**Peer Review History and Correspondence**

# MS Title: Straight from the scientist's mouth - Plain language summaries promote laypeople’s comprehension and knowledge acquisition when reading about individual research findings in psychology

# Author Name: Martin Kerwer, Anita Chasiotis, Johannes Stricker, Armin Günther, and Tom Rosman

**Editor First Decision—Revise & Resubmit**

Oct. 30, 2020

Dear Dr. Kerwer,

I want to begin by apologizing for the delay in making this streamlined decision. The delay was entirely my fault and I apologize for any problems this may have caused for you. I have now read over your manuscript, “Straight from the scientist’s mouth - Plain language summaries promote laypeople’s comprehension and knowledge acquisition when reading about individual research findings in psychology” as well as the previous decision and reviews that you submitted. I found the manuscript to be quite strong, and had only a few small for revisions. I believe all of the changes I request are easy to make, and the rest are suggestions for you to decide whether to take or leave. I am confident that if you are willing to make these changes, this manuscript can be accepted for publication. I therefore encourage you to submit a revised version to Collabra: Psychology.

I will list my recommendations here, separated into two categories. The first are changes that I would need to see before accepting your manuscript (though, as always, you are free to respond with an explanation for why you are choosing not to make this change and I will consider your arguments). The second are suggestions for you to consider. In your resubmission, please include a document with a point-by-point response to both the points I list here, outlining your response to each point.

I want to begin by saying that I was very impressed with this manuscript. The writing is very clear. The pre-registeration is helpful and the results are presented in a way that makes it easy to tell what was pre-registered and what wasn’t. And your interpretation of the results was well-calibrated to what your study could and couldn’t test. It was a pleasure to read this manuscript.

Required changes (or compelling counterarguments):

1. One major issue that I would like to see you tackle head-on more than you currently do is the question of whether increasing readers’ trust in (or their perceived credibility of) scientific reports is the right goal. In some places in your manuscript, you use language that implies that this is necessarily a good thing (e.g., phrases like “sustaining trust in science” imply that the correct level of trust is a high one). In other places, you do either imply or in one case directly address the possibility that high levels of trust may not always be warranted. I would like you to add some text (e.g., at least one paragraph in the intro, probably around the top of page 3, and also at least one paragraph in the discussion) where you articulate the issue at play here. You are free to take whatever position you want, of course, or not take a position one way or the other, but I would like readers to see an explanation of the distinction between the view that it is always a good thing to increase the public’s trust in a particular scentific field or finding vs. the view that public trust should be earned and increasing trust when the field or finding are not solid could be a bad thing.

I noticed in the materials you submitted for streamlined review that you took out some language that related to this issue in interpreting your results (e.g., removing the sentence “This might lead to the problem that not only scientifically valid, but also questionableresearch which only pretends to apply a scientifically rigorous method would be trusted more.” and removing some parenthetical remarks in another sentence). Personally, I think the manuscript would be stronger with that language back in. However, whether you should put it back in or not depends on what position you take in the introduction and discussion about whether increasing trust/credibility is necessarily the right goal in all cases. (One could even argue that we should be skeptical that increasing accuracy/understanding is always a good thing - if the claims that readers are more accurately understanding are actually not warranted, increasing understanding could be counterproductive! I probably wouldn’t go that far, though, because even if the claims are exaggerated, readers should at least be able to accurately decipher what the author wants them to believe - the skepticism should come after understanding, not undermine understanding.)

I would like to ask that you flag this issue – that there could be disagreement about whether higher levels of trust/credibility are necessarily a good thing (or, if you take the position that they’re aren’t necessarily a good thing, then flag your position) – in the abstract and in the conclusion as well.

1. Please describe what the advertisements for recruiting participants to the study said. This will help readers get a better idea of who might have been more likely to select into the study, and thus of the population of people to which your findings are more likely to generalize. Please also consider whether the recruitment materials/advertisements may have introduced a selection bias that could affect the interpretation of your results, and include any relevant discussion of such possible biases in your discussion.
2. For all of your tables, please add more information so that a reader could interpret the table without referencing the main text/methods. This includes providing more information about the scales for the responses (e.g., 1 to 8, etc.), and more explanation of what the variables mean (e.g., what is “semester” in Table 2). Please consider the point of view of something flipping through your paper who just wants to look at the tables – I realize you can’t convey all the relevant details of the method in the tables, but please try to make them as easy to understand as possible on their own. You may want to solicit the help of someone who isn’t familiar with the paper (e.g., have them read just the abstract and then look at the tables, and ask them what other information could be provided in the table to make them easier to interpret).
3. Please provide more information for readers about the size of the effects in the results section. Assume your readers do not know how to interpret the statistics you report, and provide some aid in understanding how practically meaningful the effects are. Please also incorporate a consideration of effect sizes in your own interpretation of the results – the statistical significance is useful for understanding how likely the result would have been if the null hypothesis is true, but a more comprehensive interpretation of the result would also include some consideration of practical significance/effect size. (Related to this, I wonder if most readers will know what a likelihood ratio test is – maybe consider giving readers extra information especially about this test and how to interpret its results?)
4. Related to this, I would like you to consider how your null results should be interpreted. Right now, you don’t say much about them, and I’m not sure if readers will interpret them as evidence of absence, or just absence of evidence. Can you provide more information about how you interpret the null results, and what you think they can or can’t tell you about evidence of absence? This comes up in a few places.
5. On page 35, you say that the results “illustrate that plain language summaries…” – this is in the context of describing the exploratory findings. I think “illustrate that” is too strong for the exploratory results.
6. One aspect of your study that I found interesting to think about is that this is indeed an experiment, but because you used actual original abstracts and plain language statements, you did not have experimental control over the differences between these two conditions. This challenged me to think about whether this could introduce “confounds” and I came to the conclusion that “confounds” is the wrong way to think about this – the better way to think about is that the independent variable is not whether the summary presented is in scientific vs. plain language, but you manipulated whether readers saw the authors’ original abstract vs. plain language summary, with all of the other changes this entails. The fact that you used actual abstracts/summaries is a positive feature, and it also changes how we should think about the specific independent variable being manipulated – whatever else changes systematically between authors’ original abstracts and plain language summaries (at this journal, and more generally) was also different between the conditions here. That means we can’t be sure exactly what the mechanism is that is driving the effect, but we can be more confident that it translates to actual original abstracts vs. plain language summaries. Given how central this is to understanding what conclusions can be drawn from your study, and what the conceptual independent variable is in your study, I would strongly recommend that you include a paragraph or two discussing this issue. I could imagine it would be useful to have a paragraph around the place where you describe the study/hypotheses, and then another paragraph in the discussion revisiting this issue.
7. I was very glad to see that you replaced the PLS acronym with the full phrase. I just found one place where you still use the acronym (p. 5) - please replace it and check and see if there are others.

Suggestions:

1. On page 2, you state that plain language summaries “may be seen as a safeguard against the sometimes misleading or distorting press releases from universities or other research institutions”. However, my understanding (and my experience) is that much of the exaggeration in press releases actually comes from the authors themselves (the people who write the press releases typically interview the authors, who provide the exaggerated claims). Perhaps more importantly, I don’t think you need this point, and including it (especially in the very first paragraph) may distract from the actual focus of the paper, since you do not actually compare plain language summaries to press releases. I would suggest removing this point.
2. I wonder if adding some data visualizations would help (see point 4 above). In some cases, it is hard to grasp the meaning of the result, and the uncertainty/variability. Data visualizations that include individual data points (e.g., raincloud plots, or something like that) might make it easier for readers to figure out how practically significant they think the results are. This could also be useful for null results (point 5 above).
3. In my opinion, you are over-interpreting the results of some of the exploratory analyses. For example, the results presented in the section on “The role of justification beliefs and English proficiency” are all quite similar to one another, but some cross the p < .05 threshold and others don’t. In my view, interpreting those that cross the threshold differently from those that don’t doesn’t make a lot of sense, particularly in the context of exploratory (i.e., non-pre-registered analyses), and given that the effect sizes are not that different for the significant vs. non-significant results. In addition, given that you ran a number of exploratory tests, it seems unlikely to me that the significant results would survive a correction for multiple comparisons. This isn’t the most important factor, and of course interpreting the results of exploratory tests is a subjective exercise, so I will leave it up to you to decide how you want to interpret these, but for what it’s worth I didn’t agree with your interpretation. Another relevant factor to consider for many of the exploratory results is that your power for detecting between-person effects is probably not very high, so null results may not be very meaningful, and all effects are likely to be imprecisely estimated (and therefore small-to-medium differences between effects are probably not reliable).
4. “we aim to close this knowledge gap” on page 12 sounds a bit bold – maybe tone it down?
5. In your sample size calculation, it might be good to say what “VPC” means (p. 23)
6. I didn’t understand the sentence beginning “To put it bluntly, individuals…” on page 34.
7. The discussion is quite long and a bit repetitive, you may want to consider shortening it.

In addition, I made a few notes on grammatical issues that I will share with you here, but I want to reiterate that overall I found the writing very clear. -p. 3: “test on the basic notion” should be “test of the basic notion” -p. 4: “whether beliefs on the justification” should be “whether beliefs about the justification” -p. 6: “and especially experimental research” – this sentence is quite long. I would cut “and especially” and start a new sentence with “Experimental research…” -p. 7: “our knowledge on how plain” should be “our knowledge of how plain” -p. 7: “public health has already” should be “public health have already” -p. 8: “agree more often to corresponding” should be “agree more often with corresponding” -p. 10: “little is known on how” should be “little is known about how” -p. 28: “more curios” should be “more curious” -p. 34: I think that “and ordinary scientific abstracts (rather short but less comprehensible texts)” should be “compared to ordinary scientific abstracts (rather short but less comprehensible texts)”

In summary, I think this is a promising manuscript and I hope you will submit a revision to Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Simine Vazire Editor in Chief Collabra: Psychology

**Author Response**

**Dec 7, 2020**

In this letter, we (the authors) outline the changes made to our manuscript “Straight from the scientist's mouth - Plain language summaries promote laypeople’s comprehension and knowledge acquisition when reading about individual research findings in psychology” in response to comments and suggestions by Simine Vazire.

*Note:* Formatted as bold text are the original comments as they were sent to us. Our responses are formatted non-bold and changes in the text are printed in italics.

**Dear Dr. Kerwer,**

**I want to begin by apologizing for the delay in making this streamlined decision. The delay was entirely my fault and I apologize for any problems this may have caused for you. I have now read over your manuscript, "Straight from the scientist's mouth - Plain language summaries promote laypeople’s comprehension and knowledge acquisition when reading about individual research findings in psychology" as well as the previous decision and reviews that you submitted. I found the manuscript to be quite strong, and had only a few small for revisions. I believe all of the changes I request are easy to make, and the rest are suggestions for you to decide whether to take or leave. I am confident that if you are willing to make these changes, this manuscript can be accepted for publication. I therefore encourage you to submit a revised version to Collabra: Psychology.**

**I will list my recommendations here, separated into two categories. The first are changes that I would need to see before accepting your manuscript (though, as always, you are free to respond with an explanation for why you are choosing not to make this change and I will consider your arguments). The second are suggestions for you to consider. In your resubmission, please include a document with a point-by-point response to both the points I list here, outlining your response to each point.**

**I want to begin by saying that I was very impressed with this manuscript. The writing is very clear. The pre-registeration is helpful and the results are presented in a way that makes it easy to tell what was pre-registered and what wasn't. And your interpretation of the results was well-calibrated to what your study could and couldn't test. It was a pleasure to read this manuscript.**

Thank you very much for this appreciation of our efforts. and the insightful suggestions you made which helped us to further improve the manuscript.

**Required changes (or compelling counterarguments):**

1. **One major issue that I would like to see you tackle head-on more than you currently do is the question of whether increasing readers' trust in (or their perceived credibility of) scientific reports is the right goal. In some places in your manuscript, you use language that implies that this is necessarily a good thing (e.g., phrases like "sustaining trust in science" imply that the correct level of trust is a high one). In other places, you do either imply or in one case directly address the possibility that high levels of trust may not always be warranted. I would like you to add some text (e.g., at least one paragraph in the intro, probably around the top of page 3, and also at least one paragraph in the discussion) where you articulate the issue at play here. You are free to take whatever position you want, of course, or not take a position one way or the other, but I would like readers to see an explanation of the distinction between the view that it is always a good thing to increase the public's trust in a particular scentific field or finding vs. the view that public trust should be earned and increasing trust when the field or finding are not solid could be a bad thing.**

**I noticed in the materials you submitted for streamlined review that you took out some language that related to this issue in interpreting your results (e.g., removing the sentence "This might lead to the problem that not only scientifically valid, but also questionable research which only pretends to apply a scientifically rigorous method would be trusted more." and removing some parenthetical remarks in another sentence). Personally, I think the manuscript would be stronger with that language back in. However, whether you should put it back in or not depends on what position you take in the introduction and discussion about whether increasing trust/credibility is necessarily the right goal in all cases. (One could even argue that we should be skeptical that increasing accuracy/understanding is always a good thing - if the claims that readers are more accurately understanding are actually not warranted, increasing understanding could be counterproductive! I probably wouldn't go that far, though, because even if the claims are exaggerated, readers should at least be able to accurately decipher what the author wants them to believe - the skepticism should come after understanding, not undermine understanding.)**

**I would like to ask that you flag this issue – that there could be disagreement about whether higher levels of trust/credibility are necessarily a good thing (or, if you take the position that they’re aren’t necessarily a good thing, then flag your position) – in the abstract and in the conclusion as well.**

Thank you for pointing out this central issue. Let us first refer to the issue of trust in scientific findings in general. In our personal opinion, a certain amount of trust in science by the public seems to be necessary. If this trust does not exist, scientific findings are at risk of being marginalized when it comes to decision making processes and this may also lead to proliferation of conspiracy theories (anecdotal evidence for this sad truth can be currently seen in the unwillingness of some individuals and governments to uptake scientific advice during the COVID-19 crisis). However, we think that trusting claims just because they come from science is problematic, too. In our view, plain language summaries may, however, be a way to earn this trust. To provide a more nuanced and elaborated view on this matter, we extended the following paragraph in the introduction on page 3 (new content is underlined):

*We argue that sustaining trust in science is especially relevant in (social) psychology—a discipline which investigates topics of high relevance to the public, but which also struggles with a replicability crisis (Klein et al., 2018; Open Science Collaboration, 2015; Świątkowski & Dompnier, 2017) and has been stricken by several misconduct cases lately (Callaway, 2011; Stricker & Günther, 2019; Świątkowski & Dompnier, 2017). Without this trust, scientific findings are likely at risk of being marginalized, which may even lead to a proliferation of conspiracy theories. On a more fine-grained level, one may, however, question if high trust in science is desirable in all circumstances—especially if there is uncertainty with regard to the robustness of scientific findings in a certain discipline. Since plain language summaries are a way of achieving higher transparency of the research process (Barnes & Patrick, 2019; Kuehne & Olden, 2015), we argue that they may provide a way to earn trust for those studies that actually deserve it. A recent study corroborates this argument (Carvalho, Elkins, Franco, & Pinto, 2019): scientific articles that provided plain language summaries had a higher methodological quality compared to articles that did not. In line with this reasoning, the American Psychological Association (APA) emphasized that “translating psychological science to the public” may help in addressing common misconceptions on the supposedly lacking rigor of psychological science (Kaslow, 2015).*

Our position is, as can be seen here, that we generally see trust in science as a good thing, but that it is also important to establish effective measures to ensure that the claims put forward in plain language summaries are actually warranted.

Possible ways to address these issues at the author and journal level are discussed on page 38:

*The authors’ ability to transparently communicate their findings without exaggerations is, therefore, of utmost importance when it comes to justifying this increased trust in plain language summaries. We argue that psychological science needs to establish (more detailed) guidance and best practices on how to communicate and report the quality of evidence to lay audiences to support researchers in this task. Journals might further improve the adequateness of the claims provided in plain language summaries by making them part of the review process. This is already done in some journals (e.g., Diabetes Therapy, n.d.). A different approach at the journal level is to have independent writers compose plain language summaries (e.g., King, Pewsey, & Shailes, 2017).*

However,we should not forget that the articles we presented to our participants were actually published in a peer-reviewed journal and that a high perceived credibility in their findings might indeed be justified. Providing laypersons with plain language summaries might just allow them to appreciate the (overall) quality psychological science has. We added this more positive note when discussing our findings on the easiness effect on page 38f:

*Taking a more positive stance, one might also argue that participants were simply able to more strongly appreciate the quality of the studies we presented when they received plain language summaries with subheadings. This would imply that trust can be already earned by means of providing plain language summaries. As a theoretical underpinning for this argument, we draw on the theoretical distinction between first-hand and second-hand evaluations of knowledge claims put forward by Bromme, Kienhues, and Porsch (2010). First-hand evaluations focus on directly evaluating the veracity of knowledge claims. For example, one might investigate the logical coherence of a study’s argument, evaluate its methodological approach (e.g., design, sample size, etc.), or compare it with other studies on the same subject (Bromme et al., 2010). In contrast, second-hand evaluations do not focus on the knowledge claim itself, but on the credibility of the corresponding source (Bromme et al., 2010). One might, for example, check whether a certain claim was brought forward by a known conspiracy theorist or by a renowned scientist—and choose to only believe the claim in the latter case. We argue that plain language summaries may be regarded as a way to facilitate first-hand evaluations by helping individuals to overcome barriers associated with technical terminology (cf. Nunn & Pinfield, 2014). In fact, understanding what was (not) done in a study, enables laypeople to make, at least to a certain extent, an informed decision about its respective quality. Current findings by Hoogeveen, Sarafoglou, and Wagenmakers (2020) support this view as they suggest that laypersons are indeed able to evaluate the credibility of research findings (to a certain extent) if they are presented in plain language. In their study, based on non-technical summaries, laypersons were able to predict how likely study findings were to be replicated and also took adequately into account (if provided) information on the strength of evidence. Since our participants reported an increased ability to evaluate the corresponding studies for plain language summaries with subheadings (H4a), more first-hand evaluations will likely have taken place if this type of research summary was presented. If we assume that most peer-reviewed journals, such as JSPP, publish, at least to a large extent, high-quality research, first-hand evaluations of our participants were likely to be mostly positive. Consequently, such evaluations would result in a higher perceived credibility of the corresponding findings. Naturally, to enable laypersons to make such (basic) first-hand evaluations on the quality of the research at hand, researchers need more guidance on how to present and discuss their research as transparently and accurately as possible in plain language summaries.*

And a note as to the role open science can play in this regard is also given on page 39:

*We may also provide laypeople with decision aids that further support them in their first-hand evaluations. We argue that open science badges might be a viable option to provide laypeople with this information as adherence to open science practices is considered to be an indicator of scientific rigor (Prager et al., 2019). In fact, open science practices reduce researchers’ degrees of freedom, which effectively reduces questionable research practices such as HARKing (hypothesizing after the results are known; Chambers, 2019), for example.*

Moreover, we also included passages to flag our position now in the conclusion (and we also changed the order of arguments here):

p. 45

*Second, in line with the easiness effect of science popularization, we found that individuals perceived plain language summaries with subheadings as more credible and had higher confidence in their ability to make decisions based on those plain language summaries. If this increased perceived credibility is always necessarily a good thing is, however, debatable—at least if there are no effective measures in place to ensure that the claims put forward in plain language summaries are actually warranted by the empirical evidence of the corresponding study.*

And in the abstract (p. 2):

*Moreover, in line with the easiness effect of science popularization, individuals perceived plain language summaries as more credible and were more confident about their ability to make a decision based on plain language summaries. If and under which circumstances this higher perceived credibility is justified, is discussed together with other practical implications and theoretical implications of our findings.*

Finally, we put some of the information in brackets on page 4 back in.

1. **Please describe what the advertisements for recruiting participants to the study said. This will help readers get a better idea of who might have been more likely to select into the study, and thus of the population of people to which your findings are more likely to generalize. Please also consider whether the recruitment materials/advertisements may have introduced a selection bias that could affect the interpretation of your results, and include any relevant discussion of such possible biases in your discussion.**

As a response to this comment, we have added additional information on the recruitment materials used in our study on page 18:

*In the advertisements for recruiting participants, we stated that our study would examine the extent to which non-scientists find psychological research comprehensible. Moreover, participants were informed that they would have to read summaries of English research articles on social and political psychology in our study.*

One main implication of the information provided in the recruitment materials regarding selection bias is that English language proficiency was probably quite high. This is already stated in the exploratory findings section (see page 41).

The information might, however, also result in a higher baseline level of interest in the topic at hand. This issue is now discussed in a footnote on page 42f:

*Furthermore, we forced participants to read specific research summaries in our study (instead of letting them choose articles based on personal preferences6). On the one hand, one might argue that observed differences in outcomes, for example, in full text access rates, might be even larger when individuals do an information search (e.g., on Google Scholar) out of their free will and on a self-selected topic with an information need of their own.*

*Footnote 6: This effect might, however, be counteracted by the fact that our participants knew beforehand that research on social and political psychology was the topic of our study. As a consequence, self-selection might have occurred regarding personal interest in the topic.*

We also added a new paragraph on this issue in our discussion of epistemic emotions and full text access rates on page 40:

*However, our participants knew beforehand that research on social and political psychology was the topic of our study, which is why self-selection might have occurred. This might also lead to more positive emotions in general and a higher amount of requested links (even though the proportion was still quite low) compared to a general population sample. Therefore, the extent to which these findings can be transferred to other populations remains an open question for future research.*

1. **For all of your tables, please add more information so that a reader could interpret the table without referencing the main text/methods. This includes providing more information about the scales for the responses (e.g., 1 to 8, etc.), and more explanation of what the variables mean (e.g., what is “semester” in Table 2). Please consider the point of view of something flipping through your paper who just wants to look at the tables – I realize you can’t convey all the relevant details of the method in the tables, but please try to make them as easy to understand as possible on their own. You may want to solicit the help of someone who isn’t familiar with the paper (e.g., have them read just the abstract and then look at the tables, and ask them what other information could be provided in the table to make them easier to interpret).**

Thank you very much for pointing out that there is a need for clarification here. We have revised all our Tables and Figures to make them more easy to understand for our readers.

For example, in Table 2 and Table 3, we included an additional column on the measurement of variables to give our readers a better idea of the meaning of these variables and the specific scores we reported. Moreover, additional notes on Table 2 are now provided.

e.g. regarding semester:

*Number of semesters participants studied with regard to the primary degree1 they currently pursued*

*1 102 students (61.45 %) were enrolled in Bachelor degree programs, 57 students (34.34 %) in master degree programs and 7 (4.22 %) in other study programs.*

In Table 4 and 5, we provide additional information on the scale of outcome variables in the corresponding Table notes:

*comprehensibility/credibility were measured on 1 to 8 semantic differentials and ability to evaluate/to make a decision on 1 to 8 Likert scales*

*all epistemic emotions were measured on 1 to 5 Likert scales*

And on the content of Table 4 and 5 in general in their captions.

E.g. Table 4

Old:

*Results of confirmatory analyses on H1, H3, H4, H5.*

New:

*Results of confirmatory analyses on comprehensibility (H1), credibility (H3), ability to evaluate (H4) and ability to make a decision (H5). Estimates are based on mixed models with (contrasts of) fixed effects for type of presented research summary (plain language summary with/without subheadings and ordinary scientific abstract) and random effects for participant and study.*

Table 6 seems to be the most complex table in our article. We added “OSA” in row 5 and 7 of the “added parameters” column to indicate more clearly which random slope is predicted here, but also provide now much more information on how to read this table in the corresponding notes (and row numbers that correspond to the following note):

*In column 1, information on the parameters that were tested in the corresponding LR test are provided. First, tests on the significance of random intercepts (row 1) and random slopes for both experimental conditions (i.e., with plain language summary with subheadings as a reference category) were conducted (row 2, 3). If these tests on random effects were significant, we conducted subsequent LR tests to examine if epistemic justification beliefs (row 4, 5) or English proficiency (row 6, 7) were able to explain variation in random intercepts (row 4, 6) or random slopes for ordinary scientific abstracts (row 5, 7). Since all LR tests on random slopes for the effect of plain language summaries without subheadings were non-significant (row 3), the corresponding tests with epistemic justification beliefs and English proficiency were not conducted and these rows are therefore omitted from this table.*

Additionally, the caption of Figure 2 provides now more information on its content:

*Overlapping curved arrows indicate that randomization took place for the order of blocks (curved arrows above blocks), the order of studies within blocks (curved arrows above studies) and the type of presented research summary (exemplary curved arrows within Study 1).*

1. **Please provide more information for readers about the size of the effects in the results section. Assume your readers do not know how to interpret the statistics you report, and provide some aid in understanding how practically meaningful the effects are. Please also incorporate a consideration of effect sizes in your own interpretation of the results – the statistical significance is useful for understanding how likely the result would have been if the null hypothesis is true, but a more comprehensive interpretation of the result would also include some consideration of practical significance/effect size. (Related to this, I wonder if most readers will know what a likelihood ratio test is – maybe consider giving readers extra information especially about this test and how to interpret its results?)**

We agree with you that information on standardized effects would be interesting for our readers. It is indeed a problem that no established standard procedure for computing standardized effects in mixed models exists (at least to our knowledge). This issue is comprehensively outlined in an excellent blogpost by Jacob Westfall (<http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/>). Based on this blogpost that outlines five options for computing effect sizes in mixed model designs, we concluded that the following option to which Jacob Westfall refers as *dR* (but strongly recommends not to use this term in publications) will serve our purposes best, and we included a statement on this on page 23 in the methods section:

*To facilitate the interpretation of our results, we computed standardized effect estimates. For this purpose, we divided effect estimates by the residual standard deviations at the text-level (i.e., the proportion of variance that cannot be explained by the fixed effects of research summary type or the random effects for studies and participants; see Westfall, 2016).*

We think that providing this information gives a reasonable idea of the practical significance of our findings in the results sections. The corresponding information is now included in Tables 4 and 5, as well as in the respective results sections.

Page 26

*As can be seen in Figure 3A, these differences were quite large (e.g., mean comprehensibility scores for plain language summaries with subheadings and scientific abstract differed by more than 0.600 residual standard deviations at the text-level, see Table 4).*

Page 28

*All effects on these measures were, however, considerably smaller (though still practically relevant) when compared to the corresponding effects that were obtained for perceived comprehensibility and ranged from 0.135 to 0.287 residual standard deviations (see Table 4).*

*The expected difference between plain language summaries without subheadings and ordinary scientific abstracts did not emerge in our study (see Table 4 and Figure 3). Corresponding effects were very small and not practically relevant (less than 0.040 residual standard deviations).*

Page 30

*In terms of residual standard deviations, the largest differences between plain language summaries and ordinary scientific abstracts were obtained for confusion (0.334 without subheadings and 0.571 with subheadings), whereas effects on boredom were considerably smaller (0.021 without subheadings and 0.173 with subheadings).*

Moreover, as a response to your second suggestion (see below), we included raincloud plots for all confirmatory outcomes in Figure 3. This will further support readers in interpreting our findings. The only exception is the knowledge acquisition test because we are confident that percentage scores on correct/incorrect answers are already quite indicative when evaluating practical significance.

Finally, we provide additional information on likelihood ratio tests on page 23f.

*For all outcome variables that were included in confirmatory analyses, we explored individual differences and whether these individual differences could be explained by justification beliefs and English language proficiency in two steps. To do this, we employed likelihood ratio (LR) tests. LR tests compare two nested models (e.g., models that include justification beliefs/English language proficiency as predictors or not) based on the ratio of their likelihoods. More specifically, they can be used to determine if additional parameters (e.g., random slope variances) significantly improve the model fit. The anova function in R provides an approximately chi-square distributed test statistic for LR tests. The degrees of freedom of this test statistic’s distribution are the number of parameters tested (i.e., in our case, the number of additional fixed effects or random effect [co]variances).*

1. **Related to this, I would like you to consider how your null results should be interpreted. Right now, you don’t say much about them, and I’m not sure if readers will interpret them as evidence of absence, or just absence of evidence. Can you provide more information about how you interpret the null results, and what you think they can or can’t tell you about evidence of absence? This comes up in a few places.**

First, we hope that Figure 3 which we included based on your second suggestion is helpful in this regard. We agree that more guidance may be needed here. Since our sample size calculation was based on a Cohen’s *d* of .5 (medium-sized effect), we included the following paragraph on page 42:

*How should the non-significant findings in our study be interpreted? Since our confirmatory hypotheses were preregistered and the power calculation for these hypotheses was based on a Cohen’s d of .5, we are reasonably sure that we did not fail to find any medium-sized effects (at text-level). Especially evidence on the absence of effects for differences between the plain language summaries without subheadings and the ordinary scientific abstracts condition (H3b-H5b) seems quite strong. On three related outcomes, all findings were clearly non-significant and the data (distribution) on these outcomes appears very similar (see Figure 3). Notwithstanding, given that we only report evidence of one single study, we cannot fully rule out that (possibly smaller) effects may exist at the population-level, which we were unable to detect. We are less confident regarding the robustness of our non-significant (and significant) exploratory findings on participant-level covariates (i.e., justification beliefs and English language proficiency) as evidence of absence. This is due to the fact that the calculations of our power analyses did not explicitly target such higher-level effects in mixed models. Therefore, effects might be estimated with less precision on this level and should be interpreted with great caution.*

1. **On page 35, you say that the results “illustrate that plain language summaries…” – this is in the context of describing the exploratory findings. I think “illustrate that” is too strong for the exploratory results.**

Yes, we agree. We changed the wording accordingly.

Old:

*These differences in full text access illustrate that plain language summaries are not only of theoretical relevance in the sense of affecting response patterns in self-report measures, but also of practical relevance as they influence how individuals deal with and access research findings.*

New:

*Nonetheless, these differences in full text access might point towards the practical relevance of providing plain language summaries. Providing plain language summaries did not only influence response patterns in self-report measures, but also how individuals dealt with and accessed research findings.*

1. **One aspect of your study that I found interesting to think about is that this is indeed an experiment, but because you used actual original abstracts and plain language statements, you did not have experimental control over the differences between these two conditions. This challenged me to think about whether this could introduce “confounds” and I came to the conclusion that “confounds” is the wrong way to think about this – the better way to think about is that the independent variable is not whether the summary presented is in scientific vs. plain language, but you manipulated whether readers saw the authors’ original abstract vs. plain language summary, with all of the other changes this entails. The fact that you used actual abstracts/summaries is a positive feature, and it also changes how we should think about the specific independent variable being manipulated – whatever else changes systematically between authors’ original abstracts and plain language summaries (at this journal, and more generally) was also different between the conditions here. That means we can’t be sure exactly what the mechanism is that is driving the effect, but we can be more confident that it translates to actual original abstracts vs. plain language summaries. Given how central this is to understanding what conclusions can be drawn from your study, and what the conceptual independent variable is in your study, I would strongly recommend that you include a paragraph or two discussing this issue. I could imagine it would be useful to have a paragraph around the place where you describe the study/hypotheses, and then another paragraph in the discussion revisiting this issue.**

Thank you for this meaningful and elaborate analysis of the mechanisms at work here. We have added a paragraph when introducing our hypotheses on page 9:

*It is essential to consider in this context that plain language summaries without subheadings can directly be derived from plain language summaries with subheadings by excluding existing subheadings (see Method). This means that these conditions do indeed only differ in this one specific aspect (i.e., the presence of subheadings). When it comes to comparisons between scientific abstracts and both types of plain language summaries, the types of presented research summaries differ fundamentally with regard to many aspects, not just subheadings or use of technical terms. These aspects (e.g., usage of statistics or provision of background information) are thus not experimentally varied or controlled. As a consequence, (only) the type [italics] of presented research summary (of the same scientific publication) is varied here and not one specific aspect of the research summary.*

And in the discussion of our results on page 36f:

*The practical relevance of these findings is quite high since we did not use artificial materials to illustrate the benefits of plain language summaries but employed published (and thus, “real”) plain language summaries and scientific abstracts. These ‘gains’ in the ecological validity and practical significance of our findings are, however, at the expense of a less strict control of differences between plain language summaries and scientific abstracts. We do not know exactly what differs between those conditions—or what drives this effect—as the authors wrote both types of research summaries themselves based on a short guidance provided by JSPP. Adherence to such guidance has been shown to be rather low (see Kadic et al., 2016) and different authors might differ fundamentally in their understanding of how to write plain language summaries. Differences between plain language summaries and scientific abstracts may have existed, for example, in linguistic characteristics (e.g., usage of technical terms), formal characteristics (i.e., sentence length) or content (e.g., a more comprehensive introduction to the background of the research question at hand or qualitative descriptions of statistical results). Analyzing in what way exactly published plain language summaries and scientific abstracts differ regarding these aspects is an intriguing research question of its own that was, however, beyond the scope of this study.*

1. **I was very glad to see that you replaced the PLS acronym with the full phrase. I just found one place where you still use the acronym (p. 5) - please replace it and check and see if there are others.**

Thank you very much. We replaced the acronym with the full phrase here and carefully checked the manuscript for any other occurrences.

**Suggestions:**

1. **On page 2, you state that plain language summaries "may be seen as a safeguard against the sometimes misleading or distorting press releases from universities or other research institutions". However, my understanding (and my experience) is that much of the exaggeration in press releases actually comes from the authors themselves (the people who write the press releases typically interview the authors, who provide the exaggerated claims). Perhaps more importantly, I don't think you need this point, and including it (especially in the very first paragraph) may distract from the actual focus of the paper, since you do not actually compare plain language summaries to press releases. I would suggest removing this point.**

Thank you for providing this interesting input on the topic of press releases. We will keep this in mind for our future research on plain language summaries. Consequently, we removed this point.

1. **I wonder if adding some data visualizations would help (see point 4 above). In some cases, it is hard to grasp the meaning of the result, and the uncertainty/variability. Data visualizations that include individual data points (e.g., raincloud plots, or something like that) might make it easier for readers to figure out how practically significant they think the results are. This could also be useful for null results (point 5 above).**

Thank you very much for this suggestion. Raincloud plots seem indeed to be a very informative approach to data visualization. As mentioned above in response, we have now included raincloud plots for all continuous confirmatory outcomes in a new Figure (Figure 3). These plots are based on residual scores after controlling for random effects of participants and study.

Caption of Figure 3:

*Raincloud plots for continuous confirmatory outcomes. Residual scores for comprehensibility (A), credibility (B), ability to evaluate (C) and ability to make a decision (D) are depicted separated by experimental condition (PLS = plain language summary). Residual scores were obtained from a mixed model that controlled for study and participant as random factors.*

1. **In my opinion, you are over-interpreting the results of some of the exploratory analyses. For example, the results presented in the section on “The role of justification beliefs and English proficiency” are all quite similar to one another, but some cross the p < .05 threshold and others don’t. In my view, interpreting those that cross the threshold differently from those that don’t doesn’t make a lot of sense, particularly in the context of exploratory (i.e., non-pre-registered analyses), and given that the effect sizes are not that different for the significant vs. non-significant results. In addition, given that you ran a number of exploratory tests, it seems unlikely to me that the significant results would survive a correction for multiple comparisons. This isn’t the most important factor, and of course interpreting the results of exploratory tests is a subjective exercise, so I will leave it up to you to decide how you want to interpret these, but for what it’s worth I didn’t agree with your interpretation. Another relevant factor to consider for many of the exploratory results is that your power for detecting between-person effects is probably not very high, so null results may not be very meaningful, and all effects are likely to be imprecisely estimated (and therefore small-to-medium differences between effects are probably not reliable).**

This is a fair point. We have revised the corresponding paragraph and we now report also non-significant regression coefficients for significant LR tests in the text to allow our readers to draw their own conclusions based on the data. Additionally, we chose a more neutral wording for significant effects.

p. 32f

*More specifically, for plain language summaries with and without subheadings, effects of justification by authority (b = .14, p = .072) and justification by multiple sources (b = .15, p = .053) (closely) failed to reach significance, whereas the effect of beliefs in personal justification ~~resulted in a lower~~ on perceived comprehensibility was significant and negative (b = -.18, p = .020). Interestingly, these effects reversed (at least on a descriptive level) for ordinary scientific abstracts, where a significant interaction was found for justification by multiple sources (b = -.16, p = .035), but not for personal justification (b = .14, p = .080) or justification by authority (b = -.11, p =.173). Since all these effects were of similar size and power issues likely exist for exploratory analyses on reader characteristics (see Discussion), we strongly advocate against overinterpreting the statistical significance of findings here.*

p. 33f

*More specifically, ~~stronger~~ beliefs in justification by authority ~~resulted~~ had significant effects ~~in overall higher scores~~ on all three outcomes (b = .30 - .34, all p < .01). Effects of personal justification (b = -.13 - .09, p = .09 - .39) and justification by multiple sources (b = -.03 - .0002, p = .77 - .98) were non-significant.*

Moreover, we state now more clearly which non-significant findings relate to LR tests (which is why no fixed effects are reported for these predictors) on page 33:

*While justification beliefs were not able to predict differences in knowledge acquisition (as indicated by a non-significant LR test, see Table 6),…*

Additionally, we revised or wording when interpreting the results of exploratory analyses in the discussion on page 41:

Old:

*Our results therefore indicate that English language proficiency may constitute a main obstacle for comprehending any kind of research summary in non-native speakers.*

New:

*Although these findings are exploratory and should therefore be interpreted with adequate caution, they point towards English language proficiency as a major obstacle for comprehending any kind of research summary in non-native speakers.*

Finally, we point towards potential power issues regarding these covariates in our discussion (see our response to comment 5) and made additional changes regarding our wording on exploratory analyses as a response to comment 6.

1. **“we aim to close this knowledge gap” on page 12 sounds a bit bold – maybe tone it down?**

Yes, this wording seems to be indeed somewhat inappropriate—especially for an exploratory outcome. We revised the corresponding sentence on (now) page 13:

Old

*[...]we aim to close this knowledge gap in the present study by means of exploratory analyses.*

New

*[...]we aim to examine this type of behavioral consequence by means of exploratory analyses.*

1. **In your sample size calculation, it might be good to say what “VPC” means (p. 23)**

Thank you very much for pointing this out. We do now provide a brief definition of this term in Footnote 5:

*The acronym VPC is short for variance partitioning coefficient and denotes “the relative magnitude of the estimable variance components” (Judd, Westfall & Kenny, 2017, p. 18).*

1. **I didn’t understand the sentence beginning “To put it bluntly, individuals…” on page 34.**

Thank you very much for this feedback. As you note in your next suggestion, our discussion was quite long before this revision. Since it has been considerably extended in this revision and we deem these extensions to be more important than the argument we made here on page 34, we removed the corresponding paragraph.

1. **The discussion is quite long and a bit repetitive, you may want to consider shortening it.**

As a response to this suggestion, we removed the paragraph on the interplay of text complexity and text structure on page 34 (see our response to the previous suggestion).

Moreover, we shortened other parts of the discussion. For example, by removing the following statement:

*we also gained some interesting insights on how text structure might also contribute to perceived ‘text easiness’.*

And by removing two paragraphs on (formerly) page 34 and 35 because they were a bit repetitive and mainly based on exploratory findings (which might be less robust, see above):

*~~Some of our findings [..] to larger audiences.~~*

*~~From a theoretical perspective [...] more strongly.~~*

**In addition, I made a few notes on grammatical issues that I will share with you here, but I want to reiterate that overall I found the writing very clear.**

Thank you very much. We apologize for these grammatical errors. We have now corrected them.

**-p. 3: “test on the basic notion” should be “test of the basic notion”  
-p. 4: “whether beliefs on the justification” should be “whether beliefs about the justification”  
-p. 6: “and especially experimental research” – this sentence is quite long. I would cut “and especially” and start a new sentence with “Experimental research…”  
-p. 7: “our knowledge on how plain” should be “our knowledge of how plain”  
-p. 7: “public health has already” should be “public health have already”  
-p. 8: “agree more often to corresponding” should be “agree more often with corresponding”  
-p. 10: “little is known on how” should be “little is known about how”  
-p. 28: “more curios” should be “more curious”  
-p. 34: I think that “and ordinary scientific abstracts (rather short but less comprehensible texts)” should be “compared to ordinary scientific abstracts (rather short but less comprehensible texts)”**

**In summary, I think this is a promising manuscript and I hope you will submit a revision to Collabra: Psychology. I look forward to receiving your revision.**

**Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.**

**If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.**

**We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.**

**Sincerely,**

**Simine Vazire  
Editor in Chief  
Collabra: Psychology**

**Editor Final Decision—Accept**

Dec 7, 2020

Dear Dr. Kerwer,

I have now had a chance to read over your revised manuscript “Straight from the scientist’s mouth - Plain language summaries promote laypeople’s comprehension and knowledge acquisition when reading about individual research findings in psychology”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns I raised - I really appreciate how thorough your revisions and explanations were. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely, Simine Vazire Editor in Chief Collabra: Psychology