between studies needs to be more clearly explained.

CHARGE QUESTION 5: <i>Are the stimuli used appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
	<p>In addition, the development process for the graphic images needs to be more clearly described. What was the formative testing process that went into designing and then finally choosing those images? What was the charge given to the designer that developed those images? Were images keyed to the specific text warnings? Were issues regarding diversity of the model facsimiles part of the development (e.g., regarding age, gender, apparent race or ethnicity)?</p> <p>There is an element of artificiality in this kind of experimental study, of course, but the decision to use a cigarette named “brand” versus a name brand was not clear. There also could be concerns about use of a blue box for this cigarette brand stimuli as different colors communicate different features of the product. For example, some older research by the tobacco industry suggests that blue could convey information that the product is “safer” or more geared toward males (again the research is old). While there were not differences between conditions in the package color, the findings may be different with differently colored or logoed packages (or packages for real brands). Some consideration of this issue is warranted.</p> <p>Decision processes and development of the print ad is also needed (e.g., why music? why a male model?).</p>
Reviewer #5	<p>The stimuli used were appropriate for an online study. The images were presented on both a cigarette pack and on a mock print advertisement, which replicates how they will appear in the real world. Having a 3D-style pack that rotates is appropriate and desirable. However, the Appendix did not properly depict what the pack presented to study participant actually looks like. Appendix 1 shows a flattened illustration for a printing company. That is fine to show, but at least one example of a control image and one of a treatment pack should be depicted in the Appendix. I know that this image will not be movable given this is a report but show the image from 3-4 angles to show readers what the images looked like in the study.</p>
Reviewer #6	<p>Yes. You could consider giving some details about the development of the photorealistic images and how they were determined as concordant with the messages.</p>

CHARGE QUESTION 6: <i>Are the outcomes measured appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
Reviewer #1	<p>The problem with the outcome measures is that there is no presentation of a theory on why they are important. Of particular concern is the relevance of health beliefs. This is the key difference between the text only vs. graphic warning labels. See above.</p>
Reviewer #2	<p>As I commented for Study 1, the conceptual model(s) that orients measurement for Study 2 is never clearly defined. At the end of a single sentence in section 2.3 Instrument Development, the authors cite a bunch of studies to support their measurement strategy. The report would be stronger if they provided citations separately for each measure used and an indication of how it fits under a framework for message effects and/or the conceptualization of “understanding” (given the FDA mandate to increase</p>

CHARGE QUESTION 6: <i>Are the outcomes measured appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
	<p>public understanding). My comments around outcomes that are also used in Study 1 apply to Study 2, so I will not repeat them. However, there are also some new measures for Study 2 that I have not seen before and that also would benefit from some information about their validity and prior use. For example, is “Understandability” the same thing as “clarity”? If so, that construct is relatively common in studies of perceived effectiveness and could be cited as such. The new question on attention grabbing is commonly used and should be cited.</p> <p>The authors should include citations for each specific measure, as most are not standard (e.g., learning, new knowledge, informativeness, factuality). This will help with interpretation of results.</p> <p>It is not clear how the measurement of B1 (Before today, had you heard about the specific smoking-related health effect described in the warning?) was done for warnings where multiple health effects are mentioned in the warning statements. Were participants asked to interpret the entire gestalt of the warning statement, including if it mentioned multiple health outcomes (e.g., SG warning on “lung cancer, heart disease, emphysema and may complicate pregnancy”)? If so, there is a lack of “fit” between some warning statements and question B1. There is also a lack of fit when the warning does not address a specific health effect (i.e., “Quitting smoking now greatly reduces serious risks to your health;” “Cigarette smoke contains carbon monoxide”). These issues should be addressed in the limitations.</p> <p>Measurement of recall is okay, but the authors would ideally describe the specific type of assessment (which some call “recognition”) and, in the limitation section, potential biases associated with this type of recall vs. other types (e.g., confirmed recall, cued recall). For a good discussion of the strengths and weaknesses of different approaches to self-reported recall, see Niederdeppe J. Conceptual, empirical, and practical issues in developing valid measures of media campaign exposure. <i>Communication Methods and Measures</i>, 8(2), 138-161. 2014.</p>

CHARGE QUESTION 6: Are the outcomes measured appropriate given the study’s purpose?	
REVIEWER	COMMENT
Reviewer #3	The believability criterion was not included in this study and that is problematic I think because these results undermined the legitimacy and utility of the warnings. Facticity was a problem with the new warnings in Study 1 and they are a problem here as well. The bottom line in this and in the prior study is that the new information is not as well accepted as the old information and so as a warning to smokers and potential smokers it will be less effective than a warning based on established claims. Of course, an established claim of negative health consequence is “old hat” and one cannot show improvements in learning as it is already overlearned. But new warnings will require support or will fall by the wayside in terms of their acceptability. The counterargument that this new warnings standard will be better than the existing warnings is without question going to be true but whether these new labels would be as effective as some established warnings that are already accepted is not tested and is a legitimate counter hypothesis. Another counterargument from the data of this and the prior study is that beliefs are affected by the new labels and so are actually pretty effective? But the comparison establishing this claim is a weak set of existing SG warnings and the beliefs shown to be affected are ones that are variations on the wording in the new warnings already (and repeated multiple times). So, the counterargument is a weak one I think. Testing should have included some false beliefs or beliefs not a part of the warning set to show that the effects of the new warning labels are on the targeted beliefs and not a general halo on any and all smoking related beliefs.
Reviewer #4	The outcomes are geared toward the goal of the study, but more information on the origin of the measures and validity needs to be considered. There are general references provided in the aggregate for the measures but linking specific measures to a specific citation is important for determining the appropriateness of that measure.
Reviewer #5	Yes, given the outcomes listed in the original legislation, the outcomes chosen are appropriate. I also appreciate that labels with multiple warning elements were evaluated with scales rather than a single double-barreled or triple-barreled question. The study goes beyond just one or two measures of the novelty of this warning information, but also probes the credibility, understandability of the images, as well as their ability to grab viewers’ attention. This last factor is a key element of many health communication theories.
Reviewer #6	Yes. It would be helpful for there to be some theoretical expansion on some of the constructs measured. The rationale for the new knowledge, learning and health beliefs is clear, but I would have liked to have seen some more rationale for including thinking about the risks, believability, and factuality.

CHARGE QUESTION 7: Are the study participants included appropriate given the study’s purpose?	
REVIEWER	COMMENT
Reviewer #1	Yes. This study population represents a good trade off. While it is not a representative sample of the population, it is diverse and easily recruitable. As the study uses an unbiased allocation procedure across study groups, the study participants are appropriate to address the research questions.
Reviewer #2	Generally, yes. The focus on adolescents who smoke and are susceptible to smoke is standard for this kind of study, as is the inclusion of established adult smokers. However, it is not clear why ever-smokers who are susceptible were excluded (and only never smokers susceptible were included). Also, it is not clear why adult nonsmokers were included. The report should justify these decisions.
Reviewer #3	<p>Sample selection. The adolescent group included smokers and susceptibles which makes sense given that the transition to smoking occurs early and not later in life. But why only use smokers and not-susceptibles in young adults where the transition still is occurring. Makes sense not to include susceptibles among adults but what is the justification for non-smokers in adults (former smokers yes but never smokers)? What will this population tell us about the effectiveness of warnings other than the political acceptability of the warnings?</p> <p>The sample description in Table 2.2 says that adult non-smokers currently “do not smoke at all” but does this mean that they could be former smokers? Same for young adults. This is definitely clarified later on in the document, but it would be reasonable to make sure that it’s clear at the outset.</p> <p>In the sample’s demographics, the adult sample has a significant asymmetry in age distribution with 35-55 underrepresented.</p>
Reviewer #4	<p>The sampling frame is a significant weakness for this research. As noted in the report, convenience samples suffer from a host of structural biases that cannot be easily overcome (unless there is some sort of weighting procedure considered; but was not apparently done for this study). Moreover, the loss to follow-up over the short time frame of the study creates additional problems. The report advances some discussion points to minimize the attrition problems; but these are significant problems, nonetheless. Some analysis of drop-outs versus completers, sample representativeness, etc. would help to offset some of these concerns about the sample.</p> <p>Some information on recruitment methods are needed. For example, what were participants told about the research? Were participants kept blind to the study purpose up to a point? These issues are important to consider because each has implications for contextualizing the limitations to generalizability of the sample.</p>
Reviewer #5	The inclusion of youth who use tobacco and who are susceptible is appropriate as were the selection of adult smokers. However, I am puzzled why non-smoking adults were studied. Why would you want to ask a 60- or 70-year-old non-smoker what they think about these warning labels? There is almost no chance that they will become a smoker. I encourage a revised report to either state why they were included or to remove them from analyses.
Reviewer #6	Yes. I would clarify whether the non-smoker category includes never smokers and former smokers, or only former smokers. Since you are including non-smokers in this study (as opposed to Study 1), you

CHARGE QUESTION 7: <i>Are the study participants included appropriate given the study’s purpose?</i>	
REVIEWER	COMMENT
	might provide a brief rationale for the inclusion on non-smokers as a group in this study and any hypothesized differences.

CHARGE QUESTION 8: <i>Is the analytic approach appropriate given the design and purpose of the study?</i>	
REVIEWER	COMMENT
Reviewer #1	Yes. The analytic approach is very well considered if you assume the theoretical framework for the study. However, this theoretical framework is not presented appropriately and needs further explication. However, this can be done in a way that fits with the analytic approach and allows each of the study hypotheses to be addressed.
Reviewer #2	<p>My primary concern around the analysis is the decision to combine into a single control group the four distinct groups that were exposed to each of the four different SG warnings. This is a reasonable approach if the concern is to mitigate social desirability and testing effects, but I think it would be stronger to also test differences between specific SG warnings and GHWs that are most comparable. In other words, compare responses to the SG warning on health effects (i.e., ...lung cancer, heart disease, emphysema and may complicate pregnancy) with the GHWs on health effects (perhaps noting which comparisons include the health effects mentioned in both warnings and which do not). The SG warning on quitting can be compared with the GHW on quitting. The SG warning on pregnancy could be compared with the GHW on pregnancy. The SG warning on carbon monoxide is not comparable with any GHWs. At the very least, these more focused analyses could be treated as sensitivity analyses. This focused approach may be most meaningful for the evaluation of measure of “new information” and changes in health beliefs, but I think it would strengthen all analyses.</p> <p>Power appears reasonable, but raises questions given that the “control” group is actually four distinct groups exposed to four distinct messages. Ideally, power would address comparisons where a specific SG warning is compared with the warnings with similar content (see above comments).</p> <p>Not sure why results that do not adjust for multiple comparisons are shown. As far as I can tell, they do not provide additional meaningful information (especially as the 95% CIs are shown).</p> <p>Concerns about generalizability could be addressed somewhat by developing weights for the sample to make it more representative of age/sex/smoking status composition of the general population. This could be used for sensitivity analyses, with consistent results taken as evidence for the likely generalizability.</p> <p>[My prior comments are in response to the [methods report. Unless otherwise indicated, what follows is on the results report, although issues I raise above are pertinent to the background and methods sections of the results.]</p>

CHARGE QUESTION 8: <i>Is the analytic approach appropriate given the design and purpose of the study?</i>	
REVIEWER	COMMENT
Reviewer #3	What is not clear in the analysis plan of beliefs is whether any treatment vs. control exposure is expected to affect any and all beliefs tested. Obviously, the warning label in a condition only refers to particular beliefs and not all beliefs. So why should any other belief not referenced in the warning be affected by the warning? The partial answer to this is spillover through cognitive activation but without some idea of which beliefs are correlated with which other beliefs, there is no way of anticipating what is and is not affected by spreading activation. Not clear what exactly is being tested in the belief analysis plan e.g., all beliefs regardless of condition or all beliefs but as a function of condition. Why would the impact of one specific warning affect with a narrow focus affect all the other non-congruent beliefs?
Reviewer #4	The analytic plans represent a solid test of the study hypotheses. Correction for multiple tests is considered, and the results are presented and discussed plainly. Key results could also be presented via figures to improve communication of this densely packed set of results.
Reviewer #5	The analytic approach seemed appropriate. There were two primary comparisons: comparing the new warnings to the existing Surgeon General ones and comparing the change in health beliefs over two time periods using a difference in difference approach. Both are appropriate and yield slightly different information.
Reviewer #6	Yes. I’m slightly concerned about the number of times the participants were exposed to the health beliefs assessment and whether there could be an interaction with the condition. All participants were exposed to a large list of health effects of cigarettes that are, by design, novel. Then only those in the treatment condition are being exposed to the warnings with the new/novel information on it. It concerns me slightly, that there might be a priming effect of the health belief assessment for those participants in the treatment condition.

CHARGE QUESTION 9: <i>Are results presented consistent with the analytic approach?</i>	
REVIEWER	COMMENT
Reviewer #1	Yes.
Reviewer #2	The results are consistent with the analytic approach described.
Reviewer #3	The effects on beliefs from session 1 to session 3 are of course even more difficult to produce statistically significant effects but there is indeed some evidence of the remaining effect on beliefs even as long as two weeks out. Obviously, there are always competing explanations having to do with attrition rates and test sensitization among other things. But these are potentially strong findings from such a small and unfamiliar warning dose.
Reviewer #4	The analytic approach was very clearly described, and the execution and presentation follow from this approach.

CHARGE QUESTION 9: Are results presented consistent with the analytic approach?	
REVIEWER	COMMENT
Reviewer #5	Yes, the way that the results are presented is consistent with the planned analyses.
Reviewer #6	In the presentation of results, rather than order the tables by condition, which has no inherent meaning to the reader, consider leaving the control at the top and then ordering the remaining rows by the highest (new information, or highest mean). It allows the reader to see the labels from most effective to least effective on some of the main outcomes.

CHARGE QUESTION 10: Are there any concerns with the results presented?	
REVIEWER	COMMENT
Reviewer #1	My only concern is the need for an appropriate theoretical framework to justify the analytic approach and to allow the reader to put the results in context. To me, the results from this study, when taken with the results of Study 1, allow a conclusion that graphic text messages are much better than text only messages as they can lead to a modification of health beliefs which is a step further along in the behavior change process. Thus, the graphic warning labels are much more likely to lead to quitting than the text-only messages. I suggest that the two studies be considered in the same report so that such a conclusion can be drawn. I would put the combined methodology into an appendix to such a report.
Reviewer #2	<p>As far as I can tell, there was no assessment of whether the characteristics of the treatment and control groups were significantly different (i.e., whether randomization worked).</p> <p>The range of days from baseline to completion of session 2 and session 3 should be included. Just the median is currently provided.</p> <p>Table 3-1 would ideally include statistical tests for differences in sample composition across sessions. While most characteristics appear similar, there appears to be some meaningful attrition by age and smoking status over time.</p> <p>Differential attrition could be partly addressed by including as control variables the specific age X smoking status variables that appear to differ over time (e.g., a single category for never smoker susceptible adolescents) rather than what I interpret as controlling for age and smoking status as separate adjustment variables.</p> <p>The assessment of differential attrition by treatment and control groups is only done at the group level. There is no assessment of whether the sociodemographic and smoking status composition of the groups becomes more dissimilar over time. Indeed, such differences could emerge even if the overall attrition rate was the same across groups. The report would be</p>

CHARGE QUESTION 10: Are there any concerns with the results presented?	
REVIEWER	COMMENT
	<p>strengthened by this kind of assessment and inclusion of statistical controls for the characteristics that become significantly different over time.</p> <p>Table 3-3. Show the means for each SG warning done separately, as well as averaged together (see prior comment about analysis of specific SG warnings).</p>
Reviewer #3	<p>The strong effects obtained on judgments that the information presented in the new warnings is new would have been more persuasive to me if for example some false information was included but was not a part of the warnings or information that was not a part of the warnings was included. In order to see that the “thinking about risk” outcome is to some degree not completely dependent on the “new information” judgment which it follows in the questionnaire, some measure of association between the new information measure in the thinking about risks measure should be provided. What is also clear is that the perceived factualness of the warning is less accepted when the information is seen as new. So just as in the Study 1 people identify information that they hadn’t seen before as new but also less likely to be a fact rather than opinion and therefore less likely to be accepted which is a part of the learning process. What happened to believability?</p> <p>The fact that the recall item at session 3 was better recalled than the specific textual warning from the Surgeon General’s warnings is not at all surprising. The respondents were exposed to their specific warning within condition four times – session 1 and session 2 one on a cigarette pack and the other on an advertisement-- so to find that the recall of the text that they saw and read after four exposures versus no exposures is minimal evidence of the recall ability of the materials.</p> <p>Also, this is a recognition test not a recall test. Recognition tests are a lot easier than recall tests would be. A third factor of course is the presence of a visual image supporting the text in a way that is presumably concordant with the text that is being recalled. So, several factors argue in favor of successful recall that is recognition of the message to which they were exposed.</p>
Reviewer #4	<p>No concerns beyond the fact that the study was not powered to detect subgroup differences. While the report is very clear about this decision, it is unfortunate because to understand what works and for whom (and potentially why) are important questions given widespread distribution of the warning labels.</p>
Reviewer #5	<p>There are many results, but the report walks the reader through them systematically. The primary exception is that the power analyses were very confusing. I make some suggestions in the table below about how this could be more clearly presented.</p>
Reviewer #6	<p>Potentially. In addition to differential attrition by condition, I would like to make sure there wasn’t differential attrition among key demographic groups between time points, it wouldn’t impact the treatment/control findings, but it might hamper the generalizability of the findings if particular demographic groups were more likely to drop out. Looking at Table 3-1 it appears that low income individuals were dropping out more than higher income individuals, and that higher education</p>

CHARGE QUESTION 10: <i>Are there any concerns with the results presented?</i>	
REVIEWER	COMMENT
	participants were retained better than lower education. I would be curious to know if these demographics dropped out at different rates per condition.

CHARGE QUESTION 11: <i>Are potential limitations of the study appropriately identified?</i>	
REVIEWER	COMMENT
Reviewer #1	The section on limitations is appropriate.
Reviewer #2	<p>When raising the issue of the ecological validity of responses to online images of warning statements, the authors could cite studies that indicate convergent validity (see comments from Study 1).</p> <p>Concerns about generalizability could be addressed by citing studies showing that patterns of response in experiments are generally consistent across population sources. Also, as a sensitivity analysis, the authors could weight observations so that they are more similar to the profile of smokers and nonsmokers in the general population and evaluate the consistency of results.</p> <p>The limitations should better describe potential measurement issues raised above (e.g., “fit” of B1 for some warnings; measurement issues described for common measures with Study 1), including implications for interpreting results. See Study 1 comments on considerations of content validity around measurement of “understanding.”</p>

CHARGE QUESTION 11: Are potential limitations of the study appropriately identified?	
REVIEWER	COMMENT
Reviewer #3	<p>In discussing the demographics of the sample at times one through three I think it would be reasonable to compare the characteristics of the sample to more representative samples that have been gathered by other sources simply as a way of saying that the sample is close or far from established samples of smokers and susceptibles among young people.</p> <p>The attrition from time one to time three is substantial as would be expected but it is incumbent upon the researchers to discuss a little bit about who it is that remains in terms of things like a priori beliefs and how those change over time with the attenuated samples of sessions two and three. For example, adolescents who are susceptible to smoking seem to drop off from session 1 to session 3 and the percentage of adult non-smokers seems to increase from session 1 session 3. So, by session 3 the susceptible adolescents are down, and the adult non-smokers is up. The attrition by condition does not look appreciable to me on any criteria other than the couple of demographic differences that I noted above. But one of the things that makes a lot of sense to me is to report attrition as a function of session 1 beliefs. So, for example if there is evidence to show that those who are more accepting of the negative health consequences remain in the sample than they are going to be more likely to be attuned to the messages that the warnings carry being more engaged attentive and accepting. If that is the case then the very favorable outcomes observed are overstated.</p>
Reviewer #4	Major study limitations are described in the conclusions. Some of these limitations are quite serious (e.g., sampling) and the report does a reasonable job of explaining and describing implications of these limits.
Reviewer #5	Yes, key limitations are identified.
Reviewer #6	Mostly. I was interested in the possible impact of an interaction between testing and condition. All participants saw a list of health beliefs (many of which are considered novel), and then those in the treatment group saw a warning that potentially reinforces one of those novel health beliefs, while those in the control see information that is mostly not considered novel. I was wondering if there was a concern about priming people with the pre-test, and it being reinforced in the treatment condition, but not the control.

CHARGE QUESTION 12: Are the conclusions drawn from the study well supported by the data presented?	
REVIEWER	COMMENT
Reviewer #1	Yes, the conclusions presented follow from the analyses and results. However, they do not go far enough. From both of these studies, the authors are able to address why graphic warning labels should be preferred to text-only warning labels. This is very important to the consideration of policies on warning labels and the question should be addressed.

CHARGE QUESTION 12: Are the conclusions drawn from the study well supported by the data presented?	
REVIEWER	COMMENT
Reviewer #2	The lack of an overarching framework for and validity of the outcomes assessed makes it challenging to interpret results, particularly around factualness, which is lower for GHWs than SG warnings. The current explanation is neither based in empirical evidence nor theory.
Reviewer #3	<p>Let’s focus on the belief changes from session 1 to session 2 in contrast to the changes in the control condition. What’s clear from the results of Table 3.5 is that acceptance of the beliefs that smoking (for example) causes bladder cancer is elevated in the treatment conditions in contrast to the control conditions. But why would this be? First of all, the treatment conditions mentioned bladder cancer at both session 1 and session 2 but only in the warning condition that’s about bladder cancer. In the other 15 conditions there is no discussion of bladder cancer so why would the effect here on accepting bladder cancer as being caused by smoking be so distinctive across conditions when in fact only one condition mentions bladder cancer? Should it not be the case that the treatment condition mentioning bladder cancer should carry the weight of the impact and the other conditions mentioning other health consequences not be affected or at least affected to a much lesser degree. If that’s not the case then there is something else going on where in the warning about bladder cancer is dragging the other effects along with it on other health consequences and so the content of the warning is less consequential to belief change than one would expect. Clearly, if some form of spreading activation among potential beliefs is the basis for these effects then the particular warnings don’t matter as much to the effects on beliefs or there is some process other than the warnings that is producing these effects. I would be more convinced that it’s the content of the warnings if indeed what happened was that the bladder cancer warning as the greatest effect on the change from session 1 to session 2 than the other conditions do; alternatively that the bladder cancer condition affects the bladder cancer belief but not the other health consequences or does so to a lesser degree. This is easy to check and would make a stronger case that it is the warnings that produce the change rather than some other process obviously related to the warnings but not necessarily consistent with the content of the warnings.</p> <p>In the conclusion section, the authors argue that the skepticism attached to the graphic health warnings because they are new might disappear over time as exposure is increased. That may be true, but I think the reality is that a persuasion oriented campaign consistent with the claims that are being made here providing more elaborated evidence, information about credibility, and even testimonials to support the claims may be necessary. Otherwise, the warnings may be seen as new and unfamiliar and remain in the domain of opinion rather than fact, and less believable and accepted than would be desirable.</p>
Reviewer #4	The conclusions are generally non-numeric restatements of the findings and are accurately stated. However, some of the non-significant results for some of the warnings are discussed and explained; these sorts of explanations are largely speculative (if informed) and also do not match with other aspects of the report (where non-significant results are left unexplained). It may be best (most consistent) to not offer such explanations.
Reviewer #5	Overall, the conclusions are mostly a verbatim restatement of the results. The results are well supported by the data presented. However, there should be more synthesis of the vast amount of data presented. Also, there should be more of a summary of the merits or drawbacks of the individual warnings. I know

CHARGE QUESTION 12: Are the conclusions drawn from the study well supported by the data presented?	
REVIEWER	COMMENT
	that there is not power to compare warnings to each other, but I think the report should provide stronger guidance, based on multiple outcome variables and metrics, about whether one or more warnings should be dropped for consideration. As a whole, they fare well but 1-3 of these warnings do not score as well as the others. For instance, the Addictive and Quit Now warnings did not score as well, but did they score well enough to merit inclusion in the final list of warnings? Moreover, which one of the COPD warnings is better based on the data? I know that overall the majority of new warnings are better than the controls on the outcomes, however, I lost track of whether there are a couple of warnings fail to show a significant improvement on most of the outcomes. If so, that warning should not be recommended. Pointing that out makes it easier for the reader and regulatory agency to know if there are one or more weak warnings.
Reviewer #6	The study presents a summary of findings, which is a better descriptor than conclusions. The report does not conclude which labels are the best, or which of the results presented are used to determine the utility of the label for selection. It would be interesting to discuss whether there are certain characteristics of the labels that predicted better recall. Table 4-1 summarizes the findings nicely, but the summary does not provide any real “take-home” points.

Specific observations			
REVIEWER	PAGE	PARAGRAPH/ LINE	COMMENT
Reviewer #1			None provided.
Reviewer #2			None provided.
Reviewer #3			None provided.
Reviewer #4			None provided.
Reviewer #5	2 of 5		I would add a column showing the study condition # (0 – 16). See comment right below this row – use Table 2-1 here.
Reviewer #5	14	Table 2-1	This table is excellent – conveys a lot of information in one page.
Reviewer #5	15	Middle	The duration of exposure to the warnings should be described. Did the survey software require 5-10 seconds of exposure, or was the respondent allowed to click through at their own pace? Did the software track how long the respondent viewed the stimuli? Presenting that they viewed that page for something like an average of 9 seconds would be useful.

Specific observations			
REVIEWER	PAGE	PARAGRAPH/ LINE	COMMENT
Reviewer #5	17	Middle	The quality control description is useful, and I agree with that approach. However, I do not believe that the report lists the # or % of responses that were rejected for quality control purposes. Please show that somewhere, perhaps on this page or on the participant flow diagram.
Reviewer #5	17	Sentence above Sec 2.3	I would say “not <u>necessarily</u> representative....” These findings might actually be representative.
Reviewer #5	18	Anywhere	The mode of data collection as a study eligibility requirement should be mentioned earlier, perhaps on this page. Later in the report we learn that people taking the survey on their phone or tablet would get it discarded.
Reviewer #5	24	All	The report gives excellent data on the # of respondents (and their demographics and smoking status) to each of the 3 data collection periods. I know there is a lot of data, but it is useful.
Reviewer #5	27	Table 4-2	This whole section is really hard to follow. I believe that the researchers eventually calculated the within-person correlation – can you mention that or highlight it, so we know the actual power in this study? Was is >90% for most analyses?
Reviewer #5	28	4.3.1 after first para	This is pretty complex, and I would walk the reader through one example. The example on p.105 for difference in difference was very useful for the reader.
Reviewer #5	94	2.3	The report mentions that the comparison for the Control group is a pooled estimate for the 4 warnings (and not the participant reaction to just 1 of the SG warnings that they viewed). However, on Table 3-5 (p.106) the means for the controls are different (Session 1 – 3.35, 3.94, 4.37). I thought the control numbers should be the same if the data are pooled. I clearly did not understand the control comparison.
Reviewer #5	99	Table 3-1	Attrition across Sessions was fairly high. The good news is that it did not vary by condition. However, attrition seemed higher for smokers. I would add a nonresponse or attrition analysis for demographics and smoking behavior to show if attrition was differential.
Reviewer #6			None provided.