Evaluating Content-Related Validity Evidence Using Text Modeling

Daniel Anderson[1], Brock Rowley[1], & Sondra Stegenga[1]

[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

Topic modeling is applied with science content standards to evaluate semantic clustering. The probability that each item from a statewide assessment belongs to each cluster/topic is then estimated as a source of content-related validity evidence. We also show how visualizations can map the content coverage of the test.

Evaluating Content-Related Validity Evidence Using Text Modeling

Conceptual Framework

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). Empirical evaluations of content validity evidence generally come in the form of alignment studies (Sireci, 2007; Webb, 1997). In this presentation, we explore the use of text-mining procedures to evaluate the correspondence between the language used in content standards and the language used in test items as an additional source of content-related validity evidence.

Methods

In our application, we evaluated the concordance between the text used in the Grade 8 *Next Generation Science Standards* (NGSS) and the text used in items within a statewide alternate assessment for students with significant cognitive disabilities (United States Department of Education, 2005). We applied a text-based machine learning model known as topic modeling (see Mohr & Bogdanov, 2013). Topic modeling is akin to exploratory factor analysis with textual data, where clusters of words (topics) that are likely to co-occur are estimated through latent variables. We trained a topic model on the standards, arriving upon interpretable and meaningful topics, and then used this model to estimate the probability that each item was represented by each topic. In other words, the trained model became the "machine" which classified items. In our investigation, we expected seven topics to emerge, which generally correspond to the sub-domains represented by the standards. This number resulted in interpretable topics and was supported by the data.

Topics were estimated using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) with the *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)

Preliminary Results

In the conference paper, we will discuss in greater detail our modeling process. Due to space limitations here, we share only our preliminary results. 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 by the items. Items in the low category, for example, were estimated as slightly under-representing the Heredity and Engineering Design (ED) topics, while over-representing Motion and Humans' Activity With Earth (Humans). The medium items were somewhat problematic, with the Motion topic highly over-represented, and ED, Heredity, and Humans all highly under-represented. Across item types, Heredity and ED 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 spread equally. 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.

Conclusions and Implications

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. This method may be a valuable supplement to the evidence gathered during alignment studies. Part of the benefit of the analytic approach is that analyses could

be conducted much more regularly to inform the iterative test development process, given the cost and labor-intensiveness associated with alignment studies.

It should also be noted that our analysis and results are preliminary. Before the conference, we will obtain feedback from content experts to verify or provide guidance on our trained model, given that the validity of the procedure depends on the validity of the trained model. From a bigger-picture perspective, because these models are based on the standards, rather than any individual items, the models themselves could be made public, with members of the field providing input. Further, other test developers could apply the trained model to their item data to evaluate content coverage against a common benchmark. 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.*

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

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

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/

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

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.

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

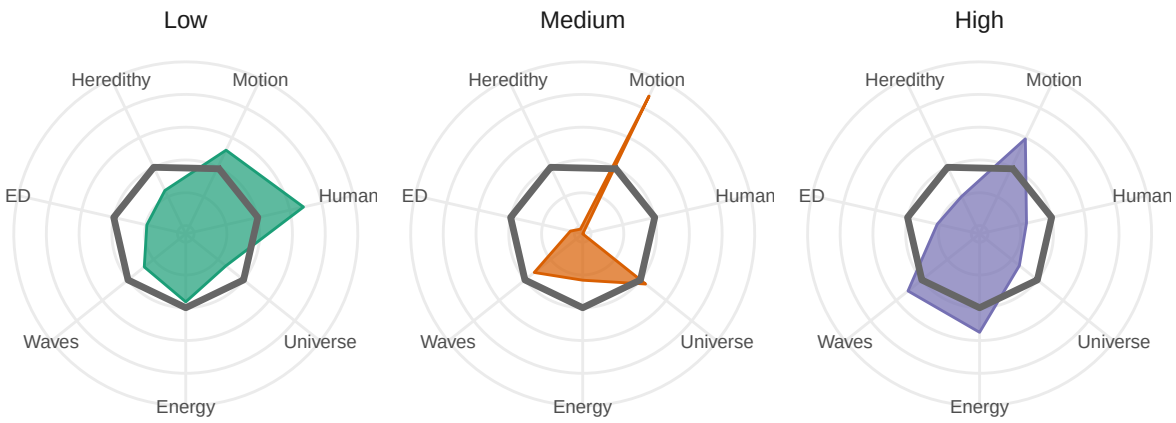

Figure 1. Overall Content Coverage

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