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

September 06, 2019

|  |
| --- |
| 1 |
| 1 |
|  |

# Author note

Correspondence concerning this article should be addressed to , . E-mail:

Abstract

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; Webb, 1999) 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 provide an illustration by building a model from the *Next Generation Science Standards*, with textual congruence evaluated against items within the Oregon alternate assessment for students with significant cognitive disabilities. We discuss the utility of this approach, and show how visualizations can be used to evaluate the overall coverage of the content standards across the test items.

*Keywords:* Validity, Content, Text-Mining, Machine Learning, textual congruence

Word count:

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

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 developmental 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 over-representation, 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 learning 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 confirmatory evidence. The method primarily addresses domain representation and relevance, and we illustrate how visual displays of the information can aid 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 experts judging the alignment between the content represented within the test items and the content represented in the corresponding content standards. Numerous methods for evaluating item-standard alignment 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 a panel of subject-matter experts 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, however, are generally difficult to design and execute. Bhola, Impara, and Buckendahl (2003) outline several challenges to alignment studies, 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 the challenges listed above, alignment studies provide a critical source of content-related validity evidence for large-scale testing programs and help ensure that the tested content matches the content taught in schools. Expert 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. For example, while Anderson et al. (2015) discussed how group dynamics may influence alignment study results, the authors still opted for all items to be reviewed by experts. Rather than using a panel-based method, however, the authors used an online platform with judgments made independently. Yet, while expert 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 text used in 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 *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.

## 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, the topics emerge from corpora of text based on the frequency of 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 pre-identified, to a more inductive approach where meaning is allowed to emerge (Mohr & Bogdanov, 2013).

Topic modeling is still relatively new in the scheme of text-based analytic research. Blei and colleagues (Blei, Ng, & Jordan, 2003) introduced latent Dirichlet allocation (LDA) in 2003, which has become perhaps the most common topic modeling estimation procedure. Prior to LDA being introduced, inductive or latent themes in text were achievable primarily through qualitative analysis (Nikolenko, Koltcov, & Koltsova, 2017). Topic modeling, however, has potential in an array of contexts. Quinn, Monroe, Colaresi, Crespin, 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 of these papers would not be feasible.

Topic models are estimated from a *document-term matrix*, in which the columns represent each unique term from the corpus, the rows represent each “document” evaluated, and the cells represent the counts of each word within each document. The documents represent discretized instances of the overall corpus, such as a collection of 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 proportion 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). Other examples include, in combination with sentiment analysis, improved efficiency and accuracy of the sentiment related to consumer comments of products (Lin & He, 2009), and the aggregation of results across scientific studies despite differences in terminology and fields of inquiry (Gefen, Endicott, Fresneda, Miller, & Larsen, 2017). Topic modeling has even shown emerging potential to produce similar results to traditional qualitative frameworks, such as grounded theory, when utilizing a semi-supervised form of LDA in which researchers restrict potential topic assignment to predefined intervals (Nikolenko et al., 2017). The increasing accuracy of such techniques provides support for potential use in expanded applications.

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 themes/subdomains. Building the topic model includes a number of decision points, most importantly, the number of topics to be estimated, and we discuss the specifics of our model in detail in the Method section. However, once the model is built, the probability that any given text is represented by each topic can be estimated. In other words, the model is built from the content standards independent of the test items, and this model is then used to estimate the textual congruence of the items and standards (through model-based predictions using the item text). The relative representation of the topics across the test items can then be evaluated, along with the mapping of individual items to topics. An item found to not be represented by any topic would have the probability spread equally across all topics, which may be a candidate for further review and potential revisions (although it is also possible that the item represents the standards in ways outside of the text). These analyses may provide additional triangulating information with expert judgments, but perhaps more importantly, provide a viable source of diagnostic information during test and item development given that the analyses themselves are relatively fast and less cost- and labor-intensive than alignment studies. For example, items found to have no text conforming with any topic could be flagged for further review, which may identify out-of-scope items (e.g., aligned to a different grade level) before reaching a full alignment review. Similarly, the relative representation of topics across the items could be evaluated, with the test refined to ensure a roughly equal inclusion of all areas represented in the standards.

# Method

We evaluate the textual congruence of the middle school (Grades 6-8) Performance Expectations which are the standards outlined by the *Next Generation Science Standards* (NGSS), and the Grade 8 Alternate Assessment based on Alternate Achievement Standards (AA-AAS) in Oregon, as described more fully below. In what follows, we describe our data sources, including the NGSS and AA-AAS, as well as our topic model. We specifically 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. 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

As this study dealt with test items, The NGSS are based on the *Framework for K-12 Science Education* (Council, 2012) and are built around Performance Expectations, rather than more narrowly-defined standards of academic content knowledge, in each of grades K-5 and grade bands 6-8 (middle school) and 9-12 (high school). NGSS Performance Expectations are situated within one of four Domains (Physical [PS], Life [LS], Earth/Space [ESS], and Engineering Design [ED]) and organized into Arrangement Clusters (groups of interrelated Performance Expectations, e.g., Earth’s Place in the Universe [cluster ESS1]). Together, Performance Expectations unify three dimensions of science learning, requiring that students understand and apply *science and engineering practices* as they master *disciplinary core ideas* and engage *crosscutting concepts*, which span science broadly (National Research Council, 2015).

We created two document-term matrices, with one representing the content standards (and each NGSS Performance Expectation represented 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). Although preliminary models included Webb’s words, the words associated with scientific concepts (e.g., evidence, genetic, mass, motion) became conflated with common scientific verbs (e.g., design, describe, interpret, analyze) making the individual topics more difficult to discern. We therefore opted to keep the model as focused on the content as possible by removing these words.

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 pre-processed using the same steps as listed above for the test items, using all text represented within each item (i.e., item prompt and three answer options) and each test 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.

## 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 (note, there is a 1% reporting cap, not participation cap). Oregon does not include disability type as part of the eligibility requirements for AA-AAS participation. Rather, 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 ability to generalize learning across settings. SWSCDs require highly specialized services related to both their education and, often, social and medical needs. The majority of these supports are provided by educators across a variety of settings.

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. As mentioned previously, 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, non-verbal, gesturing).

## Procedure

Our process included (a) using topic modeling to identify key domains/topics within the statewide content standards based on frequency of word co-occurrence; (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 items to the modeled domains/topics, including the relative representativeness of each domain/topic within the test. The number of topics was determined through a combination of statistical analyses and expert judgment. The evidence gained from these analyses can then supplement formal alignment studies, while also potentially being useful as a diagnostic tool during test development.

### 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 mentioned above, Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003), is perhaps the most common estimation procedure for topic models and was the approach used here.

This implies that each document (performance expectation or test item) includes different proportions of each modeled topic (including 0% or 100%). Words can be shared between topics, but may be more strongly associated with specific topics over others. LDA simultaneously estimates both the mixture of topics within a document and the mixture of words within a topic. The matrix reports the estimated probability that each word belongs to (or was generated by) a given topic, while the 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 words within each topic (typically 10-20 words), based on the 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 expert judgment[[1]](#footnote-1). 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 to 25 topics for each model fit indicator.

Following the evaluation of topics, we found a range of topics which 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 experts. 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.

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

In this section, we first discuss the selection of the optimal number of topics. We then discuss the mapping of topics to standards, and words to topics. Finally, we 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 the results by whether they were written to be of *low*, *medium*, or *high* difficulty.

## Number of Topics

Figure ?? 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 appeared more nuanced and suggested differing, but similar ranges of topics. After six 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 six topics should be extracted (as displayed by the shaded rectangle in the background of Figure 1). Each topic solution (3-6) was therefore evaluated based upon expert judgment for substantive meaning.

Two science content experts with strong science education expertise and familiarity with the NGSS evaluated the 3-6 topic solution models independently for substantive meaning. Differences in labels assigned to topics were resolved through post-hoc collaboration. In general, the content experts began with a three-topic solution and first considered each word within a given topic individually, including their acceptation and known use within the NGSS. Second, the two experts considered their combined meaning within each extracted topic as being a representation of a possible construct (or topic of study) in the broad field of science, while noting the value for each word, and gave each topic a substantive scientific name (see Figure ??). Third, and finally, the experts examined each named topic for independence, overlap, and stability across solutions. This process was applied for each topic-solution set between three to six topic solutions.The substantive names between each competing topic-solution were compared for independence, overlap, and stability (i.e., did each topic in the lower solution remain, were new topics created that were independent of others and that had substantive scientific meaning).

The content experts independently settled on the five-topic solution as best representing the data, and their justifications were similar, as determined through collaborative de-briefings 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 amongst 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 ?? displays substantive labels determined by the two content experts for each of the topics derived from the final (five-topic) model.

## Mapping Topics to Standards and Words to Topics

Figure ?? displays a heatmap of the 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 (paste0(round(ess1\_1$gamma[1], 2)\*100, "%")) by Topic 1: *Analyzing data and using evidence to understand organisms and systems*. Performance Expectation ESS2.4, however, is represented partially (paste0(round(ess2\_4$gamma[1], 2)\*100, "%")) by Topic 1, and partially (paste0(round(ess2\_4$gamma[2], 2)\*100, "%")) 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 ??, according to their estimated 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 to 18th words all had the same value, only the 13th to 15th would be displayed based on a random selection of that value range. Each topic has been labeled according to its identified substantive label. Note that roughly equivalent plots were used for each of the 3-6 topic solutions when arriving upon the final topic model through expert judgment. As can be seen, some words were represented across multiple topics (e.g., *evidence* is most strongly associated with the Organisms and Systems topic, but is represented in the top 15 words in every topic outside of Scientific and Technological Solutions). These are generally words that represent cross-cutting concepts in science. These words could be removed to increase the distinction between each topic, 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). We therefore opted to keep these words in the model. It is also worth mentioning that the model is estimated with every word including a beta value for every topic, and this distinguishing characteristic of a topic is the relative weight assigned to the topic for a given word (through its beta value).

## 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, were used when making predictions. These predictions can then be used to evaluate content coverage within the assessment (i.e., coverage of the identified topics).

Figure ?? displays the probability of the text used within a random sample of nine items being generated from each of the five modeled topics. Random items 2, 5, and 7 all did not include any text that could be classified by our model, and the probability was equally distributed across the five topics. Note that this was expected, and is likely to be a more commonly observed outcome when working with data from AA-AAS assessment systems that intentionally limits textual density within items. The equal spread of probability across topics is therefore not an indication that the items did not align with the content standards, but rather that the text represented in the item (if any) was not represented within our model. Random Items 4, 6, and 8 all clearly aligned with a single topic, while Random Items 1, 3, and 9 had probabilities split between two topics. Mapping these topics back to the standards they composed therefore provided additional information about the items and their linkage to the content standards.

The model-based predictions from the text used within the items were also used to evaluate content coverage across the test items. In our application, test items were theoretically designed to be of *Low*, *Medium*, and *High* difficulty, as discussed in the methods. We were therefore interested in whether the coverage of the derived topics was equal across each type of item. Figure ?? displays a summary of the overall topic coverage by item type, with the average probability of items representing each of the five topics displayed via radar plots and bar charts. 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, were estimated as slightly under-representing all topics, with the exception of *Organisms and Systems*, which was highly over-represented. 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 coverage was considerably better. Although the evidence was not definitive, this type of information could be useful to guide subsequent investigations of content representativeness.

# Discussion

Content validity 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. 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 that served as one of the primary data sources for this study. 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 (Council, 2012). Interestingly, and perhaps appropriately given how the tri-dimensional 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., a deeper 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 triangulating evidence of the alignment for individual items. If the same information was provided by such separate 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 subject-matter expert. 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. From a time- and cost-benefit perspective, 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 both the number of topics 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 ?? and ?? for topic solutions between 2 to 16 topics within a GitHub repository that houses all the code for our analyses[[2]](#footnote-2). Ideally, the topic model itself could be refined over time, perhaps through a semi-supervised approach to topic modeling (Lu & Zhai, 2008), that would place predefined constraints on topic assignment based on expert consensus across 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, Muegge, Hamner, & Thompson, 2013). In our study, we built a preliminary model using text from the NGSS. We then evaluated 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 NGSS. In our approach, two expert reviewers came to consensus on a five-topic solution. It is possible, however, that alternative content experts 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 subject matter experts providing input to ensure they understand the tool they are helping to develop, its purpose, and how their input may influence the final model. Once some form of consensus was reached, an online Application Programming Interface (API) could even be created so 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 the topics. We view these cases as a flag for further review. In other words, the item may align well with the content standards and contribute meaningfully toward the content-related validity evidence of the test, but not include any text. A more long-term solution would be to utilize other areas of machine learning, such as image recognition. This would include first creating a dataset with items with no text, and providing the standard(s) with which the item was judged as being aligned (through expert 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. Expert 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 (and the image recognition algorithm could be used to learn the topics, so the algorithms would work together), 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 being the “gold standard”.

Along these lines, the extent to which textual congruence established through topic modeling methods correlates with human-based judgments is not yet clear. Although not a clear one-to-one correspondence, it would be helpful to understand how, for example, estimates of categorical concurrence from Webb’s model correspond with topic representation (see Webb, 1997). As mentioned, we view topic modeling as potentially being beneficial as an *additional* source of content-related validity evidence, as well as a useful diagnostic tool during item development. However, understanding how model-based predictions do or do not align with human judgments may help refine the model while also providing a source of validity evidence for its use.

Readers may consider what value the topics add in terms of summarizing the Performance Expectations. As noted above, the middle school NGSS Performance Expectations are already grouped into 12 Arrangement Clusters across the four domains. These arrangement clusters represent a more fine-grained description of performance expectation groupings (relative to the five topics we identified) and are based upon pre-defined and theoretical criteria. These theoretical considerations could potentially be useful within a semi-supervised LDA application. The topics we identified were based solely on the co-occurrence of words within the Performance Expectations but then reviewed post-hoc for substantive and theoretical meaning. These are alternative approaches to the same basic goal. The topics identified, however, did reflect the contents of the Performance Expectations in a succinct and comprehensive way. It is also worth mentioning that our approach does not provide a clear linking between an item and an individual performance expectations. That is, the specificity of the information provided is at the topic (sub-domain) level, rather than to a specific skill. This is helpful for evaluating the relative representation of these topics across the test items, but may be less helpful for a specific item. However, if a given item is found to not be represented by *any* of the topics, it may warrant further investigation. While features outside of text (e.g., graphics) could relate to the content standards, flagging these items for further review may help detect items in need of revision before going to full alignment panel reviews.

Finally, our specific application utilized the NGSS content standards and a science test. In some ways, science is “easy” to evaluate 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, the image recognition approach 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 with latent Dirichlet allocation is one of several text-mining procedures that could provide additional information and context to content-related validity studies. The results could also be used diagnostically during item development to help identify areas of under- and over-representation, and particular items that may be missing critical vocabulary. If such an method were applied in operational use, 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 numbers and probabilities to words and items may communicate a more “objective” approach, but as discussed above, this is not necessarily the case. As a general diagnostic tool, and as a supplemental source of evidence, however, we believe the approach discussed here holds considerable promise. The practical costs associated with alignment studies are high. Ideally, the alignment study should essentially provide confirmatory evidence, while ad-hoc text analyses could serve as a more diagnostic and exploratory.

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. (2014). AERA, apa, & ncme. *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., … others. (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*(Jan), 993–1022.

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.

Council, N. R. (2012). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press: Washington, DC.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84.

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in r with annotated code: Studying online reviews and the stack exchange community. *CAIS*, *41*, 21.

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. doi:[10.18637/jss.v040.i13](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*.

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*(Apr), 361–397.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 375–384). ACM.

Lu, Y., & Zhai, C. (2008). Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on world wide web* (pp. 121–130). ACM.

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.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter. *Poetics*, *41*(6), 545–569. doi:[https://doi.org/10.1016/j.poetic.2013.10.001](https://doi.org/https://doi.org/10.1016/j.poetic.2013.10.001)

National Research Council. (2015). *Guide to implementing the next generation science standards (committee on guidance on implementing the next generation science standards, board on science education, & division of behavioral and social sciences and education eds.)*. Washington DC: National Academies Pres.

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>.

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. *National Center for Research on Evaluation, Standards, and Student Testing. Retrieved June*, *2*, 2005.

Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, *1*(3). doi:[10.21105/joss.00037](https://doi.org/10.21105/joss.00037)

Silge, J., & Robinson, D. (2017). *Text mining with r: A tidy approach*. "O’Reilly Media, Inc.".

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299–321.

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>

Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research monograph no. 6.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. Research monograph no. 18.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>

Wickham, H. (2017). *Tidyverse: Easily install and load the ’tidyverse’*. Retrieved from <https://CRAN.R-project.org/package=tidyverse>

1. It is important to note that we consider the topic model preliminary. All code used to estimate the models (and produce this manuscript) will be made publicly available upon publication via a GitHub repository. [↑](#footnote-ref-1)
2. The repository is currently set as private for peer-review purposes, but will be made public upon publication. [↑](#footnote-ref-2)