Revision Summary

Thank you and the reviewers for your constructive feedback regarding our manuscript. In accordance with the recommendations of the reviewers, we have substantially revised our manuscript. In particular, we have bolstered our review of past research on content-related validity evidence in order to make the specific contributions of the machine learning procedure clearer, as well as the implications of this kind of procedure for educational assessment. Below we summarize our revisions. In addition to the revisions that we outlined, we carefully edited and made revisions to the manuscript in accordance with reviewers' comments and recommendations.

Literature Review

- The introduction of the manuscript has been substantially revised with the focus specifically on validity evidence based on test content, in accordance with the recommendations from Reviewer 3.
- When introducing topic modeling, we have re-framed the discussion around previous research by specifically citing applied examples.
- An example of how the method could be applied diagnostically has been added.

Method

- We have clarified the use of document-term matrices to make it clear that a single document-term matrix was used for the content standards, which was used for the model training and that we then created a document-term matrix for the items to evaluate their relation to the estimated topics. Specifically, we clarified that a new document-term matrix would need to be calculated for the evaluation of any new test items.
- We have clarified that each of the criteria used to evaluate competing numbers of topics are specific measures of model fit.
- We have clarified that subject matter experts arrived on the five-topic solution independently, but did also have a de-briefing period to discuss alternative topic model solutions.
- We note that Webb's depth of knowledge verbs were eliminated in an effort to keep the topics concentrated on content-related vocabulary. We specifically mention that these words were included in preliminary models, but that their inclusion obscured substantive interpretation of topics.
- The paragraph describing the revisions made to Oregon's alternate assessment was removed.
- We have noted that we evaluated both field test and operational items because we wanted to flag those that did not appear to be generated by any topic for further evaluation.
- We re-organized the sub-sections of this section and added clarifying statements about the contents of the sub-sections according to the recommendations of Reviewer 3.

Results

- The y-axis on Figure 1 is now included.
- We have included further discussion on words that appear across multiple topics, specifically mentioning that these terms generally are representative of the cross-cutting concepts in science represented in the NGSS standards. While their removal may result in more distinct topics, it would also come at the cost of not representing the standards as well. We specifically provide an example with the term *evidence*.
- Gamma values, which essentially represent the weight of a given topic for a given document, have been added to all non-zero cells in Figure 2.

Discussion

- We now explicitly mention the possibility that alternative subject matter experts may have arrived at a different optimal number of topics, and use this as further evidence of the need for multiple views to help crowdsource a solution and reach a consensus model. We also mention in the same section, however, that general training may need to first be provided so the subject matter experts understand the tool they are helping to develop, its purpose, and how their input may influence the final model.
- We have included a new paragraph discussing that the extent to which our method would correlate with judgments from human raters is not yet clear, and that future research should address this gap.
- We have extended our discussion on mapping of individual items to topics to help address the "so what" question by specifically mentioning that it may be useful to flag items that do not align with any topic, and that evaluating topic representation across items can be useful. However, this is balanced by the information learned by an individual item mapping onto a specific topic being lower than what is obtained in alignment studies, because the mapping is to a sub-domain level rather than to an individual skill.
- We have clarified that if the results of topic modeling and alignment studies do not converge additional data may need to be collected, "perhaps with an additional subject-matter expert".
- When discussing crowdsourcing a model, we clarify that crowdsourcing may be beneficial for the topic model estimated from the text within the NGSS. The text used from the Oregon alternate assessment used here was an example of an evaluation of a test against this model, but it is the model that could be benefit from additional refinement through crowdsourcing. And test written to measure the NGSS could then be evaluated against this model.
- We have expanded our discussion on the application of this method to mathematics, emphasizing that keywords in mathematics are still important and that mathematical symbols could be included as words. These symbols could therefore be important "keywords" in identifying if the item is representative of the standards. However, it is also noted that this is mostly speculative, and that further research would be needed to understand how well it actually works.