

Evaluating Content-Related Validity Evidence Using Text Modeling

Daniel Anderson¹, Brock Rowley¹, & Sondra Stegenga¹

¹ University of Oregon

Author Note

Correspondence concerning this article should be addressed to Daniel Anderson, 5262 University of Oregon. E-mail: daniela@uoregon.edu

Abstract

The alignment of test items with content standards is a critical source of content-related validity evidence within high-stakes testing and accountability frameworks. Typically, alignment studies are conducted with panels of experts providing qualitative judgments on the degree of alignment with each item in the test and the representative content standards. Various summary statistics are then calculated from these judgments (e.g., categorical concurrence, balance of representation; Webb, 1999). In this paper, we propose an alternative approach that capitalizes on text-based machine learning procedures, specifically topic modeling, to identify text-based clusters within the content standards. The probability that each item from a statewide assessment aligns with each cluster/topic can then be estimated as an additional source of content-related validity evidence. We discuss this approach, and show how visualizations can be used to evaluate the overall coverage of the content standards by the test items.

Evaluating Content-Related Validity Evidence Using Text Modeling

For the last two decades, the landscape of education has evolved through the passage of legislation aimed at improving student outcomes through standards driven accountability such as the No Child Left Behind Act (No Child Left Behind, 2002) and its more recent renewal as the Every Student Succeeds Act (see McGuinn, 2016). This has led to a vast increase in standardized educational assessment use. As part of this movement an intensive focus on alignment has also emerged due to the need for congruence between assessment, standards, curricula, and instruction in order to impact student outcomes (e.g. Porter, 2002). Specifically, the goal of alignment is to “create a coherent educational system that conveys a clear and unified message about expectations and goals” (???).

The alignment of test items to content standards is imperative to content validity. Content-related validity evidence 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). Commonly, alignment studies are used to determine content validity for standards-based assessments, with panels of experts judging the alliance between the content represented in the standards and the content represented in the test items (Sireci, 2007; Webb, 1997). Alignment studies are difficult to execute and design, even under ideal conditions, and present researchers with potential sources of error. The chosen methodology, number of participants, professional judgments, consensus through discussion or averaged by ratings, costs associated with travel, and time are elements for consideration. This process can be time consuming and cost intensive (Anderson, Irvin, Alonzo, & Tindal, 2015).

According to Kane (2006) content validity is necessarily based on judgment, more subject to confirmation bias, and appears to lack the objectivity available in construct-related evidence. This paper adds to the literature on content-related validity evidence with recent

methodological advancements related to mining and analyzing textual data. Evidence of textual congruence between test items and content standard can then supplement, though not replace, information from alignment studies through efficient but effective means. Specifically, we propose using machine-learning based approaches, specifically topic modeling, to mine the educational standards for groups of words that co-occur frequently. These data-derived clusters can then be evaluated for substantive meaning, and the extent to which the text within individual test items corresponds with these clusters can be directly evaluated as an additional, alternative source of content-related validity evidence.

Topic modeling is a statistical method for garnering meaningful insight from data (???). Specifically, it is a probabilistic method for identifying latent topics in text based documents. Its history stems from content analysis and latent semantic analysis but, rather than the researcher predetermining the topics to be analyzed and coded, the topics emerge from corpora of text based on frequency of co-occurrence (i.e., text-based correlations). Topic modeling has advanced the field of text analysis from simply identifying specified words through a deductive approach where topics are pre-identified to a more inductive approach where meaning is allowed to emerge through a corpus of texts (???). Topic modeling is still relatively new in the scheme of text-based analytic research, with the first article by Blei published in 2003 (Blei, Ng, & Jordan, 2003). Prior to these advancements in machine learning and more specifically topic modeling, inductive themes or latent meanings in text were achievable mainly through qualitative analysis (???). Topic modeling, however, is demonstrating potential for use in a broad array of contexts with good success, including use in business for sentiment analysis of consumer comments about products and services (???), understanding political themes across a range of documents (e.g. Hagen, 2018), and aggregating results across scientific studies despite differences in terminology and fields of inquiry (???). Topic modeling has even shown emerging potential to produce similar results to traditional qualitative frameworks such as grounded theory on qualitative survey data when utilizing a semi-supervised form of latent dirichlet analysis (LDA) (???). The

increasing accuracy of such techniques thereby provide support for potential use in expanded applications.

In this paper we explore the use of topic modeling and specifically LDA as a source of content-related validity evidence to evaluate the textual congruence (alignment) between test items and content standards. We adopt a similar theoretical framework to alignment studies (Porter, 2002; Webb, 1997, 1999), but from a text-based machine learning perspective, which holds potential as a triangulated source of content-related validity evidence.

Method

In this section, we describe the measures used for our specific application, as well as the standards they were designed to measure. We then discuss our analysis (topic model) in detail, including how we arrived upon the specified number of topics within the standards, and how the results can be used to evaluate the overall textual congruence between the content standards and the test items.

Content Standards

brief description here

Measures

Our application utilizes the statewide alternate assessment based on alternate achievement standards (AA-AAS) for student with the most significant cognitive disabilities (United States Department of Education, 2005) in one western state. We evaluate the concordance between the text in the Grade 8 *Next Generation Science Standards* (NGSS) and the text used in the AA-AAS item stems and response options. As Ysseldyke and Olsen (1997) note, “There is more variability in the skill levels and needs of this 1% of the students than there is in the rest of the total student population” (p. 16). Correspondingly, the development of items followed a staged process, where content standards were first identified, and then three versions of essentially the same item were developed to be of, theoretically, *low*, *medium*, and *high* difficulty. In science, key vocabulary is a critical component of demonstrating knowledge and all *high* items were written to include this vocabulary.

However, this was not the case for the *low* or *medium* items and we therefore presumed, a priori, that the textual match of the *high* items with the content standards would be greater than the *low* or *medium* items.

Analyses

In evaluating the concordance between the language used in the content standards and the language used in the test items, our approach is to use a text-based machine learning model, specifically topic modeling (see Mohr & Bogdanov, 2013), to mine the standards and evaluate the topics represented therein. Once this model is trained, we can estimate the probability that each item is represented by each topic. In other words, the model learns the patterns of words from the standards, and we can then evaluate whether the words used in the items correspond to those patterns.

Topic modeling is akin to exploratory factor analysis, where latent variables (topics) are estimated based on the probability that the words within the topic will co-occur. In our investigation, we expected seven topics to emerge, which generally correspond to the sub-domains represented in the Grade 8 NGSS standards: (a) Heredity, (b) Earth and the Universe, (c) Earth and Humans, (d) Motion and Stability, (e) Energy, (f) Waves, and (g) Engineering Design. Post-hoc explorations of the words represented within each topic confirmed this structure. Topics were estimated using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We removed stop words (common words like of, a, the, and, is) from the onix, SMART, and snowball lexicons (Lewis, Yang, Rose, & Li, 2004; Onix, 2018; Snowball, 2018), as implemented in the *tidytext* R package (Silge & Robinson, 2016). Additionally, we removed verbs associated with Webb's depth of knowledge levels (Webb, 2002). These removals helped ensure the topics were clustered around content-related words, rather than specific verbs prevalent throughout the standards, or overly common words. Topics were estimated using *textmodeling* package (Grün & Hornik, 2011) within 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 the full proposal, we will discuss in greater detail our modeling process and how we arrived upon our seven topics. Due to space limitations here, we share only our preliminary results from that model.

Figure 1 displays the overall content coverage of the test, separated by items that were theoretically designed to be of *Low*, *Medium*, and *High* difficulty. Essentially, the average probability of items representing each of the seven topics is displayed on the log scale, specifically $\log(p(x_i + 1))$. The thick gray band represents the expected probability, if all topics were equally represented. Items in the low category, for example, are estimated as slightly under-representing the heredity and engineering design topics, while over-representing motion and humans activity with earth. The medium items are somewhat problematic, with the Motion topically highly over-represented, and Engineering Design, Heredity, and Humans all highly under-represented. Across item types, Heredity and Engineering Design were universally under-represented.

Figure 2 displays the probability of a random sample of nine items aligning with each of the seven topics. Random Item 2, 5, 6, and 8 all did not include any text that could be classified by our model, and the probability that the item aligned with each topic was equally spread. Note that this does not imply the items did not align with a given topic, but that the text represented in the item was not represented by our topic model. Random Items 3, 4, and 7 all clearly aligned with a single topic, while Random Items 1 and 9 had their probability split between two topics.

Discussion

Content validity is critical to the overall evaluative judgment of the validity of a test for a given use. This paper introduces a new method using text mining procedures to evaluate the concurrence between language used in the content standards and language used in the test items. From a cost-benefit perspective, it is much cheaper to conduct an analysis

of data in-house than to conduct alignment studies. These analyses could even be conducted during item and test development to inform the developmental process. However, the analyses are not intended to *replace* the evidence gathered during alignment studies, but rather to *supplement*. Part of the benefit of the analytic approach, however, is that they could be conducted much more regularly to inform the iterative test documentation/validation process.

It should also be noted that our analysis and results presented here are preliminary. Before the conference, we plan to obtain feedback from content experts to verify or provide guidance on modifications to our trained model, given that the validity of the procedure depends on the validity of the trained model (i.e., all the topics make sense and sufficient topics are extracted to cover the content represented in the standards). From a bigger-picture perspective, however, because these models are based on the standards, rather than any individual items, the models themselves could be made public with other provided the opportunity to provide input. Further, they could actually apply the model to their own data to evaluate content coverage or topic concurrence for individual items in a manner similar to that shown here. For the conference paper, we plan to provide much more detail about the modeling, its strengths and limitations, and a more in-depth illustrations of the results of our application.

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.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.

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

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.

McGuinn, P. (2016). From no child left behind to the every student succeeds act: Federalism and the education legacy of the obama administration. *Publius: The Journal of Federalism*, 46(3), 392–415.

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

No Child Left Behind. (2002). Act of 2001, 20 usca 6301 et seq.

Onix. (2018). Onix text retrieval toolkit api reference: Stop word list 1. <http://www.lextek.com/manuals/onix/stopwords1.html>.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and

practice. *Educational Researcher*, 31(7), 3–14.

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/>

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

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477–481.

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>

Ysseldyke, J. E., & Olsen, K. (1997). Putting alternate assessments into practice: What to measure and possible sources of data (nceo synthesis reports).

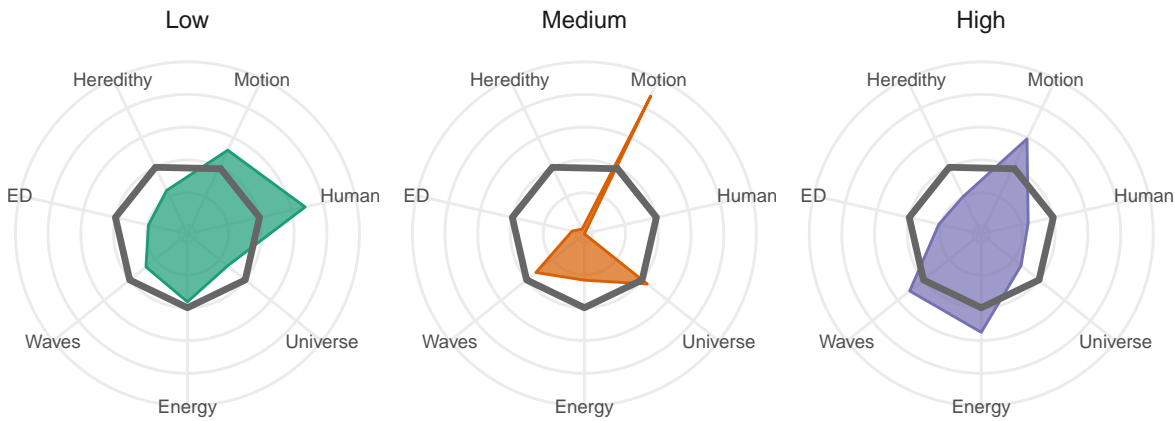


Figure 1. Overall Content Coverage

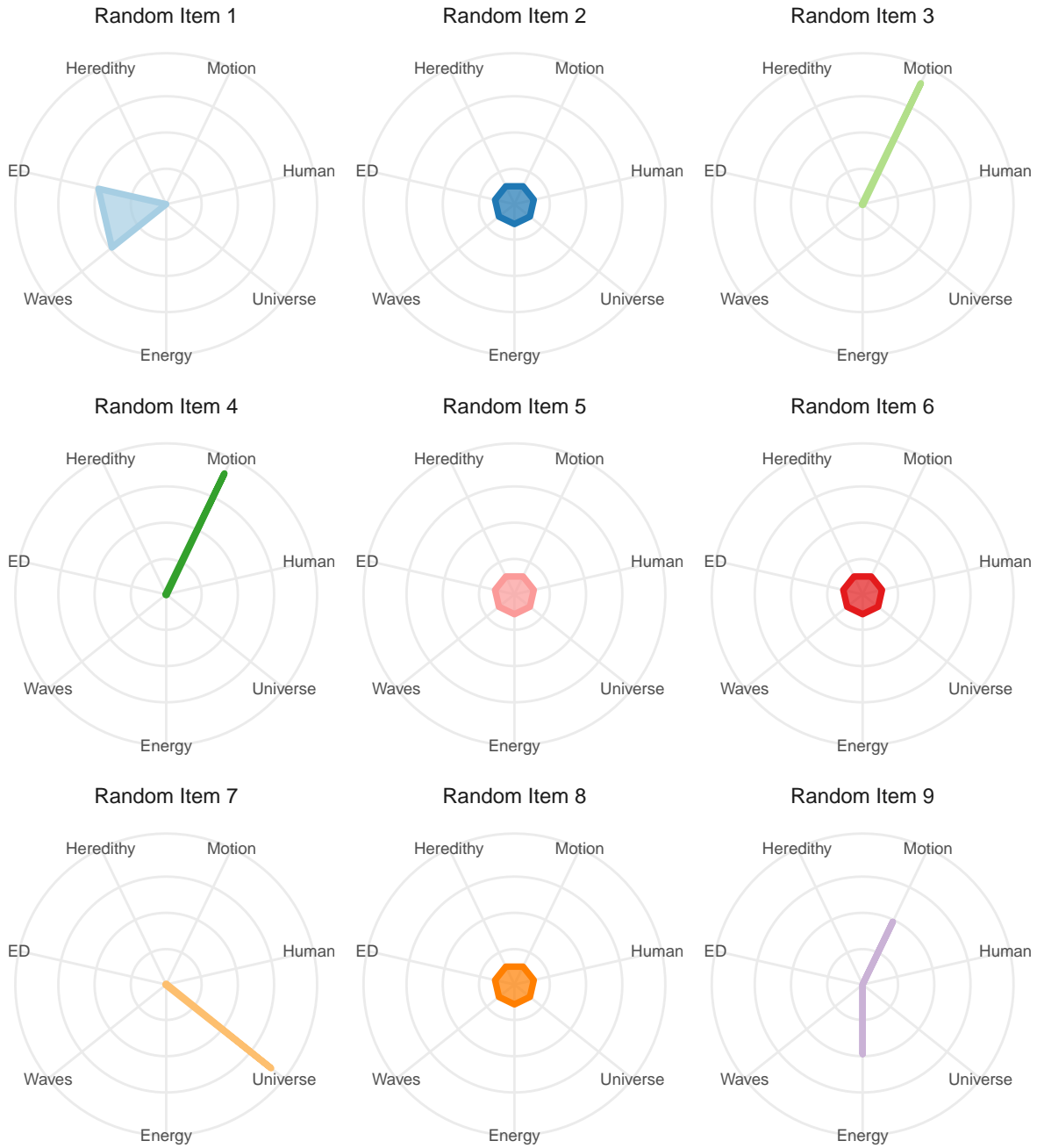


Figure 2. Probability of topics by item: Random sample of nine items