Evaluating Content-Related Validity Evidence Using Text Modeling

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

[1] University of Oregon

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.

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* (A. E. R. Association, Association, Measurement in Education, & others, 2014). Empirical evaluations of of content validity evidence for high-stakes tests generally come in the form of alignment studies, with panels of experts judging the alignment match between the content represented within the test items and the content represented in the corresponding standards (Sireci, 2007; Webb, 1997). In this presentation, we explore the use of text-mining procedures to evaluate the language used within the content standards and the corresponding language used in the test items (i.e., the item stems and response options). We demonstrate that these procedures can be successfully applied to not only identify correspondence between items and standards, but also to evaluate the content coverage and standards representation across the items.

Methods

Our particular application corresponds to evaluating the content-validity evidence for 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 concurrence between the text in the

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

Grade 8 *Next Generation Science Standards* (NGSS) and the text in the item stems and response options for the corresponding statewide AA-AAS. 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.

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 **(CITE)**, 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 word clusters (topics) are determined based on their semantic coherence **(CITE)**. Post-hoc investigations of the topics then provide substantive meaning. In our investigation, we expected seven topics to emerge, which generally correspond to the sub-domains represented in the Grade 8 NGSS standards.

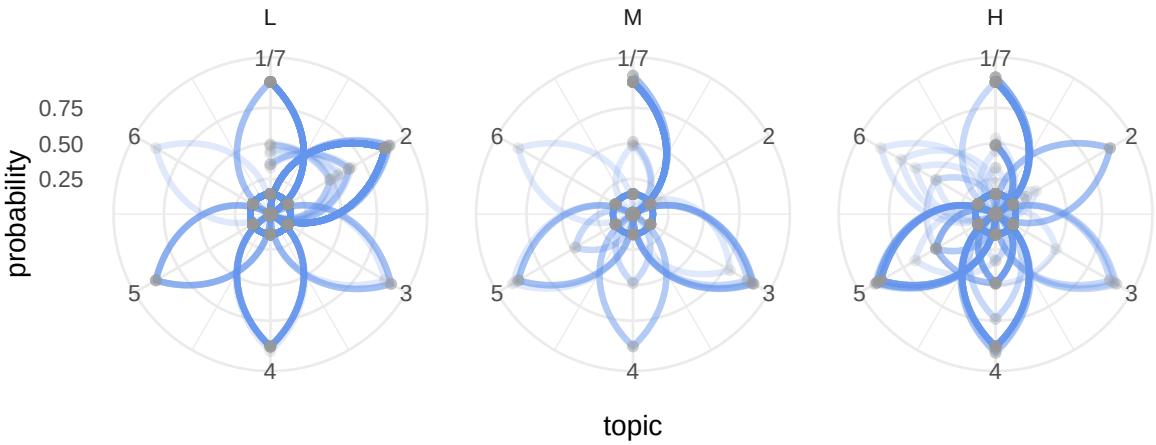
Latent Dirichlet Allocation (LDA) was used... **(CITE)**. We removed stop words according to... **(CITE)**.

Data were analyzed using using R (R Core Team, 2018) with the *textmodeling* package (Grün & Hornik, 2011). All code used for the analyses will be made publicly available.

Preliminary Results

Conclusions and Implications

Overall, this paper will discuss a new and innovative proposed approach to establishing validity through analyzing and categorizing text data via modern data tools of R and RStudio text analysis and structural topic modeling. It does not replace current methods for establishing validity but demonstrates initial promise to add strength to the development design. In a world of constantly evolving and growing data and data sources it is imperative as educational researchers that we not only begin to explore new methods that hold promise for increasing efficiency and accuracy with data analysis but also ensure that methods engage an element of translatability. By increasing the accuracy of constructs and improved validity we provide the opportunity to increase utility and translatability over a range of consumers in the educational community. In addition, modern technology provides an array of methods and open source resources and tools, such as R and RStudio, that not only provide the ability to capture and categorize the data but also visualize the findings. Again, this is an imperative piece in a world that has seen vastly increased calls for the translation of data and research to practice and practice to research.



References

Association, A. E. R., Association, A. P., Measurement in Education, N. C. on, & others. (2014). AERA, apa, & ncme. *Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.*

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.

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.

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