






Evaluating Content-Related Validity Evidence Using a Text-Based Machine Learning Procedure

Daniel Anderson  and Brock Rowley , University of Oregon,
Sondra Stegenga,  University of Utah, P. Shawn Irvin,  University of Oregon, and
Joshua M. Rosenberg,  University of Tennessee

Validity evidence based on test content is critical to meaningful interpretation of test scores. Within high-stakes testing and accountability frameworks, content-related validity evidence is typically gathered via alignment studies, with panels of experts providing qualitative judgments on the degree to which test items align with the representative content standards. Various summary statistics are then calculated (e.g., categorical concurrence, balance of representation) to aid in decision-making. In this paper, we propose an alternative approach for gathering content-related validity evidence that capitalizes on the overlap in vocabulary used in test items and the corresponding content standards, which we define as textual congruence. We use a text-based, machine learning model, specifically topic modeling, to identify clusters of related content within the standards. This model then serves as the basis from which items are evaluated. We illustrate our method by building a model from the Next Generation Science Standards, with textual congruence evaluated against items within the Oregon statewide alternate assessment. We discuss the utility of this approach as a source of triangulating and diagnostic information and show how visualizations can be used to evaluate the overall coverage of the content standards across the test items.

Keywords: machine learning, text-mining, textual congruence, validity

Validity evidence based on test content is a critical component of the “overall evaluative judgment” (Messick, 1995, p. 741) of the validity of test scores for a given use, and is one of the five major sources of validity evidence outlined by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Content-related validity evidence is generally defined by the definition, representation, and relevance of the domain, as well as the appropriateness of the test development process (see Sireci, 1998). In large-scale statewide accountability testing, the domain definition is largely provided by the adopted statewide content standards, which are distilled into a set of test specifications. Domain representation and relevance address similar issues, with the former indicating the degree to which items within the domain are represented by the test items, including areas of under- or overrepresentation, and the latter representing the degree to which a given item is relevant to the domain. Finally, evidence of content-related validity can be strengthened by specific elements of the test development process, including interim reviews of items by content, measurement, and bias/sensitivity experts (Sireci & Faulkner-Bond, 2014).

In this paper, we illustrate a method for obtaining content-related validity evidence using a text-based machine learn-

ing model trained on (i.e., estimated from) a set of content standards. The method can be used both diagnostically during test development, and as a source of triangulating and confirmatory evidence along with ratings from subject matter experts (SMEs). The method primarily addresses domain representation and relevance, and we illustrate how visual displays can aid in the interpretation of findings. If used diagnostically, the method can also provide evidence for developmental processes. In what follows, we first discuss empirical methods for gathering content-related validity evidence before introducing the specifics of our proposed method. We then illustrate our approach using a statewide alternate assessment for students with significant cognitive disabilities.

Validity Evidence Based on Test Content

Empirical evaluations of content-related validity evidence for statewide accountability tests generally focus on domain representation and relevance through alignment studies. Alignment studies typically include panels of SMEs judging the alignment between the content represented within the test items and the content represented in the corresponding standards. Numerous alignment study methodologies have been proposed, including the *Achieve* model (Rothman, Slattery, Vranek, & Resnick, 2002), the *Surveys of Enacted*

Curriculum (SEC; Porter & Smithson, 2001), and perhaps most prominent, the *Webb* model (Webb, 1997). For a review of these models, see Sireci and Faulkner-Bond (2014). While each model differs from the others to some degree (particularly the SEC model, which also considers the curriculum as part of the overall systems-level alignment), they are fundamentally similar in that each model includes panels of SMEs evaluating items across a number of dimensions with respect to the content standards. Analyses of the resulting data are then used to provide both item- and test-level information (i.e., domain representation).

Alignment studies are generally difficult to design and execute. Bhola, Impara, and Buckendahl (2003), for example, outline several challenges, including specificity in alignment criteria, training, and representation of items within each performance classification (e.g., *Did Not Meet*, *Nearly Meets*, *Meets*, or *Exceeds*). Anderson, Irvin, Alonzo, and Tindal (2015) further discuss challenges related to group dynamics, practical costs, and methods for aggregating judgments across raters (i.e., group consensus, averages, or, as the authors propose, latent trait methods). Despite these challenges, alignment studies provide a critical source of content-related validity evidence for large-scale testing programs and help to ensure that the tested content matches the content that is intended to be taught in schools. SME judgment remains the fundamental means by which items are evaluated, and with good reason—features of the item judged holistically may be missed by evaluating only specific features of the item. Yet, while SME judgment remains the gold standard, other sources of information can still add to the evidentiary base supporting content-related validity.

We operate from a theoretical framework that assumes the domain representativeness of a test, and to a lesser extent, domain relevance can be estimated by evaluating the overlap in the text used in both the test items and the content standards the test was designed to measure. In other words, subdomains within the content standards are assumed to be defined by specific keywords. If we can identify these keywords, and which keywords are most strongly associated with each subdomain, we can investigate the relative representation of these keywords across the items in the test. We refer to this evaluation as an investigation of the *textual congruence* between the set of test items and the corresponding content standards. We evaluate textual congruence through the application of a text-based machine learning model, specifically *topic modeling*, estimated through latent Dirichlet allocation (LDA).

A Text-Based Machine Learning Approach

Topic modeling is a probabilistic machine learning model for identifying latent topics (i.e., themes, subdomains) in text. Its history stems from qualitative content analysis and latent semantic analysis (Mohr & Bogdanov, 2013). Rather than the researcher predetermining the topics to be analyzed and coded, however, the topics emerge from a text corpus based on the frequency of word co-occurrence (i.e., text-based correlations). Topic modeling has advanced the field of text analysis from identifying specified words through a deductive approach, where topics are preidentified, to a more inductive approach where meaning is allowed to emerge (Mohr & Bogdanov, 2013).

Topic modeling is still relatively new in text-based analytic research. Blei, Ng, and Jordan (2003) introduced LDA in 2003, which is now a widely used estimation procedure. Prior to LDA, inductive or latent themes in text were achievable primarily through qualitative analysis (Nikolenko, Koltcov, & Koltsova, 2017). Topic modeling, however, provides the potential to conduct text-based investigations at a scale. Quinn, Monroe, Colaresi, Crespino, and Radev (2010), for example, analyzed text data transcribed from over 118,000 political speeches from 1997 to 2004 to investigate the relative attention given to specific topics in political discourse in the United States. Similarly, Jelveh, Kogut, and Naidu (2018) estimated a topic model on text from over 80,000 economics papers. Estimates of the political ideology of the manuscript authors were then linked with these data to evaluate if their political ideology related to their academic writing, finding “a robust correlation between patterns of academic writing and political behavior” (p. 29). The scale of these investigations necessitated an automated procedure, as manual coding would be infeasible.

Topic models are estimated from a *document-term matrix*, in which the columns represent each unique word from the text corpus, the rows represent each “document” evaluated, and the cells represent the counts of each word within each document. The document-term matrix is used to estimate a set of latent variables, or topics, representing themes within the text based on patterns of word co-occurrence across documents. The documents represent discretized instances of the overall corpus, such as blog posts or newspaper articles. Quinn et al. (2010), for example, treated the text from each political speech as a document, while Jelveh et al. (2018) treated each economic paper as a document. LDA is guided by the principle that “every document is a mixture of topics...[and] every topic is a mixture of words” (Silge & Robinson, 2017, p. 90). For example, when evaluating political ideologies, Jelveh et al. (2018) estimated a set of topics from all the words represented across the economic papers, but then inspected the relative distribution of these topics across papers. The relative proportions of topics within a document were then linked back to the authors and used to predict their political leanings (estimated through campaign contributions and petition signings). Topic modeling has also shown potential to produce similar results to qualitative frameworks, such as grounded theory (Nikolenko et al., 2017), providing support for its use in expanded applications (for additional examples, see Boyd-Graber, Hu, & Mimno, 2017; Liu, Tang, Dong, Yao, & Zhou, 2016).

We explore the use of topic modeling with LDA estimation to evaluate the textual congruence between test items and content standards as a source of content-related validity evidence. Specifically, we train a topic model on the middle school *Next Generation Science Standards* (NGSS) to uncover latent topics/subdomains. Topic model estimation requires a number of decision points, most importantly the number of topics to be estimated. We discuss the specifics of our model in detail in the Method section. Once the model is built, the probability that *any* given text was generated by each topic can be estimated. The methodology therefore includes two steps: (a) estimating the latent topics/subdomains represented within the content standards, and (b) using this model to generate item-level predictions to topics (i.e., the probability that the text within a given item was generated by each of the topics). Importantly, the model is built directly

and solely from the content standards. While a certain degree of judgment is involved in the creation of the model (as explained in more detail in the Method section), it is independent of any specific test and, once built, could be applied to any test addressing those standards. Our specific topic model was developed with the guidance of two science SMEs, but potentially hundreds or even thousands of SMEs could theoretically be involved in the refinement of the model through crowdsourcing methods (see Arganda-Carreras et al., 2015; Bentzien, Muegge, Hamner, & Thompson, 2013) leading to a consensus model developed with input from the field.

Once built, the topic model can be used to estimate the textual congruence of test items (from any test) and the content standards through model-based predictions. That is, the text within an item (item stem and answer options) is evaluated against the topic model, and the probability that the text was generated by each topic is estimated. Note that these predictions are at the topic level, rather than the individual standard level, and thus provide a coarser and fundamentally different representation of content-related validity evidence than item-standard alignment studies (i.e., textual congruence at the topic level would be unlikely to provide adequate evidence for high-stakes peer review processes in accountability testing; United States Department of Education, 2018). Once the item-level predictions have been made, however, the relative representation of topics across items can be evaluated, along with the mapping of individual items to topics. These analyses may provide triangulating information with SME judgments, but perhaps more importantly, provide a source of diagnostic information during test and item development. For example, if the text represented within the item was not represented within any of the modeled topics, no prediction would be made and the probability would be distributed equally across all topics. These items could be flagged for further review, which may identify out-of-scope items before reaching a full alignment review (although it is also possible that the item may represent the standards in ways outside of the text). Similarly, the relative representation of topics across items could be evaluated, with the test refined to ensure a roughly equal inclusion of all modeled topics to adequately represent the overall standards.

Method

We evaluate the textual congruence between the middle school (grades 6–8) *NGSS* and the Grade 8 Alternate Assessment based on Alternate Achievement Standards (AA-AAS) in Oregon, as described more fully below. We also evaluate the correspondence between results obtained from our text-based model versus SMEs, and thus, briefly describe the source of these data. The text-based analysis addresses a fundamentally different aspect of content-related validity evidence than alignment studies, given that congruence is evaluated between items and topics (which are composed of multiple standards), rather than between items and standards. However, it is worth considering the extent to which the methodologies generally do, or do not, arrive upon similar substantive conclusions. In what follows, we detail how we evaluated the number of topics to be extracted and how the results can be used to evaluate the overall textual congruence between the test items and the content standards. We also detail how we linked our topic model data with alignment study results. Note that our specific application has some unique

aspects related to the measure itself, but the methodology should readily extend to any situation in which an evaluation between a set of items and a set of content standards is needed.

Data Sources

The *NGSS* are based, in part, on the *Framework for K-12 Science Education* (National Research Council, 2012) and are built around performance expectations for each of grades K-5 and grade bands 6–8 (middle school) and 9–12 (high school). Performance expectations are situated within one of four domains—Physical, Life, Earth/Space, and Engineering Design. Performance expectations reflect the three dimensions of science learning described in the *Framework*: understanding and applying *science and engineering practices* as they master *disciplinary core ideas* (which are specific to a science content area) and *crosscutting concepts* (which span science broadly, National Research Council, 2012).

Two document-term matrices were created, with one representing the content standards (with each *NGSS* Performance Expectation serving as a document) and one representing the text from items (item stems and answer options). The document-term matrix for the content standards was used to train (estimate) the topic model, as described more fully below, while the document-term matrix for items was used to evaluate the textual congruence between the items and the estimated topics using the fitted model. The document-term matrix for items was unique to this specific test, and any evaluation of other tests would require the creation of a new item-level document-term matrix; however, the document-term matrix for the content standards could be applied to any test.

As part of data processing, we removed stop words (common words like “of,” “a,” “the,” “and,” “is”) as represented within the *onix*, *SMART*, and *snowball* dictionaries (see Lewis, Yang, Rose, & Li, 2004; Onix, 2018; Snowball, 2018) and implemented in the *tidytext* R package (Silge & Robinson, 2016). Additionally, we removed verbs associated with Webb’s depth of knowledge levels (Webb, 2002), given their proliferation throughout the content standards. Although preliminary models included Webb’s words, the topics were difficult to identify and interpret because of the overrepresentation of these words within in each topic, rather than words relating to scientific concepts (e.g., evidence, genetic, mass, motion). We therefore opted to keep the model as focused on the content as possible by removing these process words (e.g., design, describe, interpret, analyze).

The final document-term matrix included 59 rows, one for each *NGSS* Performance Expectation, and 355 columns, one for each unique word represented in the content standards that was not a stop word or one of Webb’s depth of knowledge verbs. A similar document-term matrix was preprocessed using the same steps as listed above for the test items, using all text represented within the test items and each item serving as a document. Importantly, however, no analyses were conducted on the document-term matrix for items. Rather, the document-term matrix for items was used only to make predictions for each item to the topics derived from the content standards.

The alignment data were collected as part of the technical adequacy evidence for the United States Department of Education’s peer review of assessments process (United States Department of Education, 2018). Raters judged the

linkage between items on Oregon's alternate assessment and the corresponding statewide content standards on a three-point (0–2) Likert scale. All standards had been reduced in depth, breadth, and complexity as part of an “essentialization” process to correspond with the alternate achievement standards. All items were evaluated by four raters. For additional information on the alignment study, see the statewide technical manual published by the Oregon Department of Education (2017).

Measures

Our application utilized the science portion of the Grade 8 statewide AA-AAS in Oregon, designed for students with the most significant cognitive disabilities (SWSCDs; United States Department of Education, 2005). The AA-AAS is an alternate assessment to the statewide general assessment, with up to 1% of students eligible for reporting purposes. Any student with an individualized education program (IEP) is eligible to participate, and the recommendation for the most appropriate assessment is guided by the IEP team. SWSCDs who take the AA-AAS commonly score two or more standard deviations below the mean on standardized intelligence tests and have commensurate deficits in adaptive behavior. The disability must significantly impact students' learning and their ability to generalize learning across settings. SWSCDs require highly specialized services related to both their education, and often, social and medical needs.

The Grade 8 AA-AAS in Oregon included 48 items, with 36 used in operational reporting for accountability and 12 used in ongoing field-testing efforts to revise and improve the test annually. Text was included in our analyses from all 48 items. We included both operational and field test items because we wanted to apply the algorithm diagnostically. Items found to not have any overlap in words from any of the generated topics could be flagged for further evaluation, and we wanted to flag both operational and in-development items that had this characteristic. The Oregon AA-AAS is individually administered, with a qualified assessor providing the assessment with IEP-designated supports, and scored using a standardized protocol, with all responses scored dichotomously (correct/incorrect). Students responded to each item through a set of student materials using response modalities commensurate with their IEP and abilities (e.g., verbal, nonverbal, gesturing).

Procedure and Analyses

Our process included (a) using topic modeling to identify key subdomains/topics within the statewide content standards based on frequency of word cooccurrence; (b) applying the model to new text, in the form of words represented in the test items through either the stem/prompt or the answer options; and (c) evaluating the overall mapping of items within the test to the modeled subdomains/topics, including the relative representativeness of each subdomain/topic within the test. The number of topics was determined through a combination of statistical analyses and SME judgment.

Topic modeling. Topic modeling is akin to exploratory factor analysis (EFA), where latent variables (topics) are estimated based on the probability that the words within the topic will co-occur (see Mohr & Bogdanov, 2013). As men-

tioned above, LDA, introduced by Blei et al. (2003), is perhaps the most common estimation procedure for topic models and was the approach used here. LDA simultaneously estimates both the mixture of topics within a document and the mixture of words within a topic. The β matrix reports the estimated probability that each word belongs to (or was generated by) a given topic, while the γ matrix reports on the probability that each topic is represented within a given document. For example, in a two-topic solution, hypothetical Performance Expectation 1 may be composed of 25% Topic 1 and 75% Topic 2, while hypothetical Performance Expectation 2 is composed of 98% Topic 1 and 2% Topic 2. Each topic is then represented by the corpora of words, with each word having a different modeled probability of having been generated by the corresponding topic. Post-hoc substantive labels are generally assigned to topics through expert evaluation of the top n words within each topic (typically 10–20 words), based on the β matrix.

As with EFA, perhaps, the most difficult aspect of topic modeling is determining the number of topics (latent factors) to extract, which must be determined *a priori*. Models with different numbers of topics can provide different results and different conclusions about the underlying text. In our application, we relied on a combination of statistical evidence with SME judgment. From a statistical view, we relied upon four measures of model fit, as delineated by Arun, Suresh, Madhavan, and Murthy (2010), Cao, Xia, Li, Zhang, and Tang (2009), Deveaud, SanJuan, and Bellot (2014), and Griffiths and Steyvers (2004). Each of these are *relative* indicators of model fit (similar to information criteria) and the “best” model can only be determined by comparing competing models. Briefly, the method outlined by Arun et al. (2010) is based on the KL-Divergence of two salient distributions, viewing LDA as a matrix factorization procedure, with the goal of minimizing this value. The Cao et al. (2009) method relies on topic density through average cosine similarity (minimized), while the Deveaud et al. (2014) method uses a similar approach but relies on the Jensen-Shannon distance between topic distributions (maximized). Finally, the method proposed by Griffiths and Steyvers (2004) uses Gibbs sampling with the posterior sampled such that the harmonic mean of the sampled log-likelihoods is maximized. See Hou-Liu (2018) for a discussion on the performance of these various indicators. We evaluated models with 2–25 topics for each model fit indicator.

Following the evaluation of topics, we found a range of topics that appeared to reasonably minimize or maximize each of the criteria (as displayed in the Results section). These topics were then reviewed for substantive meaning and a final topic model was arrived upon through independent evaluations of the topic solutions by two science content SMEs. This model then represented our final trained model. The probability that each AA-AAS test item was represented by each topic was then estimated. The model was therefore trained on the words within the standards and we evaluated whether the words used in the items corresponded with the identified topics.

To evaluate the correspondence between the topic model results and SME ratings, we first estimated the mean SME rating for each item. Topics were then linked to these data via the standard the item was designed to measure, providing a link between SME ratings and the modeled topics (with the topic to standard link identified by the highest γ value for

a given standard across topics). Topic predictions were then made for each item, which estimated the probability that the text from the given item was generated by each topic. Finally, these data were linked with the SME ratings via the item ID and the topic, resulting in a final data set that included an item ID, the standard the item was designed to measure and its associated topic, as well as the mean SME alignment rating and the probability the item was generated by the given topic. Given that the standard the item was judged as aligning with was used to link standards and topics, all topic-level predictions corresponded to the given standards. In other words, it would not be possible for a model-based prediction to be made for a given item to a topic not including the standard for the given item.

We anticipated that multiple items would not include any text related to the modeled topics, leading to no topic-level prediction for the given item (i.e., equal/uniform topic probability). We were therefore primarily interested in the correspondence between model-based predictions and SME ratings when a prediction was made. However, we also hypothesized that when a topic-level prediction could not be made (given the text represented in the item), these items would have lower alignment ratings, on average, than those in which a topic prediction could be made. We evaluated this relation using a logistic regression model, with the alignment rating (collapsed from a three-point scale to a dichotomous yes/no rating) serving as the outcome, and a dichotomous indicator of whether or not a topic-level prediction was made by the model serving as the primary predictor variable of interest. We also included a set of fixed effects to control for the specific rater providing the rating (i.e., the outcome included all ratings across raters, rather than the mean rating, and we therefore included *rater* as a set of control variables). The three-point alignment scale was collapsed with a rating of 2 (on the 0–2 scale) representing alignment, and ratings of 0 or 1 representing nonalignment.

Topics were estimated using the *textmodeling* package (Grün & Hornik, 2011), while the evaluation of the number of topics to extract was conducted with the *ldatuning* package (Nikita, 2016), both of which are extensions to the R statistical computing environment (R Core Team, 2018). Data were prepared using the *tidyverse* suite of packages (Wickham, 2017), with all plots produced using the *ggplot2* package (Wickham, 2016).

Results

We first discuss the selection of the optimal number of topics, followed by the mapping of topics to standards and words to topics. We then present the results of the trained model to the items within the Grade 8 Science portion of the AA-AAS in Oregon and specifically delineate by whether items were written to be of *low*, *medium*, or *high* difficulty. Note that this delineation was based on item writing practices and theoretical characteristics relating to item difficulty, rather than empirical investigations of item difficulty. Empirically, items written to be of *low* difficulty were generally easier than those written to be of *medium* difficulty, which were generally easier than those written to be of *high* difficulty, although there was some overlap between categories. We were interested in evaluating results between these categories because of the differences in item writing practices (i.e., higher difficulty items generally included more content-specific vocabulary).

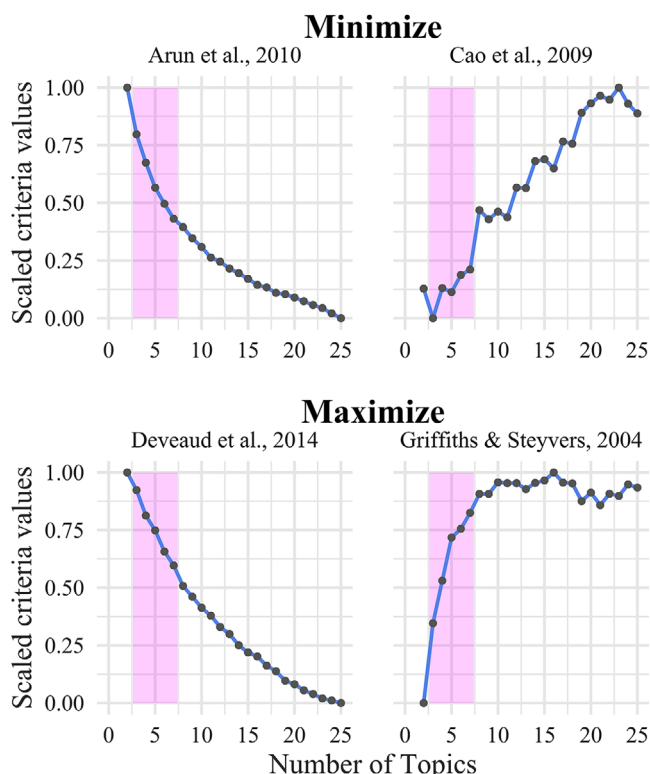


FIGURE 1. Optimal topic selection. Optimal number of topics displayed according to four separate criteria. Shaded rectangle displayed the range in which topics were evaluated for substantive meaning. [Color figure can be viewed at wileyonlinelibrary.com]

Finally, we present the logistic regression results estimating the correspondence of the topic model with SME ratings.

Number of Topics

Figure 1 displays the optimal number of topics to be extracted across each of the four criteria. As would be expected, the different criteria suggested differing numbers of topics. Relying only on the criteria suggested by Arun et al. (2010) would lead us to the conclusion that, essentially, the more topics estimated the better the fit, while the opposite conclusion would be reached if only evaluating the results relative to the criteria outlined by Deveaud et al. (2014). The Cao et al. (2009) and Griffiths and Steyvers (2004) criteria were more nuanced and suggested differing, but similar ranges of topics. After seven topics, however, a marked increase in the Cao et al. (2009) criteria was observed, indicating a poorer fit, while only a moderate increase (indicating a better fit) was observed with the Griffiths and Steyvers (2004) criteria. Taken together, these results suggested that between three (the minimum value for the Cao et al., 2009 criteria) and seven topics should be extracted (as displayed by the shaded rectangle in the background of Figure 1). Each topic solution (3–7) was therefore evaluated based upon SME judgment for substantive meaning.

Two science content SMEs with strong science education expertise and familiarity with the NGSS evaluated the 3–7 topic solution models independently for substantive meaning. Differences in labels assigned to topics were resolved through a post-hoc process of consensus. In general, the SMEs began with a three-topic solution and first considered each

Table 1. Substantive Labels Assigned to Final Topic Solution

Topic	Substantive Label
1	Analyzing data and using evidence to understand organisms and systems
2	Using scientific evidence to understand Earth systems
3	Energy
4	Genetic information
5	Scientific and technological solutions

word within a given topic individually, including their known use within the NGSS. Second, the two SMEs considered their combined meaning within each extracted topic as being a representation of a possible construct in science, while noting the β value for each word, and assigned each topic a substantive scientific label (see Figure 3). Third, and finally, the SMEs examined each named topic for independence, overlap, and stability across solutions. This process was applied for each topic-solution set between three and seven topic solutions.

The SMEs independently settled on the five-topic solution as best representing the data, and their justifications were similar, as determined through collaborative debriefings following their evaluation. Increasing the number of topics from three to four and from four to five led to the addition of substantively new and meaningful topics with little overlap among other topics. However, adding a sixth topic led to substantive overlap with at least two other extracted topics, and thus, the five-topic solution was deemed most appropriate. Table 1 displays substantive labels determined by the two SMEs for each of the topics derived from the final (five-topic) model.

Mapping Topics to Standards and Words to Topics

Figure 2 displays a heatmap of the γ values across topics for each of the 59 performance expectations evaluated. Brighter colors represent a higher likelihood of the given topic being reflected by the given standard. For example, Performance Expectation ESS1.1 (bottom) is represented almost entirely (99%) by Topic 1: *Analyzing data and using evidence to understand organisms and systems*. Performance Expectation ESS2.4, however, is represented partially (85%) by Topic 1, and partially (15%) by Topic 2: *Using scientific evidence to understand Earth systems*. These heatmaps can assist in deriving substantive meaning from the topics, given that the performance expectations that predominately make up a topic can be investigated. Once substantive meaning is assigned, however, the topics can likewise help bring new meaning back to the performance expectations, as themes may emerge that were not otherwise apparent.

The top 15 words within each topic are displayed in Figure 3, according to their estimated β values. Note that in many instances, there were ties among beta values around 15. Those selected for display were chosen randomly. For example, if the 13th–18th words all had the same β value, only the 13th–15th would be displayed based on a random selection of that β value range. Each topic has been labeled according to its identified substantive label. Note that roughly equivalent plots were used for each of the three to seven topic solutions when arriving upon the final topic model through SME judgment.

As can be seen, some words were represented across multiple topics (e.g., *evidence* is most strongly associated with *Organisms and Systems*, but is represented in the top 15 words in every topic outside of *Scientific and Technological Solutions*). These are general words that represent cross-cutting science concepts. These words could be removed to increase the distinction between topics, but this would come with the cost of not representing the standards, or the intent of the standards, as well (i.e., cross-cutting concepts are a key feature of the NGSS). Similarly, variants of the same word are occasionally represented twice (e.g., “solution” and “solutions” in the *Scientific and Technological Solutions* topic). A stemming or lemmatization approach could have been applied to the words to collapse the different forms (plural, tense) into a single word. Preliminary models explored this approach, but there was evidence that the model did not perform as well (i.e., the relation between the model and SME ratings was lower). Further, our SMEs suggested that there are substantive reasons for keeping the variation (e.g., “comparing competing solutions” is different from “developing an optimal solution”). This was largely supported by the β values; for example, the β value for the word “solutions” was .037 for the *Scientific and Technological Solutions* topic and 0 elsewhere, while the β value for the word “solution” was .018 for the *Scientific and Technological Solutions* topic and .01 for the *Earth Systems* topic. We therefore opted to keep these words in the model. Note that the model was estimated with every word including a beta value for every topic, but the distinguishing characteristic of a topic was identified by the relative weight of the β values across topic.

Predicting Items to Topics

In addition to the topic model providing information about the interrelated nature of the NGSS performance expectations, the final, selected model was also used *predictively*. That is, new text was provided to the model, and topic-level predictions were made. Specifically, each word from the new text was assigned a probability that it was generated from each topic. We used this approach with text from the Grade 8 science items from the Oregon AA-AAS. Importantly, all text from a given item, including the item prompt and the item options, was used when making predictions. These predictions were then used to evaluate content coverage within the assessment (i.e., coverage of the identified topics).

Figure 4 displays a summary of the overall topic coverage by theoretical item difficulties (*Low*, *Medium*, and *High*) via radar plots and bar charts. The test-level topic representations were estimated as the mean of the item-level topic predictions (see the top panel of Figure 5 for a sample of item-level predictions). In each display, the thick gray line represents the expected probability if all topics were equally represented (i.e., 20%). Items in the *Low* category, for example, slightly underrepresented all topics, with the exception of *Organisms and Systems*, which was highly overrepresented. This same pattern was present for the items written to be of *Medium* difficulty, although the pattern was even more severe. For the *High* items, the *Genetic Information* topic was the most underrepresented, but overall, the topic coverage was considerably better. Although the evidence was not definitive, this information could be useful to guide subsequent investigations of content representativeness.

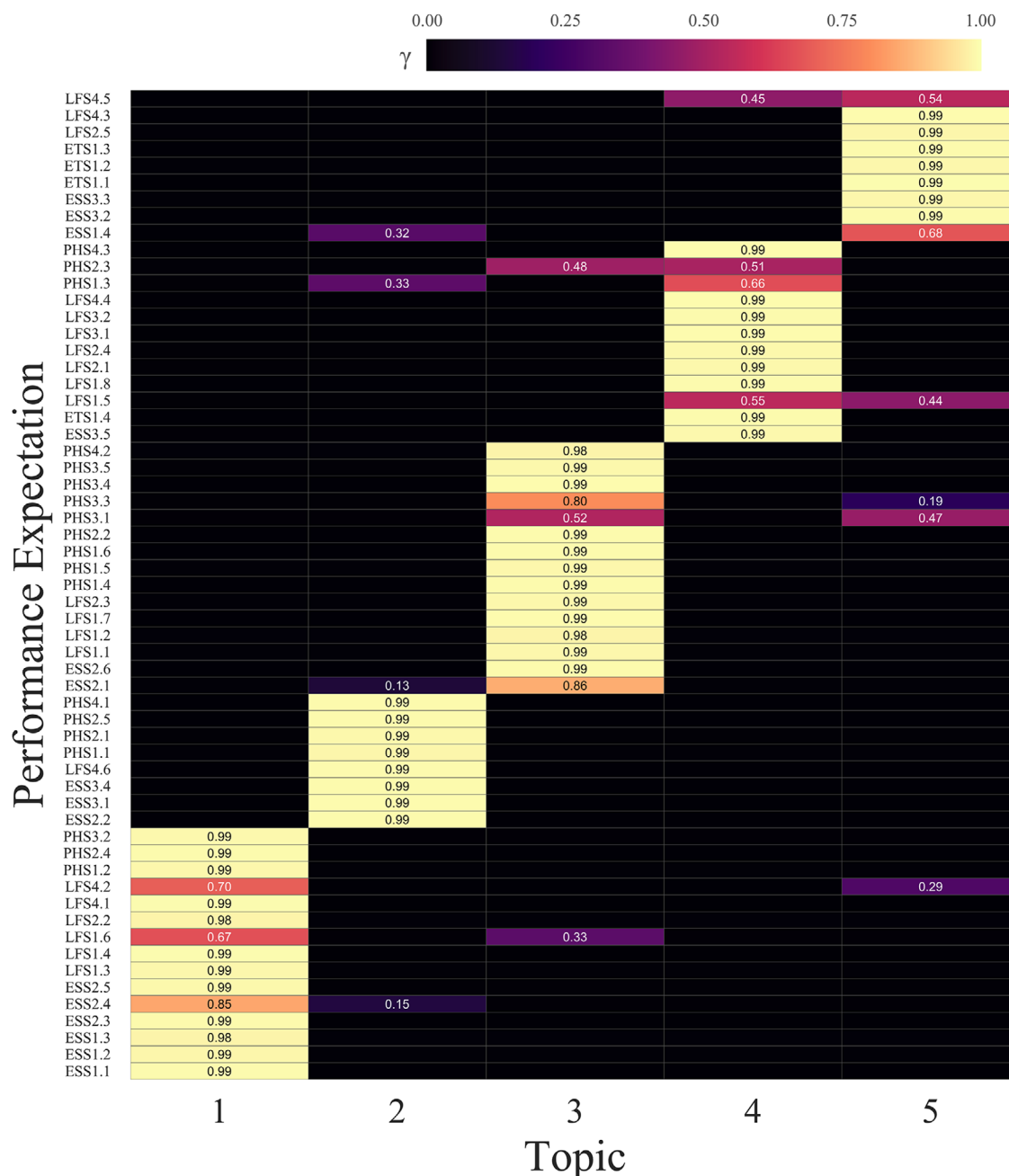


FIGURE 2. Heatmap of gamma values. Cells displayed in brighter colors have higher gamma values, representing a greater likelihood that the given topic is represented by the given standard. Actual gamma values displayed in all nonzero cells. [Color figure can be viewed at wileyonlinelibrary.com]

Correspondence with Alignment Study Results

Finally, we investigated the correspondence between the topic-level predictions from the text-based model and alignment results from SME. The top panel of Figure 5 displays topic-level predictions for each of six randomly selected items. Notice that a clear topic prediction was made for items 1, 3, and 4, while no prediction was made for items 2 and 5; these latter items are examples where the items did not include any text that could be classified by our model, and the probability was equally distributed across the five topics. Finally, item 6 was split evenly between two topics, with no clear prediction made.

The bottom panel of Figure 5 displays overlapping (kernel density) distributions of the average SME ratings, separated

by whether a topic-level prediction was made by the model. As can be seen, many items were rated with an average of 2.0, the highest possible alignment rating, regardless of whether a topic-level prediction was or was not made. However, the distribution of SME ratings when a prediction was made for the item (e.g., items 1, 3, and 4) had considerably more density at this upper extreme, while simultaneously having less density at all other values. In other words, if a prediction was made for a given item, it was more likely to have an average SME rating of 2.0, relative to items for which a prediction was not made (e.g., items 2, 5, and 6), while also being less likely to have any lower value. Importantly, however, a number of items were still judged as having adequate alignment despite no topic-level prediction being made. Items without a topic-level

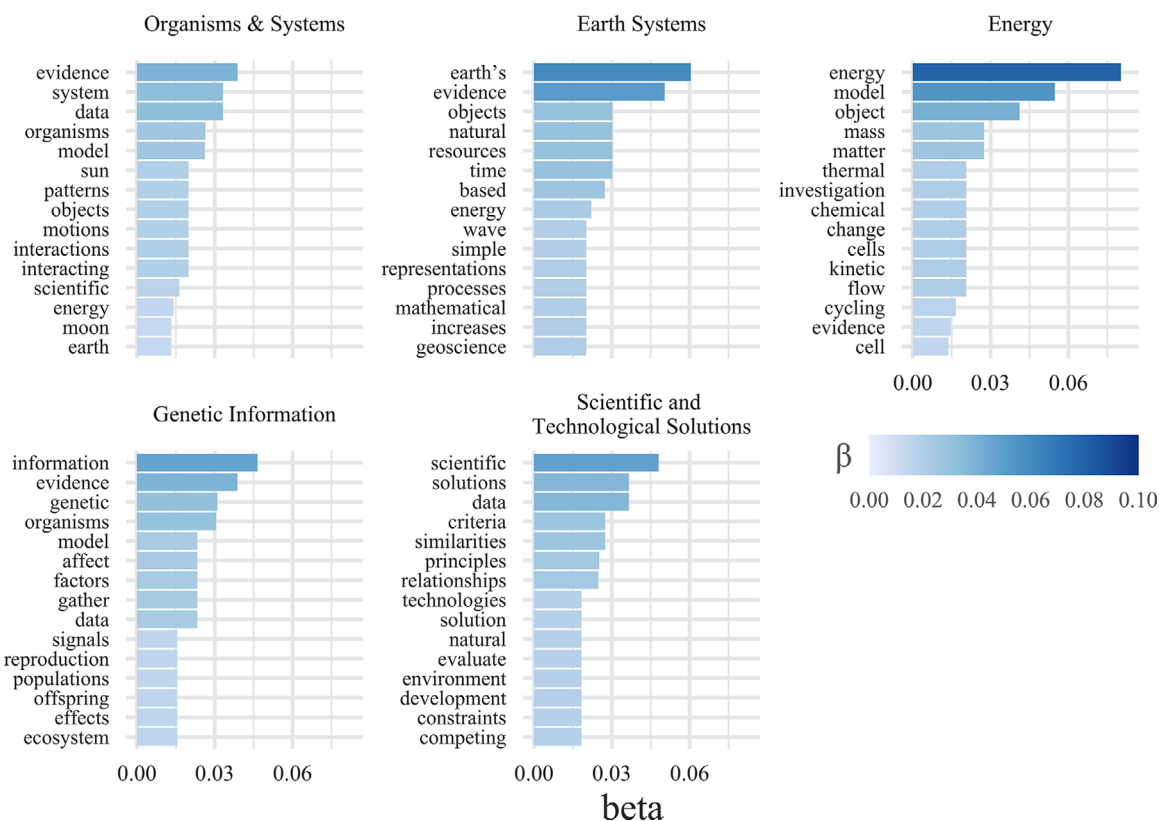


FIGURE 3. Beta values. Highest probability of words generated by topic. Note that the substantive labels were assigned post-hoc. [Color figure can be viewed at wileyonlinelibrary.com]

prediction may therefore be candidates for further review, but only these follow-up reviews would be likely to indicate lack of alignment for individual items. The logistic regression model indicated that, on average, items *with* topic-level predictions were 1.76 times more likely to be rated as aligned to the content standards by SMEs, which was significant (95% CI [1.53, 2.03], $z = 7.74$, $p < .001$).

Discussion

Validity based on test content is critical to the overall evaluative judgment of the validity of a test for a given use (Kane, 2006; Messick, 1995) and is a core element examined in alignment studies. This paper introduced a text-based, machine learning method, specifically topic modeling (Mohr & Bogdanov, 2013), to evaluate the textual congruence between content standards and test items. This approach has the potential to supplement the methods of current alignment and content-related validity studies by providing a triangulating source of evidence, demonstrating the utility of machine learning methods (topic modeling) alongside more traditional approaches, as highlighted by others (Nelson, 2020). Text mining also holds potential as a means of learning more about the content standards themselves (i.e., identifying groups of standards through previously unobserved themes), test items, and the textual link between standards and test items, as part of the iterative assessment development and refinement process.

The nature of the topics identified in this analysis is substantively noteworthy. The middle school NGSS curriculum standards are expressed through 59 performance expectations, which served as the data sources for our topic model.

These performance expectations are based upon (and include each of) the three dimensions of the standards: (a) scientific and engineering practices, (b) cross-cutting concepts, and (c) disciplinary core ideas (National Research Council, 2012). Interestingly, and perhaps appropriately given how the tridimensional performance expectations were designed to be implemented in science classrooms (National Research Council, 2015), the topics we identified reflected a blend of domains (PS, LS, ESS, and ED) and/or dimensions. For example, Topic 1 seemed to emphasize analyzing data (one of the scientific and engineering practices) to understand organisms (a topic reflected in multiple disciplinary core ideas) and systems (a focus of both disciplinary core ideas and cross-cutting concepts). Topic 2 (*Using Scientific Evidence to Understand Earth Systems*) exhibited a similar blend of the three dimensions of the standards. Topic 3 was largely focused on energy, which is reflected in both disciplinary core ideas and cross-cutting concepts. Even Topic 4, which was nearly exclusively focused on genetic information, was strongly associated with the term *evidence*, which is a core part of a scientific and engineering practice. Thus, the topics reflected the way in which the performance expectations were intended to blend the three-dimensional framework upon which the NGSS are based. The five extracted topics usefully summarize performance expectations that are highly variable in terms of the scientific content they emphasize and the outcomes they target (i.e., an in-depth understanding of disciplinary core ideas; the capacity to engage in a scientific and engineering practice).

Information from topic modeling could be used to supplement data collected through formal alignment studies by providing a source of triangulating evidence for individual items. If the same information was provided by such separate

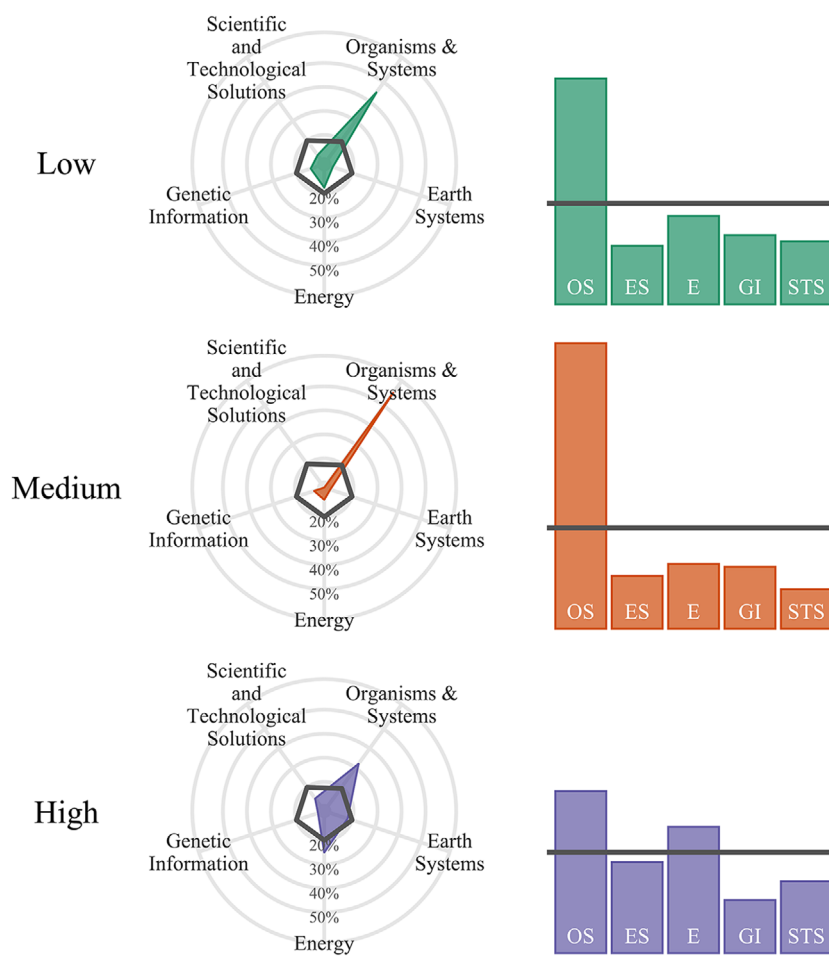


FIGURE 4. Overall content coverage. Gray bands that represent the expected probability of topics were equally distributed. Topics are over-/underrepresented to the extent the bar/polygon extends above/below the line. [Color figure can be viewed at wileyonlinelibrary.com]

sources, it may help reinforce the overall substantive conclusions about particular items or the test as a whole. If, however, the evidence did not agree, then it may serve as a prompt to collect additional data and conduct further investigations, perhaps with an additional SME. In addition, the analyses could potentially inform item and test development *during* the developmental process. That is, the analysis could serve as a diagnostic tool to better understand the content coverage of a particular test and when key vocabulary may be missing from a particular item. Topics are composed of groups of content standards (as shown in Figure 2), and a lack of representation for a given topic therefore represents a lack of representation for these standards. Using topic models as a source of diagnostic information would likely prompt item writers to include vocabulary directly corresponding with standards associated within specific topics. This may not always be desirable, however (i.e., when items are targeted at the lower tail of the ability distribution), and would likely need to be incorporated as part of item writer trainings. This also reinforces that the evidence obtained from text analyses is likely best thought of as complementary and supplemental, rather than as a replacement for any existing methods. From a time- and cost-benefit perspective, however, it is more efficient and cheaper to conduct analyses of data in-house, especially during early phases of development and refinement, than to conduct (post-hoc) formal alignment studies. Part of the benefit of an analytic approach is that the analyses could

be conducted much more regularly to inform a truly iterative test documentation and validation process.

Limitations and Future Directions

There are several limitations to the approach discussed in this paper that should be kept in mind. First, the results depend highly upon the chosen topic model, including the number of topics extracted, stop words removed, and the substantive meaning assigned to the topics. We view the topic model used here as preliminary and in need of further external validation. We have therefore published figures similar to Figures 2 and 3 for topic solutions between 3 and 16 topics within a GitHub repository that houses all the code for our analyses.¹ Ideally, the topic model itself could be refined over time with input from SMEs in the field. By making all of our work public, we hope to encourage collaboration and potentially “crowdsource” an optimal topic model to represent the NGSS—an approach that has been successfully applied in other fields (e.g., Arganda-Carreras et al., 2015; Bentzien et al., 2013).

In our study, we built a model using text from the NGSS and evaluated items from the Oregon alternate assessment against this model. However, the topic model could be used with the text from *any* test items written to measure the

¹See <https://github.com/datalorax/text-analysis-content-validity>

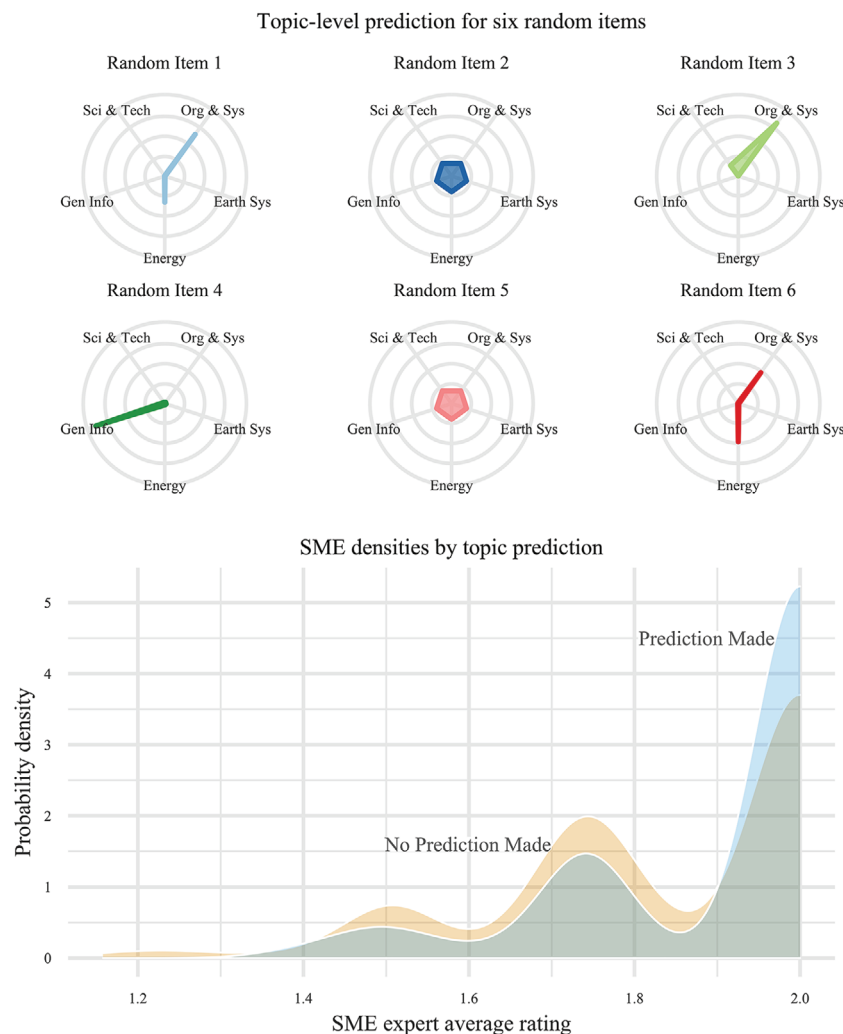


FIGURE 5. Distribution of topic model (TM) probabilities and subject matter experts (SMEs) ratings, split by whether or not a model prediction to a topic was made by the topic model. [Color figure can be viewed at wileyonlinelibrary.com]

NGSS. While our two SMEs came to consensus on a five-topic solution, it is possible that alternative SMEs may have arrived at different conclusions. Including as many voices in the conversation as possible may help reach a general consensus among the field, and models could be constructed as a general-purpose tool for any test developer writing test items to the given standards. A general training (perhaps through an online module) may first need to be provided to the SMEs providing input to ensure that they understand the tool (algorithm) they are helping to develop, its purpose, and how their input may influence the final model. Once consensus was reached, an online application programming interface (API) could be created so that test developers could submit the text from their items and obtain immediate feedback, perhaps in a similar form to the results provided here. The API could be hosted on a secure website or, alternatively, the model itself could be shared through an R package or similar outlet.

It is also the case that some items, by design in testing contexts like AA-AAS, have little to no text. In this case, our model essentially fails, because it cannot relate the item to any topic, and the probability is equally dispersed across topics. We view these cases as items to be flagged for further review. While our results indicated that these items were, on

average, significantly less likely to be rated as aligned, not all of these items received poor alignment ratings. A more long-term solution would be to utilize other areas of machine learning, such as image recognition. This would include first creating a data set with items with no text, and providing the standard(s) with which the item was judged as being aligned (through SME judgment). The algorithm could then begin to learn how different image features relate to item alignment. Again, however, this would require the algorithm being iteratively trained, with different algorithms being needed for different content areas. SME consensus would again need to be established as a source of validity evidence for the algorithm itself. However, if text mining could be used in combination with image recognition, a more powerful content-related validity evidence system could be established. We again, however, would view such a system as diagnostic and a source of triangulating evidence, with human judgment ultimately maintained as the “gold standard.”

Finally, our specific application utilized the NGSS content standards and a science test. In some ways, science is “easy” to evaluate the linkage between tests and items, because key vocabulary must be represented in each. It therefore remains to be seen how this approach would work with other highly assessed content areas like English language arts and

mathematics. Reading would seem to follow rather naturally; for example, if a standard discussed verbs, an item may ask students to identify the verb within a sentence. Because the word “verb” would occur in both, the match would be found. However, there are other instances where the mapping between the text used in content standards and items may be disparate, yet aligned. Mathematics also includes many keywords (e.g., area, measure, multiply) that may be helpful in classifying items, but the mathematical symbols themselves could also be useful by simply treating these symbols as “words” and evaluating their co-occurrence. For topics such as geometry, an image recognition approach such as discussed above may also be helpful. However, further research is needed to better understand how the method would generalize to other content areas, as well as different grades within a content area. It is possible, for instance, that the method would work well in mathematics for primary grade content, but not middle or high school content.

Conclusion

Topic modeling can provide additional information and context to content-related validity studies, as illustrated here. The results could also be used diagnostically during item development to help identify areas of under- and overrepresentation, and particular items that may be missing critical vocabulary. If such a method were applied operationally, however, it would be critical to communicate the shortcomings of the approach with stakeholders, and that the results not be relied upon in isolation. Attaching probabilities to items may communicate a more “objective” approach, but, as discussed above, this is not necessarily the case. As a general diagnostic tool and supplemental source of evidence, however, we believe that the approach discussed here holds considerable promise. The practical costs associated with alignment studies are high and should ideally provide confirmatory evidence, while ad-hoc text analyses could serve a more diagnostic and exploratory role.

Topic modeling is unlikely to ever *replace* the information obtained through alignment studies. Rather, we have argued that topic modeling can provide an additional and complementary source of content-related validity evidence (which is separate from *alignment*, although alignment data may provide content-related validity evidence). This information may be useful for diagnostic purposes and as a general indicator of the textual congruence between a set of test items and content standards. Future work should further investigate alternative topic model representations of both the NGSS standards, as well as other content standards. Although our specific model showed some promise (i.e., significantly reduced likelihood of being rated as aligned when no topic-level prediction was made), it is possible that alternative modeling representations may provide better results and, even if the model were to remain the same, further validation from a range of science SMEs would be needed before the method was operationalized.

The gathering of content-related validity evidence is dominated by a single method—alignment studies. Machine learning approaches may help to broaden the evidential base. In some cases, specifically in the early grades (pre-K; K-2), tests may not have content standards, and the approach discussed here may seem infeasible. However, formal alignment studies are also challenging, if not impossible, at these levels and,

while predictions may not be able to be made toward topics represented in content standards, much could be learned by mining the items themselves and evaluating the topics therein. In sum, while the work presented here is preliminary, text mining and topic modeling as a whole hold considerable promise for the field of measurement and psychometric research, specifically as an alternative means to gathering validity evidence.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement: Issues and Practice*, 34, 22–33.
- Arganda-Carreras, I., Turaga, S. C., Berger, D. R., Cireşan, D., Giusti, A., Gambardella, L. M., . . . Seung, H. S. (2015). Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, 9, 142.
- Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 391–402). Springer.
- Bentzien, J., Muegge, I., Hamner, B., & Thompson, D. C. (2013). Crowd computing: Using competitive dynamics to develop and refine highly predictive models. *Drug Discovery Today*, 18(9–10), 472–478.
- Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states’ content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21–29.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2–3), 143–296.
- Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7–9), 1775–1781.
- Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), 61–84.
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1), 5228–5235.
- Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1–30. <https://doi.org/10.18637/jss.v040.i13>
- Hou-Liu, J. (2018). *Benchmarking and improving recovery of number of topics in latent Dirichlet allocation models*. <https://vixra.org/abs/1801.0045>.
- Jelveh, Z., Kogut, B., & Naidu, S. (2018). *Political language in economics*. Columbia Business School Research Paper 14-57.
- Kane, M. (2006). Content-related validity evidence in test development. *Handbook of Test Development*, 1, 131–153.
- Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5, 361–397.
- Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1), 1608.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics*, 41(6), 545–569. <https://doi.org/10.1016/j.poetic.2013.10.001>

- National Research Council. (2012). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.
- National Research Council (Committee on Guidance on Implementing the Next Generation Science Standards, Board on Science Education, & Division of Behavioral and Social Sciences and Education). (2015). *Guide to implementing the next generation science standards*. Washington DC: National Academies Pres.
- Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49, 3–42.
- Nikita, M. (2016). *Ldatuning: Tuning of the latent Dirichlet allocation models parameters*. Retrieved from <https://CRAN.R-project.org/package=ldatuning>
- Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, 43(1), 88–102.
- Onix. (2018). Onix text retrieval toolkit API reference: Stop word list 1. <http://www.lextek.com/manuals/onix/stopwords1.html>.
- Oregon Department of Education. (2017). *2016-2017 technical report: Oregon's alternate assessment system. Peer review documentation: Critical elements 1-6*. Retrieved from <https://www.oregon.gov/ode/educator-resources/assessment/AltAssessment/Documents/techreport.pdf>
- Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators*. CPRE Research Report Series.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209–228.
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.
- Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3). <https://doi.org/10.21105/joss.00037>
- Silge, J., & Robinson, D. (2017). *Text mining with r: A tidy approach*. Sebastopol: O'Reilly Media, Inc.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, 5(4), 299–321.
- Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26(1), 100–107.
- Snowball. (2018). English stop word list. <http://snowball.tartarus.org/algorithms/english/stop.txt>.
- United States Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Non-regulatory guidance*. Retrieved from <https://www2.ed.gov/policy/elsec/guid/altguidance.doc>
- United States Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment of peer review process*. Retrieved from <https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf>
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Research Monograph No. 6.
- Webb, N. L. (2002). *Depth-of-knowledge levels for four content areas*. Language Arts.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from <http://ggplot2.org>
- Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from <https://CRAN.R-project.org/package=tidyverse>