PEER REVIEW HISTORY

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Which type of tobacco product warning imagery is more effective and sustainable over time? A longitudinal assessment of smokers in Canada, Australia, and Mexico
AUTHORS	Anshari, Dien; Yong, Hua; Borland, Ron; Hammond, David; Swayampakala, Kamala; Thrasher, Jim

VERSION 1 – REVIEW

REVIEWER	Leas, Eric University of California: San Diego, Division of Global Health, Family and Preventive Medicine
REVIEW RETURNED	24-Feb-2017

GENERAL COMMENTS	Seems like there might be some nuggets, but, in my view, the authors haven't dug them out yet. Hopefully the reviews help them out! This article reports evidence from a online consumer panel who rated their perception of cigarette warning labels in three countries over several waves of data. While the article is well written, I have some questions about the conclusions drawn from the data and hope the authors find the information useful. First, the authors conclude that: "changing the design elements rather than just the propositional contents of PHWs may be a more effective way to maintain warning impact." This conclusions is unclear to me for two reasons. It is unclear what the authors are trying to "effect" and how changing something like the color (vs. the images/text) would be more "effective." For instance, is it less expensive or more feasible to change design elements other than images? Is the effect stronger on the desired outcome (e.g., getting smokers to quit)? How do your results provide evidence for any beneficial effect of changing design features over changing an image, when other design features are not assessed in this study? Second, the authors conclude that: "This study also shows that PHWs with suffering and graphic imagery appear to have different routes of impact and may work in complementary fashion in achieving the intended effects of PHWs." What outcome are you

	trying to “impact?” (e.g., quitting? preventing youth from smoking?). Also, different “routes of impact” suggests causal mediation/moderation. As the public health outcome is not clear, it’s impossible for me to judge whether there are different routes by which these PHWs might be having an impact even if, as you’ve shown, the types of images achieve different ratings or behavioral responses. Third, why are the images grouped as they are (suffering/symbolic/graphic) and not in other ways? Just one other way to group the images is by type of disease presented in the image. One could reasonably expect that different diseases within one grouping could have differential ratings on the scales you’ve used. For instance, the evidence, from Australia’s pre-market studies strongly supports the notion of difference in the believability of smoking-attributable diseases; their focus groups doubted the believability of the gangrene claim, while many talked about knowing smokers who have had emphysema, making that claim much more believable (https://goo.gl/U2ooI6). The data you have are on the individual images, so why not assess the performance of each image over time? Which images held salience the longest? Did they all have the same trajectory on these scales over time? Did any have a unique trajectory? It seems like the main goal of papers like this is to tell policy-makers which images they should implement. We know that particular images can be rated as more effective at communicating health risks (including some of the images presented in the current study--see: https://goo.gl/KUPQji), but we don’t know if such effects persists or if images that do not perform as well in pre-market become more effective over time. This study starts to get at that. I could a couple scenarios where this information is very useful to a policy maker. One scenario is if a certain image performers very well in pre-market and then tanks immediately after reaching the market. Another is if an image performers terribly in pre-market and then succeeds when reaching the market.

REVIEWER	Rogers, Todd RTI International, Center for Health Policy Science and Tobacco Research
REVIEW RETURNED	07-Mar-2017

GENERAL COMMENTS	So many of the findings seem to be supportive of what is already known about the subject, and the concerns I expressed to the authors about the framing need to be addressed before this is publishable. This study assessed responses by online panels of adult smokers in Australia, Canada, and Mexico to actual, current country-specific pictorial health warnings (PHWs) that varied by imagery type. Through longitudinal assessments, the investigators sought to answer questions about the relationship between PHW features (imagery type) and affective, cognitive and motivational responses of respondents, and whether/how responses changed over time and as function of imagery type. The investigators constructed a relatively well-designed study and conducted appropriate analyses. My primary concern with this paper reflects the major limitation stated by the authors regarding, “the differences in stimuli by country and within each category, and in some cases within country over time. Hence, interpretations around cross-country comparisons

	should be tempered by this regard" (p. 13). There are not only methodological differences across countries in the study (e.g., sample characteristics, PHW characteristics, how PHWs are rotated in each country), but also idiosyncratic historical and cultural differences that militate against generalizable statements of results across countries. And yet, despite their caution to temper cross-country interpretations, the paper nonetheless includes statements that a reader would not recognize as temperate; for example, "It is notable that the relative effects of PHW imagery type on quit motivation were quite different across the countries, with no differences between graphic and suffering PHWs in Canada, whereas graphic PHWs were superior to suffering PHWs in Australia, while the reverse was found in Mexico" (p 12-13). I suggest that the authors search for all such statements in the paper and consider revising to address their temperance admonition. Other specific comments: Page 11, lines 4-15: Terms such as "desirable responses" and "superior to each other" are not defined. Do these refer to the lack of wear-out across all dependent measures? Page 13, lines 6-11: If you think that the "mixed findings" reflect country differences in the number of stimuli included in the study, why not control for this by adding that as an adjustment variable to models? Page 14, lines 9-13: If findings could be due to differential quality of images, why not independently rate image quality and control for this in the models? Table 2: Pet peeve — p values of 0.000 should be presented as <0.001. In fact, with the 95% CI values presented, you could eliminate the P>z columns and use superscripts and footnotes for p < 0.05, <0.01, <0.001.

VERSION 1 – AUTHOR RESPONSE

Responses to the Editorial requests

- Responses are in bullet points

---

1) In accordance with our editorial policies, please ensure your manuscript reporting adheres to the STROBE guidelines (http://www.equator-network.org/reporting-guidelines/strobe/) for the reporting of observational studies. This is so your methodology can be fully evaluated. Please include the STROBE checklist as a supplementary file in your next submission indicating the page/line numbers where the requested items can be found in your manuscript, and add a statement to the Methods section of your manuscript that these guidelines were followed.

- Thanks for this. We have included STROBE checklist as a supplementary file and added a statement in the Methods section (See Page 6 line 6).

2) Please clarify whether the images included in the article are subject to copyright, and in that case, whether you obtained permission to publish them. Please add a statement to your manuscript in accordance to your reply.

- The images we used for this study are in the public domain, as they are images that are printed on cigarette packs that you can purchase in each country. For that reason, they can be used for research purposes, and we assume that they can be published in scientific manuscripts as well. We have added a statement for this under the Figure 1 (See Page 21 line 5-6).

3) Although we note that the study received ethics approval by the South Carolina Institutional Review Board, we would expect that different local committees granted permission to conduct the approval in the 3 different countries. Did the ethics approval obtained include all participants? Did you obtain permissions to use/analyse the different data sets from the local source in each country? Please add a statement to your manuscript in accordance to your reply.

- We received ethics approval from the IRB at the University of South Carolina, as well as the IRB at the University of Waterloo. Our study participants came from a consumer panel used for market research, and all contacts with participants were managed by a private company (i.e., the Lightspeed GMI). The datasets we received were not of a medical nature and did not include any information that would allow us to identify participants. Because of this, our project partner in Australia deemed that the IRB approvals from other institutions were sufficient. We have added information about ethics approval from the University of Waterloo in the manuscript (See Page 18 line 15-16).

Response to the Reviewer(s)' Comments to Author

- Responses are in bullet points

Reviewer: 1

This article reports evidence from a online consumer panel who rated their perception of cigarette warning labels in three countries over several waves of data. While the article is well written, I have some questions about the conclusions drawn from the data and hope the authors find the information useful.

First, the authors conclude that: “changing the design elements rather than just the propositional contents of PHWs may be a more effective way to maintain warning impact.” This conclusions is unclear to me for two reasons. It is unclear what the authors are trying to “effect” and how changing something like the color (vs. the images/text) would be more “effective.” For instance, is it less expensive or more feasible to change design elements other than images? Is the effect stronger on the desired outcome (e.g., getting smokers to quit)? How do your results provide evidence for any beneficial effect of changing design features over changing an image, when other design features are not assessed in this study?

- The reviewer is correct. We have changed our conclusion so that it suggests this as a possible avenue for future research, and such research will be useful as over 100 countries have rotating pictorial warnings for which they have the opportunity to change warning content & design.

Second, the authors conclude that: “This study also shows that PHWs with suffering and graphic imagery appear to have different routes of impact and may work in complementary fashion in achieving the intended effects of PHWs.” What outcome are you trying to “impact?” (e.g., quitting? preventing youth from smoking?). Also, different “routes of impact” suggests causal mediation/moderation. As the public health outcome is not clear, it’s impossible for me to judge

whether there are different routes by which these PHWs might be having an impact even if, as you've shown, the types of images achieve different ratings or behavioral responses.

 We appreciate the request for greater detail to better make our point. Our key warning response measures (i.e., affect, cognition, behaviors) have theoretical and prior empirical support as mediating variables for warning effects on smoking cessation behaviors (See Page 7, lines 17-19). While we find a somewhat different pattern of responses to how participants rate graphic and suffering warnings on their packs. Our point is that no one warning type is rated consistently higher than the other on these key mediating variables. We have tried to clarify this (See Page 12 line 22 to Page 13 line 3). We also now mention how more formal tests of mediation may help determine whether the balance of imagery on warnings should be in favor of one type or another (See Page 16 lines 3-5).

Third, why are the images grouped as they are (suffering/symbolic/graphic) and not in other ways? Just one other way to group the images is by type of disease presented in the image. One could reasonably expect that different diseases within one grouping could have differential ratings on the scales you've used. For instance, the evidence, from Australia's pre-market studies strongly supports the notion of difference in the believability of smoking-attributable diseases; their focus groups doubted the believability of the gangrene claim, while many talked about knowing smokers who have had emphysema, making that claim much more believable (https://goo.gl/U2ooI6). The data you have are on the individual images, so why not assess the performance of each image over time? Which images held salience the longest? Did they all have the same trajectory on these scales over time? Did any have a unique trajectory?

 In the second paragraph of the Introduction section, we have added justification for our classification of images (See Page 4, lines 13-17). Using fear appeal theory, we classified the images based on their level of gruesome content (O'Keefe, 1990; Witte, 1992), ranging from the most frightening (graphic), less frightening (suffering), and least frightening (symbolic). We then used the negative affects rating to confirm this classification of images and it showed that such classification persists in all countries (See Page 13, lines 22-23).

It seems like the main goal of papers like this is to tell policy-makers which images they should implement. We know that particular images can be rated as more effective at communicating health risks (including some of the images presented in the current study--see: https://goo.gl/KUPQji), but we don't know if such effects persists or if images that do not perform as well in pre-market become more effective over time. This study starts to get at that. I could a couple scenarios where this information is very useful to a policy maker. One scenario is if a certain image performers very well in pre-market and then tanks immediately after reaching the market. Another is if an image performers terribly in pre-market and then succeeds when reaching the market.

 The reviewer is correct. We aim to provide evidence for policy-makers on which images may work well over time. Prior evidence on the superiority of certain types of images mainly came from pre-market experimental studies. Nevertheless, as we state in our Introduction (See Page 5, lines 4-11), there is very little research on the validity of pre-market experiments for determining pictorial warning content that is most effective after policy implementation. Our study aimed to fill that gap by embedding specific warning rating methods used in experimental research into a longitudinal study design of consumer responses post-implementation of new warnings. As we state in our conclusions, this information may be helpful for decision makers in many of the 100 countries with rotating pictorial warning label systems for cigarettes. We also underscore how this study may also inform future pre-market experimental studies on specific warning content and design features, so that results inform field trials and post-market studies.

Reviewer: 3

This study assessed responses by online panels of adult smokers in Australia, Canada, and Mexico to actual, current country-specific pictorial health warnings (PHWs) that varied by imagery type. Through longitudinal assessments, the investigators sought to answer questions about the relationship between PHW features (imagery type) and affective, cognitive and motivational responses of respondents, and whether/how responses changed over time and as function of imagery type. The investigators constructed a relatively well-designed study and conducted appropriate analyses.

My primary concern with this paper reflects the major limitation stated by the authors regarding, "the differences in stimuli by country and within each category, and in some cases within country over time. Hence, interpretations around cross-country comparisons should be tempered by this regard" (p. 13). There are not only methodological differences across countries in the study (e.g., sample characteristics, PHW characteristics, how PHWs are rotated in each country), but also idiosyncratic historical and cultural differences that militate against generalizable statements of results across countries. And yet, despite their caution to temper cross-country interpretations, the paper nonetheless includes statements that a reader would not recognize as temperate; for example, "It is notable that the relative effects of PHW imagery type on quit motivation were quite different across the countries, with no differences between graphic and suffering PHWs in Canada, whereas graphic PHWs were superior to suffering PHWs in Australia, while the reverse was found in Mexico" (p 12-13). I suggest that the authors search for all such statements in the paper and consider revising to address their temperance admonition.

- Thanks for this correction! We have replaced statements that did not recognize country differences (See Page 14 line 22 to Page 15 line 10).

Other specific comments:

Page 11, lines 4-15: Terms such as "desirable responses" and "superior to each other" are not defined. Do these refer to the lack of wear-out across all dependent measures?

- The term "desirable responses" refer to the five key outcome variables in this study, we have included these in parentheses following the term (See Page 12, line 19-20). While the term "superior to each other" was meant to the findings that graphic warnings for some outcomes were superior than suffering warnings, while the reverse was true in other outcomes. We decided to take out this term to reduce misinterpretation.

Page 13, lines 6-11: If you think that the "mixed findings" reflect country differences in the number of stimuli included in the study, why not control for this by adding that as an adjustment variable to models?

- We did the analysis stratified by country, so we could not add adjustment variables by country.

Page 14, lines 9-13: If findings could be due to differential quality of images, why not independently rate image quality and control for this in the models?

- Our limitation actually stated "…our findings could be due to the quality of the messages,…", not quality of images. We meant to refer to the textual and other message features, separate from

the images. We have clarified this now.. We now state: our findings could be due to the quality of the textual content or other message features,…", not quality of images (See Page 16, lines 9-10).

Table 2: Pet peeve — p values of 0.000 should be presented as <0.001. In fact, with the 95% CI values presented, you could eliminate the P>z columns and use superscripts and footnotes for p < 0.05, <0.01, <0.001.

- Thanks! We have corrected them.

VERSION 2 – REVIEW

REVIEWER	Eric Leas UC San Diego
REVIEW RETURNED	02-Jul-2017

GENERAL COMMENTS	Objectives and purpose of the study are clear and the manuscript is otherwise well written. I have no further comments following my previous review.