Data Schema to Formalize Education Research & Development Using Natural Language Processing

Hannah Frederick, Haizhu Hong, Amanda West, Margaret Williams, and Brian Wright
University of Virginia, acw9gs, hbf3k, hh3cy, maw3as, brianwright@virginia.edu

***Abstract*—Our work aims to aid in the development of an open source data schema for educational interventions by implementing natural language processing (NLP) techniques on publications within What Works Clearinghouse (WWC) and the Education Resources Information Center (ERIC). A data schema demonstrates the relationships between individual elements of interest (in this case, research in education) and collectively documents elements in a data dictionary. To facilitate the creation of this educational data schema, we first run a two-topic latent Dirichlet allocation (LDA) model on the titles and abstracts of papers that met WWC standards without reservation against those of papers that did not, separated by math and reading subdomains. We find that the distributions of allocation to these two topics suggest structural differences between WWC and non-WWC literature. We then implement Term Frequency-Inverse Document Frequency (TF-IDF) scoring to study the vocabulary within WWC titles and abstracts and determine the most relevant unigrams and bigrams currently present in WWC. Finally, we utilize an LDA model again to cluster WWC titles and abstracts into topics, or sets of words, grouped by underlying semantic similarities. We find that 11 topics are the optimal number of subtopics in WWC with an average coherence score of 0.4096 among the 39 out of 50 models that returned 11 as the optimal number of topics. Based on the TF-IDF and LDA methods presented, we can begin to identify core themes of high-quality literature that will better inform the creation of a universal data schema within education research.**

***Index Terms*—Data Schema, Education Research, Natural Language Processing**

I. Introduction

In the field of education, there is currently a gap between research and practice. [6] NewSchools worked with Gallup to ask a sample of 3,210 teachers, 1,163 principals, 1,219 administrators, and 2,696 students what they think of and how they use education technology inside and outside of the classroom. [10] They found that teachers, principals, and administrators all trust teachers the most for recommendations on education technology. Teachers ranked research papers low on this list because they don't place much trust in these reports that were planned and funded by the companies themselves. Teachers can also find these papers to be difficult to understand because the researchers' language is not necessarily the same as the educators' language. Laura Hamilton and Gerald Hunter (2020) saw this same trend with educational interventions as opposed to education technology - teachers tend to turn to other teachers for suggestions on academic interventions. [6] This shows a gap between research and practice because while researchers are writing reports about tools to be used in academia, educators are not fully translating this research into practice, and are thus not using these tools in the classroom.

Another problem with education research is that the field lacks a universal set of data standards and data schema. [4] There are four main issues that arise from this lack of uniformity. 1) Researchers tend to collect information on schools based on their individual projects and interests, often using particular or even proprietary terminology that is not understood or defined consistently across the field. 2) Data collection occurs on a project-by-project basis and is shared in inconsistent formats, yielding uninterpretable datasets including closed, searchable databases. 3) Data that is not publicly available may be collected throughout some of these efforts, but it is rarely distributed in a comparable or comprehensive way. 4) As researchers aim to keep up with new approaches; old datasets, descriptions, and categories are replaced, limiting the potential to evaluate trends over time. The first step towards enabling interoperability between these valuable, but unreadable, datasets is to standardize the data itself. [4]

One potential solution to these problems in education is the adoption of universally understood conceptual frameworks for the replication and validation of research papers. The PICO (Population, Intervention, Comparison, and Outcome) process is a type of data schema used in health care for defining clinical questions and evaluating clinical interventions. Successful adoption of these frameworks in the educational setting would allow for comparisons between educational interventions, and replication of these interventions would ensure that their conclusions are legitimate. InnovateEDU, a non-profit that wants to help close this gap between research and practice in education, is working with our team to build a data schema. This project is funded by the Bill & Melinda Gates Foundation. With feedback from working groups of researchers, educators, and practitioners, this schema will consolidate current standards in education. In order to facilitate the creation of this data schema, we implement NLP techniques to discover different themes and important terms that exist in educational research papers. Our work shows how representative the data schema is of educational literature, specifically from WWC and ERIC as these repositories are managed by the U.S. Department of Education. This open source schema, based on the models we form, would help educators find the studies that could lead to more helpful or useful changes in their students' education.

II. Related Work

In the field of education, Penuel et al. (2016) implement a survey-based study on research use among K-8 instructional policymakers sampled from urban schools and central offices across the nation. [11] The study revealed several important findings about how educational research was used by school and district leaders in practice. Specifically, most respondents cited using research for the purpose of decision-making, and are most likely to access research through professional associations and conferences rather than through individual researchers or U.S. Department of Education databases, including What Works Clearinghouse. Most useful research selected by respondents are about instructional practices and learning in the classroom, and also teaching and learning in specific subject matter, which support our selection of math and reading education as target subdomains. Respondents generally reported positive attitudes towards research regarding the usefulness of research, but one third of them have concerns about the potential political implications. Finally, respondents were reported to be willing to use research for decision-making, but felt they had limited access to the latest research for their needs and largely doubted their abilities to interpret the research results.

Before InnovateEDU works on developing their conceptual framework for education, it is important to consider previous framework's many advantages and disadvantages, as the limitations of one schema can contribute to the creation of another schema. Many of our aims for improving education research through a standardized data schema have already been successfully implemented in the field of health care. For example, Carpenter et al. (2012) developed a framework for articulating cancer comparative effectiveness for research data needs. [3] The framework presented by Carpenter has since served as a starting point for multiple frameworks that could, when fully implemented, address comparative effectiveness research needs, accelerate the pace of comparative effectiveness research, and enhance the adoption of research findings by the multiple stakeholders interested in improving patient outcomes.

Building on Carpenter's model, Hruby et al. (2016) developed a data schema for clinical research needs using the expert-derived framework. [7] Three data sources, ClinicalTrials.gov, EHR data requests, and EHR SQL queries, were sampled to obtain sentences and queries representing typical clinical research needs. The obtained sample data were analyzed and annotated semi-automatically through a natural language processing-assisted process and one human encoder, based on the original Carpenter model. The iterative annotation process derived a modified and enriched data schema, which was then evaluated by both a direct comparison of class preservation and eight clinical researchers. Suggestions from the eight evaluators were incorporated to construct a final participant-enriched data schema. However, Hruby et al. also found that limitations of this data schema included potential bias regarding data sources and the ambiguity introduced by abstracting medical concepts.

In addition to bias and ambiguity, there have been other potential concerns in relation to the creation of a health care data schema. Huang, Lin, and Demner-Fushman (2006) investigate the limitations of the PICO framework, the most common evidence-based schema currently in use in the health care industry. [8] The authors found that PICO has been beneficial overall, especially in regards to therapy-based interventions, and has provided a clear structure in which to filter the natural language between patient and provider. On the other hand, the authors also delve into ways to improve PICO - most importantly, Huang et al. found that situations which cannot be translated into the PICO framework are systematically ignored and under-reported in research. These limitations are important to keep in mind in order to develop a more comprehensive and inclusive data schema in education.

Finally, Erikson and Frandsen (2018) perform a rigorous systematic review, according to AMSTAR and PRISMA standards, of the effectiveness of PICO as a literature search strategy tool. [5] Researchers employed a rigorous search strategy to query for relevant PICO framework assessment studies across several major databases. They identified 2,163 records in total but only found three articles to be "eligible for further review". The three articles were then compared on four aspects (study design, relevance assessment, choice of comparator, and outcomes reported) and found to be quite different. The results showed that the number of search blocks and the inclusion of outcome-related terms would affect the quality of the literature search. However, the systematic review did not find a significant effect of this PICO search strategy compared to the alternative PIC framework, PICOS framework, SPIDER framework, and unguided search. The study concluded that more research was needed to assess the effect of using the PICO framework in the literature search.

III. Methodology

The goal of our project is to use text mining and topic modeling to find the underlying language used in educational research papers from two different clearinghouses, WWC and ERIC. We use Term Frequency-Inverse Document Frequency (TF-IDF) to find which words are most important to a paper in a corpus of papers for each clearinghouse, and latent Dirichlet allocation (LDA) to find the probabilities of allocating each paper to each topic, or set of words, depending on the words that appear in each paper.

For our preliminary analysis, we do this to see if these distributions of probabilities are different between our two corpora, or two clearinghouses, of papers. If this is the case, then it is a sign that there are latent semantic differences between the papers in the two corpora. This means that educators will likely understand and interpret papers from these two sources differently.

For our main analysis, we first find the TF-IDF scores for all unigrams and bigrams in the overall WWC corpus. We then divide our WWC corpus into mathematics and reading papers to be analyzed separately. We do this because we expect the

most significant terms in our corpus will depend on the subject of papers within that corpus. Finally, we run LDA on the full corpus of WWC papers to see if there are core groups of thematic structures in educational research papers that should be represented in the data schema.

A. Corpus Building

The corpora built for both preliminary analysis and main analysis consist of paper titles and abstracts collected from the ERIC database. We selected the most recent papers for both corpora whenever possible. For corpora that contain high-quality research, the titles and abstracts of papers reviewed by WWC are used. WWC focuses on identifying high-quality research in order to answer the question: "what works in education?" The What Works Clearinghouse Process Brief describes the ratings every eligible study receives as follows: [14]

- **Meets WWC Design Standards Without Reservations:** Studies receiving this rating provide the highest degree of confidence that the intervention caused the observed effect.
- **Meets WWC Design Standards With Reservations:** Studies receiving this rating provide a lower degree of confidence that the intervention caused the observed effect.
- **Does Not Meet WWC Design Standards:** Studies receiving this rating do not provide confidence that the intervention caused the observed effect.

Reviews of papers were downloaded from WWC directly. We use the WWC dataset feature *Study_Rating* to collect papers that met WWC standards without reservations and the ERIC API field *ieswwcreviewed* to extract papers that have not been reviewed by WWC. Some corpora are subset further into separate groups of math papers and reading papers based on metadata indicators defined by WWC. This downloaded dataset includes a feature *Topic_Mathematics* to indicate that a reviewed paper is about math, or *Topic_Literacy* to indicate that a reviewed paper is about reading.

Papers with an ERIC identification number and a math tag and/or a reading tag would be pulled from the ERIC API. When we needed to find papers that were not reviewed by WWC but were about either math or reading, we used the ERIC API to search for papers that had "Mathematics Intervention" or "Reading Intervention" in the paper's abstract. The papers that were reviewed by WWC make up the WWC corpus, while the papers that were not reviewed by WWC make up the non-WWC corpus.

When comparing WWC and non-WWC corpora, we record the distribution of intervention names that appeared in the WWC corpus, then manually duplicate this distribution in the non-WWC corpus whenever possible (see Table VII). This is done by using the *Intervention_Name* field from the WWC reviews dataset. The names of interventions (ex. "Fraction-Face-Off!") are then individually searched for in the titles of non-WWC papers pulled from the ERIC API and added to our non-WWC corpus.

There are some limitations associated with our corpus building process. WWC itself contains reviews of papers, and does not always provide access to the papers themselves. For this reason, the WWC corpus is limited to papers reviewed by WWC that have an ERIC identification number and accessible through the ERIC API. Due to copyright laws, all corpora are also limited to only titles and abstracts instead of full length papers.

B. Text Pre-processing

Each paper is represented by the title and abstract combined as one string of text. All of the text was tokenized, or split up into individual units that give meaning to the text. In the reading corpus, we removed equal signs from the tokens, while we kept the equal signs in the math corpus because it gives the text meaning related directly to math. Before modeling, the text was also lemmatized and stop words and punctuation were removed. Lemmatization is the process of collapsing different versions of the same word to a common lemma (ex. "ran" and "run" both become "run"). Stop words include frequently used words that do not have any meaning on their own (ex: "the" and "and").

C. TF-IDF Method

We use TF-IDF, or Term Frequency-Inverse Document Frequency, to find a list of the most important unigrams and bigrams within our corpus of documents. TF-IDF is incredibly popular in the field of natural language processing. Beel et al. (2016) found that of the more than 200 papers on research recommender systems published over the last 16 years, 70% implement a TF-IDF weighting system in their model. [1]

By assigning a relevance value to each word or word pair, we can also examine the cardinality of words and how one word ranks in relation to others. To improve graphical readability, we convert these TF-IDF scores to Z-scores by subtracting the mean and dividing by the standard deviation for each word's TF-IDF score averaged across all documents in the corpus.

$$TF\text{-}IDF(i,j) = TF(i,j) \times IDF(i) \tag{1}$$

$$TF(i,j) = \frac{\textit{frequency of word } i \textit{ in document } j}{\textit{total number words in document } j} \tag{2}$$

$$IDF(i) = log(\frac{\textit{total number documents in corpus}}{\textit{number documents containing term } i}) \tag{3}$$

In order to better inform the creation of a data schema, it is important to first understand what words make up the vocabulary of education research in the present. By speaking to the vernacular already in use, researchers are more equipped to adopt a universal data schema and older papers become more integratable. We aim to discover that vocabulary in the following sections with the inclusion of TF-IDF.

D. LDA Model

While TF-IDF identifies a set of significant unigrams and bigrams for documents in a corpus, the method does not reveal the latent thematical structures of the corpus. To uncover more useful insights, we proceed with the most widely used statistical topic model, latent Dirichlet allocation (LDA), first introduced by Blei et al. (2003). [2] LDA is a generative probabilistic model based upon the assumptions that each document in the corpus is a probability distribution of topics and each topic is a probability distribution of words. Topics, in the context of LDA, statistically represent the co-occurring patterns of words in the corpus. Similar to principle component analysis, LDA projects the corpus to a lower dimension and thus enables the discovery and interpretability of latent semantic meaning of a large collection of documents.

Mathematically, the LDA model first assumes there is a fixed set of K topics in the corpus and each topic is a probability distribution of vocabulary ϕ_z, which is a multinominal distribution drawn from a Dirichlet prior with hyperparameter β. Then, for a document i in the corpus, it is generated by first choosing a topic mixture θ_i, which is also a multinominal distribution drawn from a Dirichlet prior with hyperparameter α. The hyperparameters α and β control the amount of entropy in the Dirichlet distributions. Next, for each word in document i, first select a topic Z_j from θ_i and then select the word W_j from ϕ_{zj}. Therefore, for a document with N words, we can express the marginal distribution of a document as a mixture of topics:

$$P(W|\alpha,\beta) = \int P(\theta|\alpha)(\Pi_{n=1}^{N} \sum_{z_n} P(W_n|Z_n,\beta)P(Z_n|\theta))d\theta \tag{4}$$

And the joint distribution of a document as a mixture of topics can be expressed as:

$$P(\theta, W, Z|\alpha,\beta) = P(\theta|\alpha)\Pi_{n=1}^{N} P(Z_n|\theta)P(W_n|Z_n,\beta) \tag{5}$$

Therefore, according to Bayes theorem, the posterior probability distribution of a topic mixture can be formulated as the following equation:

$$P(\theta, Z|\alpha,\beta) = \frac{P(\theta, W, Z|\alpha,\beta)}{P(W|\alpha,\beta)} \tag{6}$$

The ultimate goal of an LDA model is to infer the posterior probability, which in this case is the conditional probability distribution of latent topic mixtures given the observed documents in the corpus. With the help of Bayesian estimation methods, such as variational inference and Gibbs sampling, the model can generate a probabilistic topic mixture distribution for each document in the corpus (θ_i) and a probability distribution of words for each topic (ϕ_z). Thus LDA is helpful for discovering the latent themes in a corpus.

However, since LDA is an unsupervised model, one weakness of this approach is the lack of a measure to evaluate the generated topic models. A longstanding challenge in building LDA models lies in choosing the optimal number of topics K. Traditionally, researchers rely on domain knowledge to choose the number of topics, afterwards checking if words with high probabilities for each topic meet their expectations. Due to our limited expertise in the domain of educational research, we decided to employ a less subjective metric called the coherence score. The coherence score for one topic is computed as the average cosine similarity between the top word context vectors and the centroid vector of that topic. Further, the overall coherence score is calculated as the aggregated mean over all topic coherence scores. [12]

As LDA is a probabilistic model, each run of tuning this model returns a different optimal number of topics. To counteract the stochastic nature of this model, we tune the optimal number of topics 50 different times. For each of these 50 runs, we set the search space for the number of topics to be from two to 20. Of these 19 different models, we select the one that returns the highest coherence score, as this model has the optimal number of topics for that run. Once all 50 runs are complete, we then select the most frequently occurring optimal number of topics.

IV. Results

For the corpus building of WWC research papers, 15,236 reviews of papers were initially downloaded from the WWC database, covering 2,738 unique papers in total. We assume that a unique citation indicates a unique paper. Of these unique papers, 1,943 have a unique ERIC ID. Based on the topic tag assigned by WWC discussed earlier, there are also 280 math papers, and 553 reading papers in the WWC corpus. Tables IV through VI in the Appendix show the distribution of study ratings for the full WWC corpus, the math WWC corpus, and the reading WWC corpus for our main analysis.

We then filter these 1,943 papers to just those that met WWC standards without reservation because we expect the greatest semantic differences to exist between the highest quality papers from WWC and the general quality papers from ERIC. This led to 88 math papers and 158 reading papers in the WWC corpus being studied in our preliminary analysis. Tables VII and VIII in the Appendix show the distribution of papers with and without intervention names in our math corpus and reading corpus for both WWC and ERIC.

A. Preliminary Analysis

For both the math corpus and the reading corpus, we run an LDA model with two topics to demonstrate if the two different sources of papers (WWC and non-WWC) predominantly fall into two different topics. Tables I and II show the distribution of allocation of papers for each topic in each corpus.

TABLE I
Distribution of Math Papers into Topics

	Topic 1	Topic 2
Non-WWC	49.3%	50.7%
WWC	26.9%	73.1%

Table I shows that in the math corpus, papers approved by WWC without reservation appear to fall predominantly

TABLE II
DISTRIBUTION OF READING PAPERS INTO TOPICS

	Topic 1	Topic 2
Non-WWC	55.5%	44.5%
WWC	41.6%	58.4%

(73.1%) in Topic 2, while papers that have not been reviewed by WWC evenly fall into either Topic 1 or Topic 2 (49.3% and 50.7% respectively). Alternatively, Table II shows that within the reading corpus neither Topic 1 nor Topic 2 clearly distinguishes papers approved by WWC without reservation nor papers not reviewed by WWC, with approximately 40% to 60% of papers from each clearinghouse being allocated to each topic.

B. Main Analysis

First we discuss our TF-IDF results using the overall WWC corpus. In the Appendix, Figure 1 presents the top 30 unigrams and Figure 2 presents the top 30 bigrams overall. We also analyze mathematics and reading papers from WWC separately. Figure 3 expresses the top 20 unigrams for reading compared to mathematics, and Figure 4 expresses the top 20 bigrams for reading compared to mathematics. In Figures 3 and 4 of the Appendix, blue bars denote common terms shared between reading and mathematics papers, while red bars denote words that only show up in reading papers, and green bars denote that of mathematics papers.

Next we discuss our LDA results using the overall WWC corpus. For 39 out of the 50 runs of tuning this model, the highest coherence score was achieved by choosing 11 as the optimal number of topics. The average coherence score of these 39 models is 0.4096. Therefore, we choose to run an 11-topic LDA model on our full WWC corpus. Next we generate the per-document topic assignment distribution and the per-topic word assignment distributions and evaluate the top 20 bigrams assigned to each topic. In the Appendix, Figures V through XV present the top 20 bigrams that occur in each topic, and Figure XVI demonstrates the number of documents that are assigned to each topic. We summarize and contextualize the TF-IDF and LDA findings in the conclusion.

V. CONCLUSIONS

We now discuss how the results of our TF-IDF and LDA models can aid in the development of InnovateEDU's data schema. This open source data schema should alleviate the four main issues that arise from the lack of a universal educational data schema: 1) inconsistent research terminology, 2) inconsistent data collection, 3) non-inclusive data distribution, and 4) high data turnover rates.

A. Preliminary Analysis

Our LDA preliminary analysis shows that, for mathematics studies, WWC papers are strongly associated with Topic 2, while non-WWC papers are equally associated with Topic 1 and Topic 2 (see Table I). This shows that there are strong structural differences between math papers approved by WWC and math papers not reviewed by WWC. It also makes sense that about half of these non-WWC papers were allocated to Topic 2 because even though these papers were not reviewed by WWC at this time, that does not necessarily mean that they cannot be accepted by WWC in the future.

Meanwhile, for reading studies, WWC and non-WWC papers are equally associated with Topic 1 and Topic 2 (see Table II). This shows that these differences are not that strong between reading papers approved by WWC and reading papers not reviewed by WWC. These results imply that the language used in high-quality papers on mathematics studies is different from that used in papers on standard-quality mathematics studies. Alternatively, the lexicon used in high-quality reading papers is not that different from that used in standard-quality reading papers. This also shows that there could be some intrinsic differences between math and reading papers.

B. Main Analysis

Relating to unigrams with high TF-IDF scores, many of the emerging words relate strongly to key words in the education space; "student" is by far the most important unigram, followed by "reading", "school", "teacher", and "program". We also see many words related to randomized control trials (RCTs), a common method used by papers reviewed by the WWC; for example, "intervention", "study", "effect", and "control" are all in the top 20 for overall unigrams. This finding is repeated in the unigrams broken up by mathematics and reading, with the addition of words such as "reading", "writing", and "literacy" appearing only in reading papers, and words such as "mathematics", "problem", and "result" appearing only in mathematics papers.

Bigrams, on the other hand, begin to provide more detail; in the overall corpus, "high school" is by far the most significant bigram, followed by "student achievement", "national board", "control group", and "professional development". Between math and reading papers, demographic information, such as "school district" and "high/middle/elementary school" are the most important bigrams common among the two subdomains studied. On the other hand, important bigrams that are unique to reading include "reading recovery", "significant difference", and "statistically significant". Meanwhile, important bigrams that are unique to mathematics include "problem solving", "student achievement", and "grade [of] student".

For both unigrams and bigrams, our TF-IDF analysis confirms the importance of three types of words in our WWC corpus: 1) words that are intrinsic to education papers ("student", "school", "teacher"), 2) words that refer to the type of study being presented ("randomly assigned", "control group", "statistically significant"), and 3) words that suggest the subgroup being studied ("grade", "gender", "state", "school district", "high school", "learning disability", etc.). Based on these findings, we recommend that InnovateEDU's data schema incorporate elements related to the students, the schools, and the teachers involved in educational studies. It should also incorporate elements related to the type of educational study

being presented and elements related to the group of students being examined.

With our 11-topic LDA model, we find that the top bigrams that were common across all topics include words that are intrinsic to education papers ("contain table", "table figure", "result indicate") and words that suggest the subgroup being studied ("grade [of] student", "middle school", "high school"). This supports our findings from TF-IDF scores that participants' demographic information of a study should be incorporated as elements in the data schema. We also explore which bigrams ranked highly in one topic but not so highly in the others to try to identify what each topic represents semantically. Table III summarizes our interpretation of each topic based on the top 20 bigrams that show up in each topic.

TABLE III
WWC TOPIC SUMMARIES

Topic	Documents Per Topic	Theme of Topic
1	371	Study Characteristics
2	2	Youth Intervention
3	48	Child Autism Intervention
4	2	Child Literacy Development
5	316	Higher Education
6	442	School Characteristics
7	84	Student Achievement
8	77	Language Development
9	104	Learning Disabilities
10	355	Reading Development
11	142	Early Childhood

We can identify which topics are most well-represented in this corpus based on the number of documents that are allocated to each topic. The most prevalent topics in this corpus are Topic 6 (School Characteristics), Topic 1 (Study Characteristics), Topic 10 (Reading Development), and Topic 5 (Higher Education) in descending order of the number of documents in each topic. As a result, we recommend that these four themes be the most represented in the data schema. This universal schema should help resolve the issue of inconsistent research terminology by documenting the definition of each element in the data dictionary. By standardizing research terminology with the validation of multiple working groups, this schema should increase data interoperability and inclusivity, thus alleviating the issues of inconsistent data collection and non-inclusive data distribution. More work may be needed to understand the evolution of education research; this can be accomplished by tracking how the needs of the data schema change after it has been deployed.

Regarding next steps, we plan to run and tune our LDA model on titles and abstracts within the ERIC database using the ERIC API and other top educational research journals (ex. the Journal of Research on Educational Effectiveness by SREE) using the Virgo API. This would allow us to increase our sample size of papers by multiple orders of magnitude (over 300,000 ERIC papers spanning 2013-2020). This should help overcome the limitations of analyzing only a small set of text for each paper instead of the full text of each paper. By expanding our corpus of research papers, we expect the optimal number of topics to increase and the semantic meaning of each topic to change. If the WWC is effectively representative of high-quality research within the education sphere, these additional databases could give a more all-encompassing representation of general education research as it currently exists.

ACKNOWLEDGMENT

We would like to thank InnovateEDU for leading this project and working with us, the steering committee and working groups for providing invaluable feedback, and the Bill & Melinda Gates Foundation for funding this project. We would also like to thank Lane Rasberry, Daniel Mietchen, the Department of Education, and the UVA Library for meeting with us on multiple occasions to discuss the project.

REFERENCES

[1] Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., & Nürnberger, A. (2013, October). Research paper recommender system evaluation: a quantitative literature survey. In Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation (pp. 15-22).

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. the Journal of machine Learning research, 3, 993-1022.

[3] Carpenter, W. R., Meyer, A. M., Abernethy, A. P., Stürmer, T., & Kosorok, M. R. (2012). A framework for understanding cancer comparative effectiveness research data needs. Journal of clinical epidemiology, 65(11), 1150-1158.

[4] Campbell, J. and Copa, N. (2019). Standardizing School Innovation Data: Report and Recommendations. Retrieved October 04, 2020, from https://www.aemcorp.com/

[5] Eriksen, M. B., & Frandsen, T. F. (2018). The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: A systematic review. Journal of the Medical Library Association, 106(4). doi:10.5195/jmla.2018.345

[6] Hamilton, L.,& Hunter, G. (2020). Educators Turn to Colleagues for Help with Struggling Students. Retrieved October 04, 2020, from https://www.rand.org/pubs/research_reports/RR2575z9-1.html

[7] Hruby, G. W., Hoxha, J., Ravichandran, P. C., Mendonça, E. A., Hanauer, D. A., & Weng, C. (2016). A data-driven concept schema for defining clinical research data needs. International journal of medical informatics, 91, 1-9.

[8] Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of PICO as a knowledge representation for clinical questions. In AMIA annual symposium proceedings (Vol. 2006, p. 359). American Medical Informatics Association.

[9] Menard, A. (2020). What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers. Retreived October 04, 2020, from https://fantasticanachronism.com/2020/09/11/

[10] New Schools Venture Fund (2019). Making meaning of the NewSchools-Gallup survey of educator and student perceptions of ed tech. Fordham Institute for Advancing Educational Excellent.

[11] Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., ... & Allen, A. R. (2016). Findings from a National Study on Research Use among School and District Leaders. Technical Report No. 1. National Center for Research in Policy and Practice.

[12] Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6.

[13] Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC medical informatics and decision making, 7(1), 1-6.

[14] WWC Process Brief: The Study Review Process. Retrieved April 3, 2021, from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_brief_review_052417.pdf

APPENDIX

TABLE IV
BREAKDOWN OF FULL WWC CORPUS INTO STUDY RATINGS

Ineligible for review	Does not meet WWC standards	Meets WWC standards with reservations	Meets WWC standards without reservations	Additional source not reviewed	Not rated	Total
459	706	214	441	117	6	1943
23.62%	36.34%	11.01%	22.70%	6.02%	0.31%	100%

TABLE V
BREAKDOWN OF MATH WWC CORPUS INTO STUDY RATINGS

Ineligible for review	Does not meet WWC standards	Meets WWC standards with reservations	Meets WWC standards without reservations	Additional source not reviewed	Not rated	Total
38	86	44	86	25	1	280
13.57%	30.71%	15.71%	30.71%	8.93%	0.36%	100%

TABLE VI
BREAKDOWN OF READING WWC CORPUS INTO STUDY RATINGS

Ineligible for review	Does not meet WWC standards	Meets WWC standards with reservations	Meets WWC standards without reservations	Additional source not reviewed	Not rated	Total
61	188	105	170	29	0	553
11.03%	34.00%	18.99%	30.74%	30.74%	0.00%	100%

TABLE VII
BREAKDOWN OF MATH WWC CORPUS AND MATH NON-WWC CORPUS

WWC Corpus		NonWWC Corpus	
88		88	
100%		100%	
With Intervention Names	Without Intervention Names	With Intervention Names height	Without Intervention Names
29	59	16	72
32.95%	67.05%	18.18%	81.82%

TABLE VIII
BREAKDOWN OF READING WWC CORPUS AND READING NON-WWC CORPUS

WWC Corpus		NonWWC Corpus	
158		158	
100%		100%	
With Intervention Names	Without Intervention Names	With Intervention Names height	Without Intervention Names
98	60	55	103
62.02%	37.97%	34.81%	65.19%

Fig. 1. Top TF-IDF Unigrams for Papers in the WWC Corpus

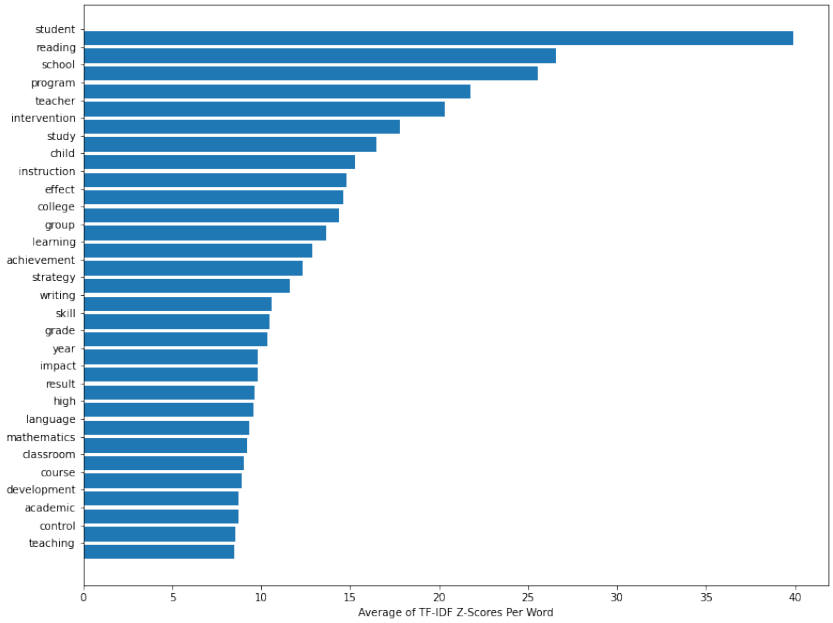

Fig. 2. Top TF-IDF Bigrams for Papers in the WWC Corpus

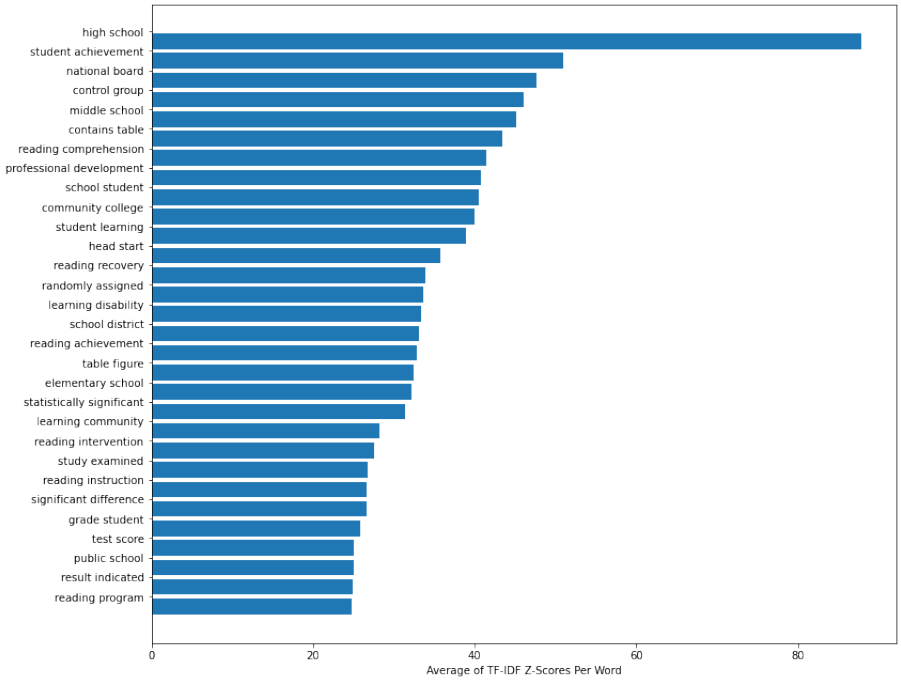

Fig. 3. Breakdown of Top Math & Reading TF-IDF Unigrams

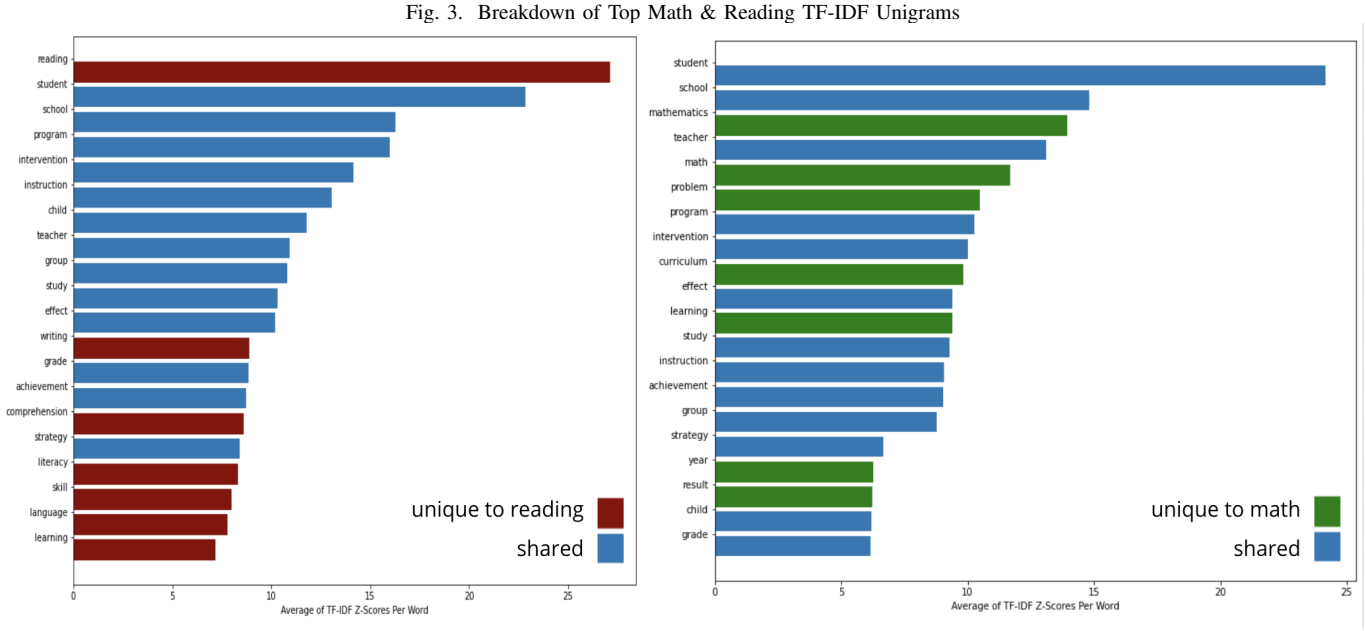

Fig. 4. Breakdown of Top Math & Reading TF-IDF Bigrams

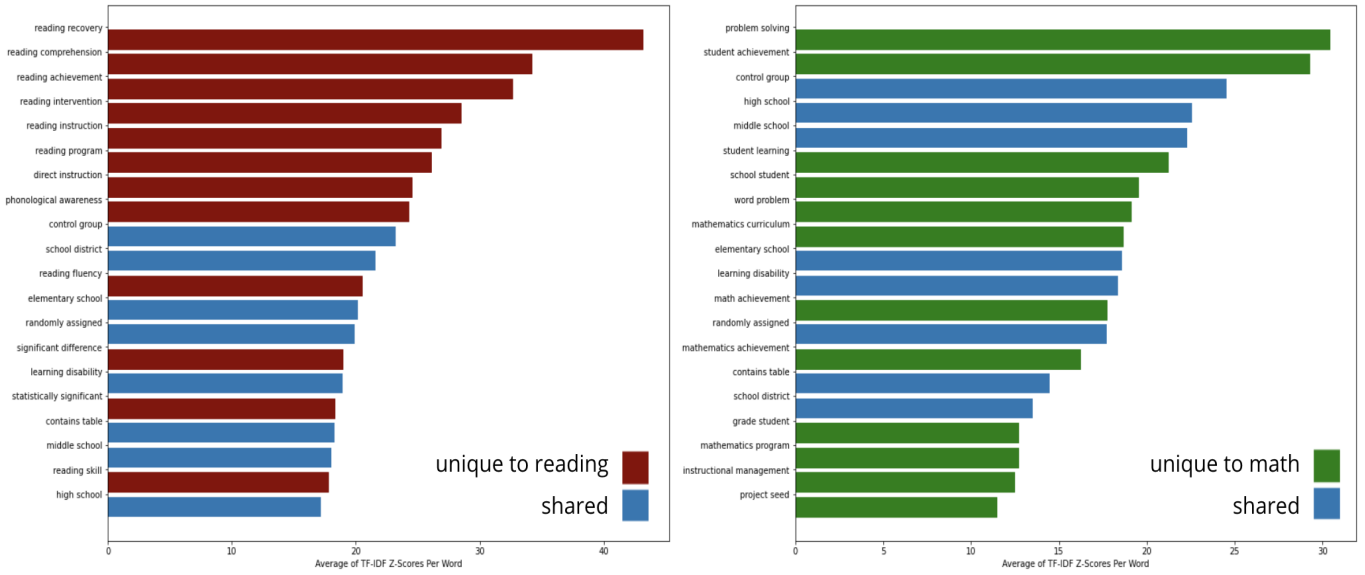

Fig. 5. Top 20 Bigrams in Topic 1 of the WWC Corpus

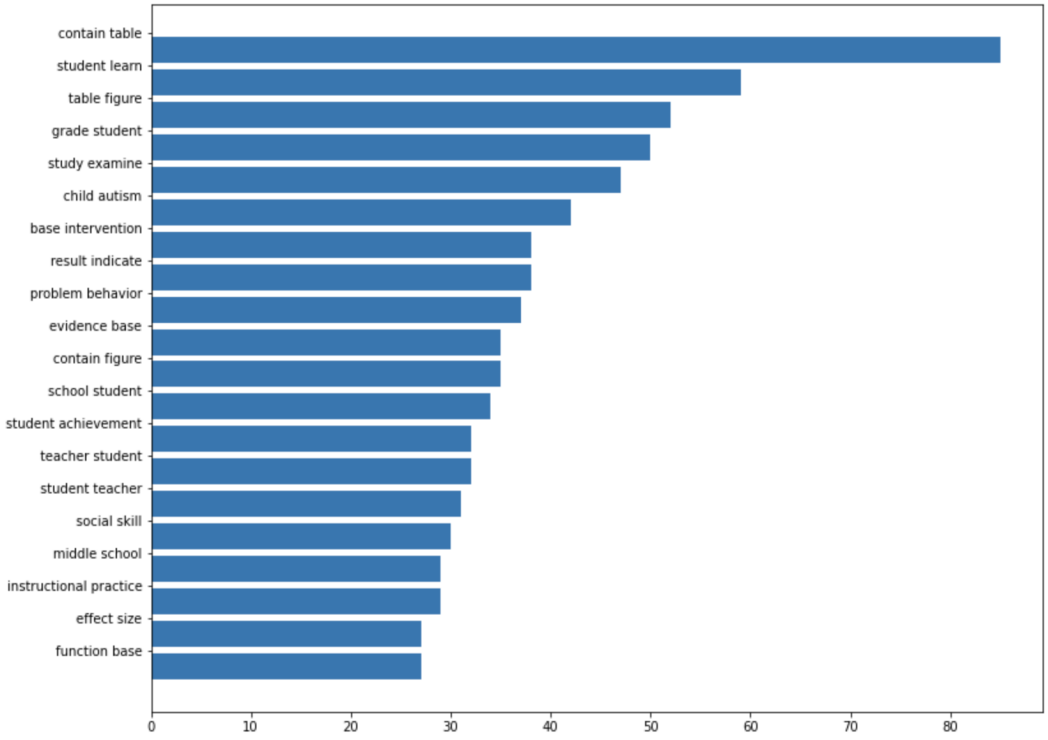

Fig. 6. Top 20 Bigrams in Topic 2 of the WWC Corpus

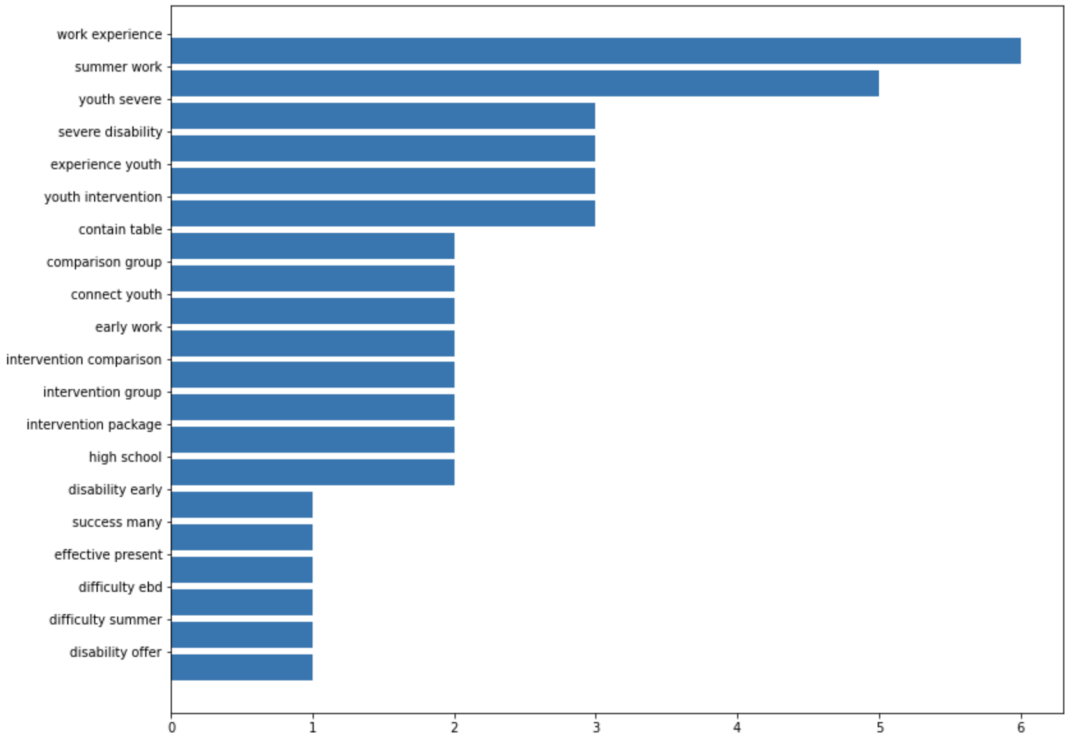

Fig. 7. Top 20 Bigrams in Topic 3 of the WWC Corpus

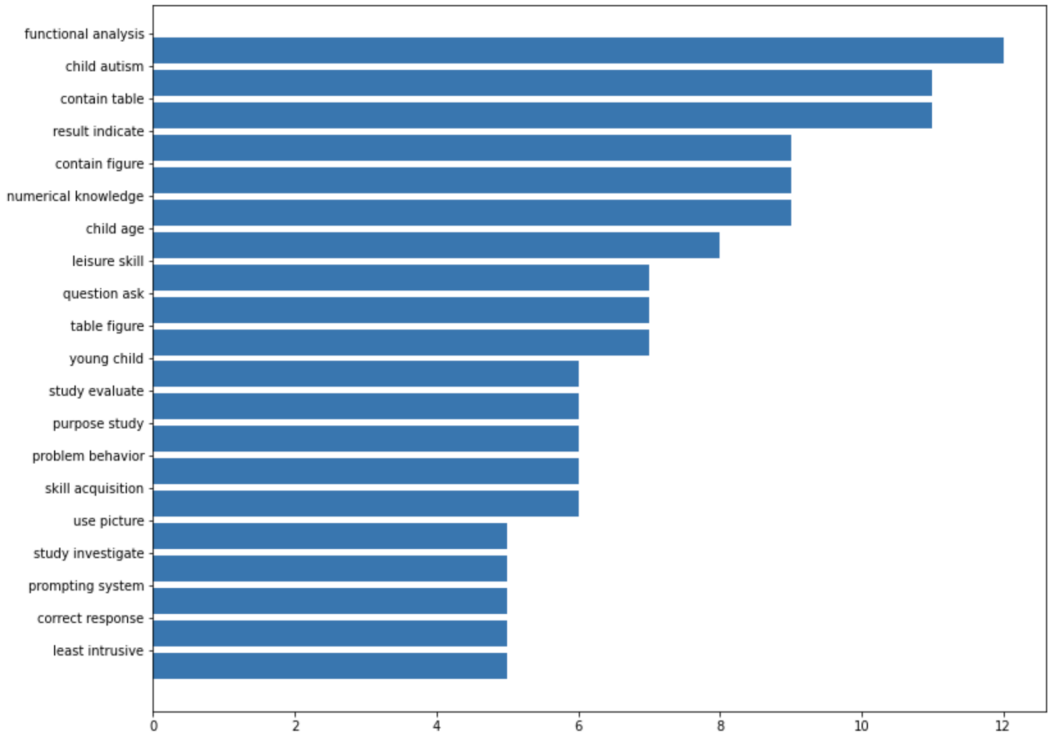

Fig. 8. Top 20 Bigrams in Topic 4 of the WWC Corpus

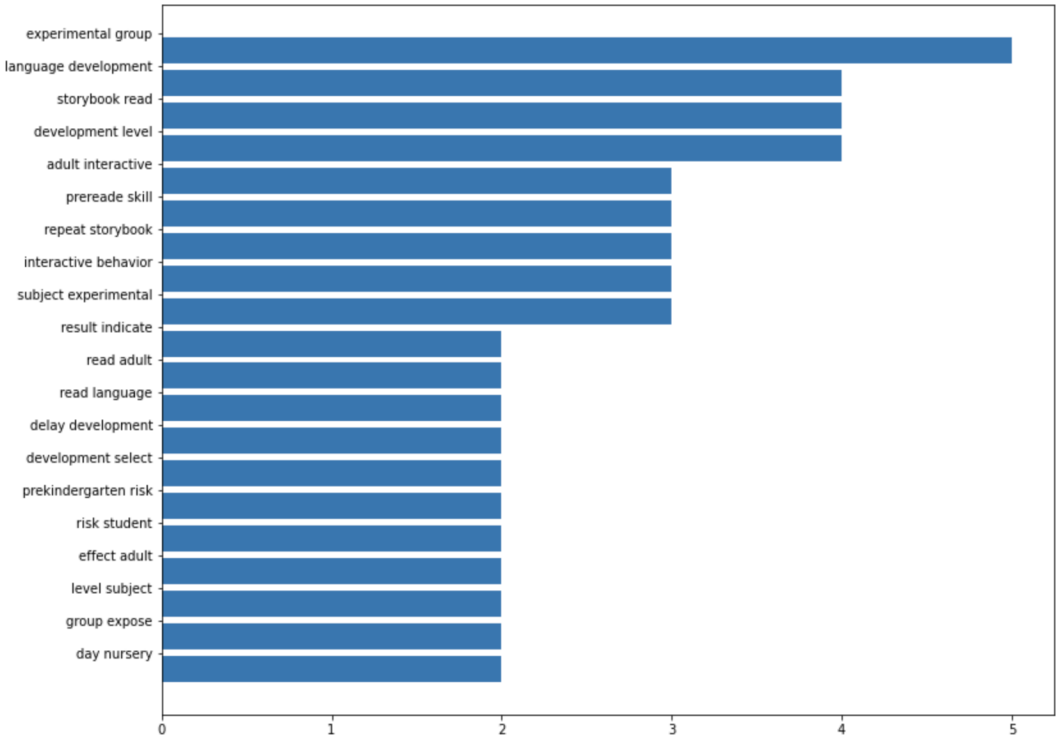

Fig. 9. Top 20 Bigrams in Topic 5 of the WWC Corpus

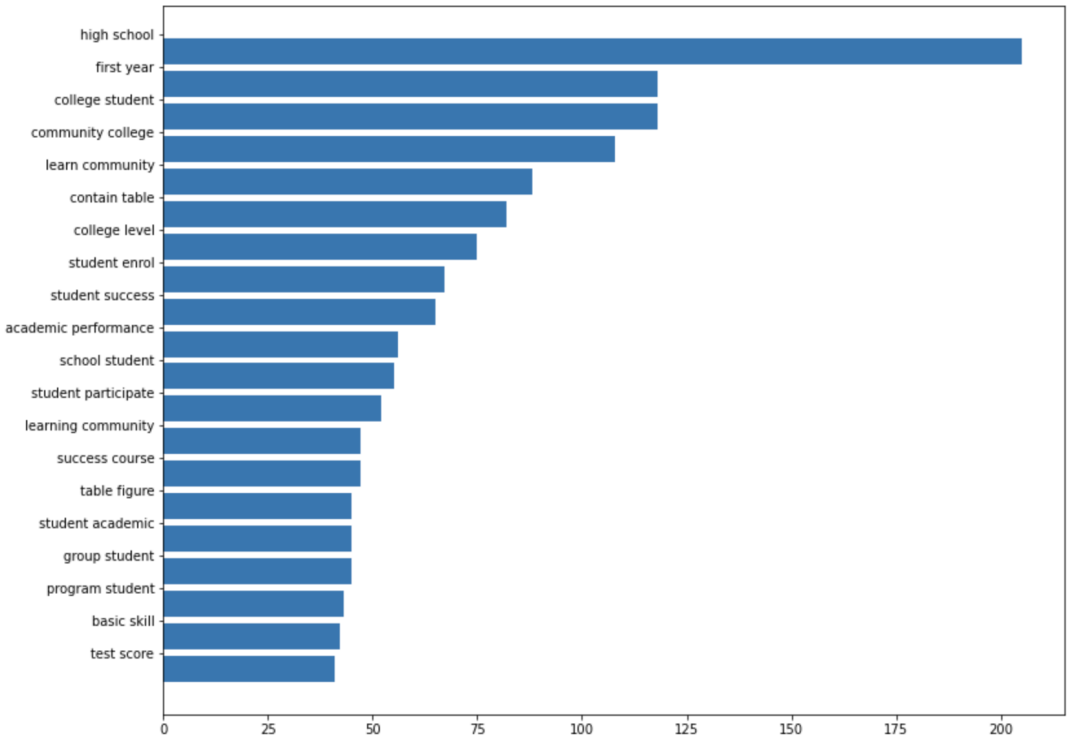

Fig. 10. Top 20 Bigrams in Topic 6 of the WWC Corpus

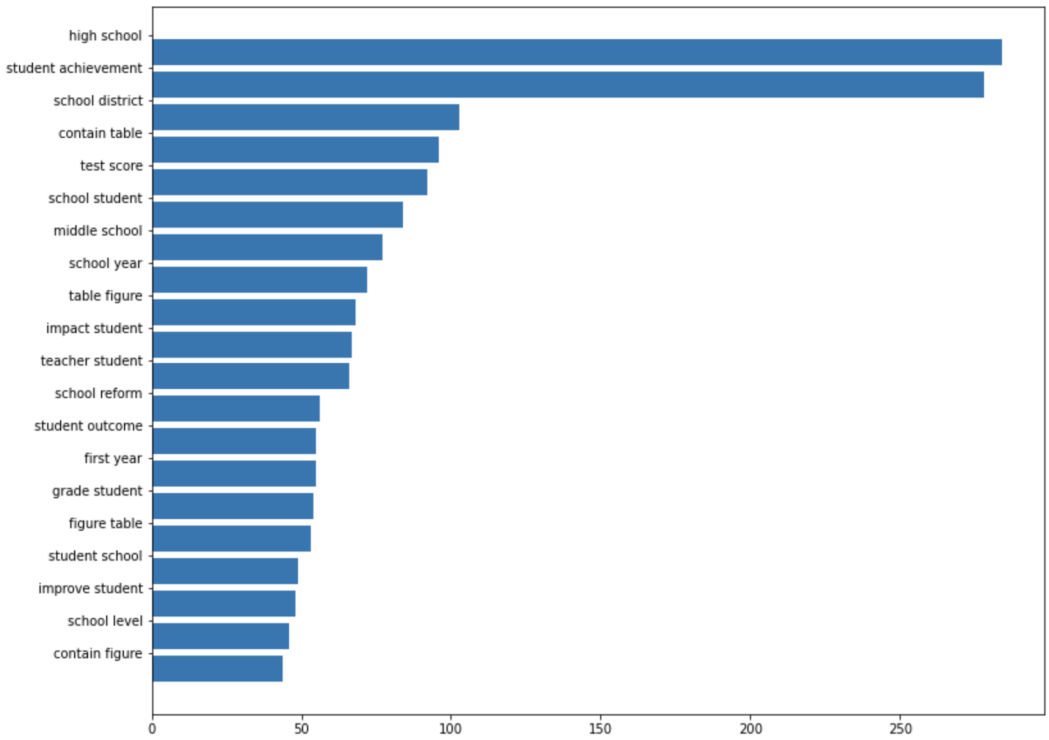

Fig. 11. Top 20 Bigrams in Topic 7 of the WWC Corpus

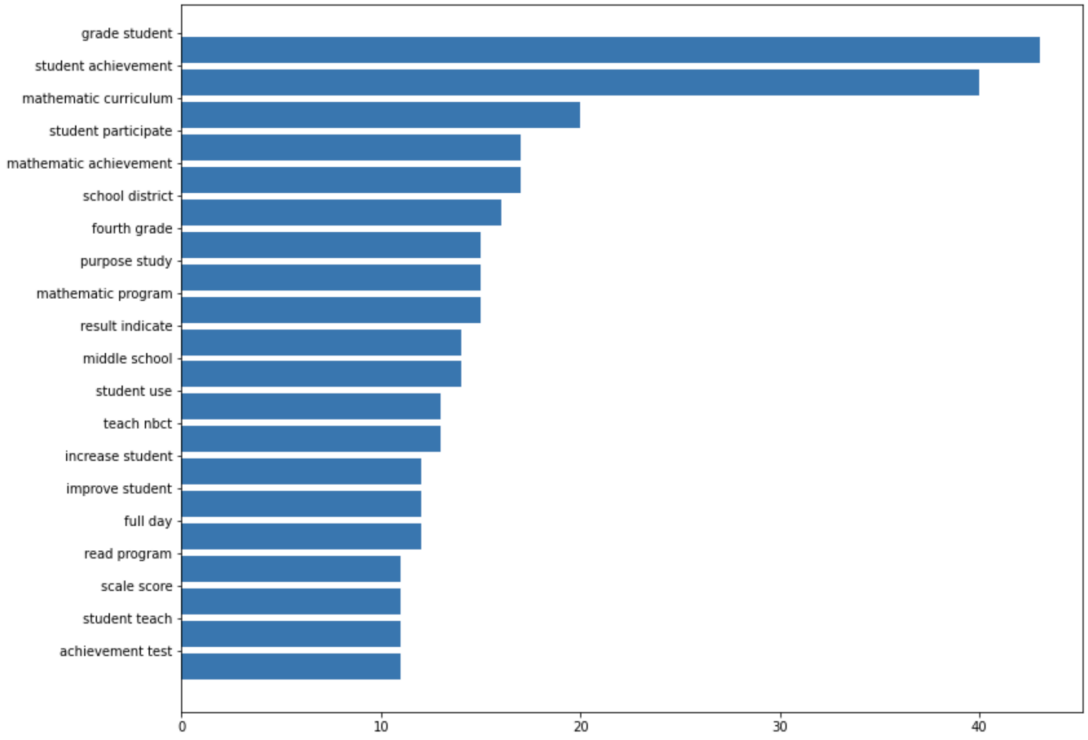

Fig. 12. Top 20 Bigrams in Topic 8 of the WWC Corpus

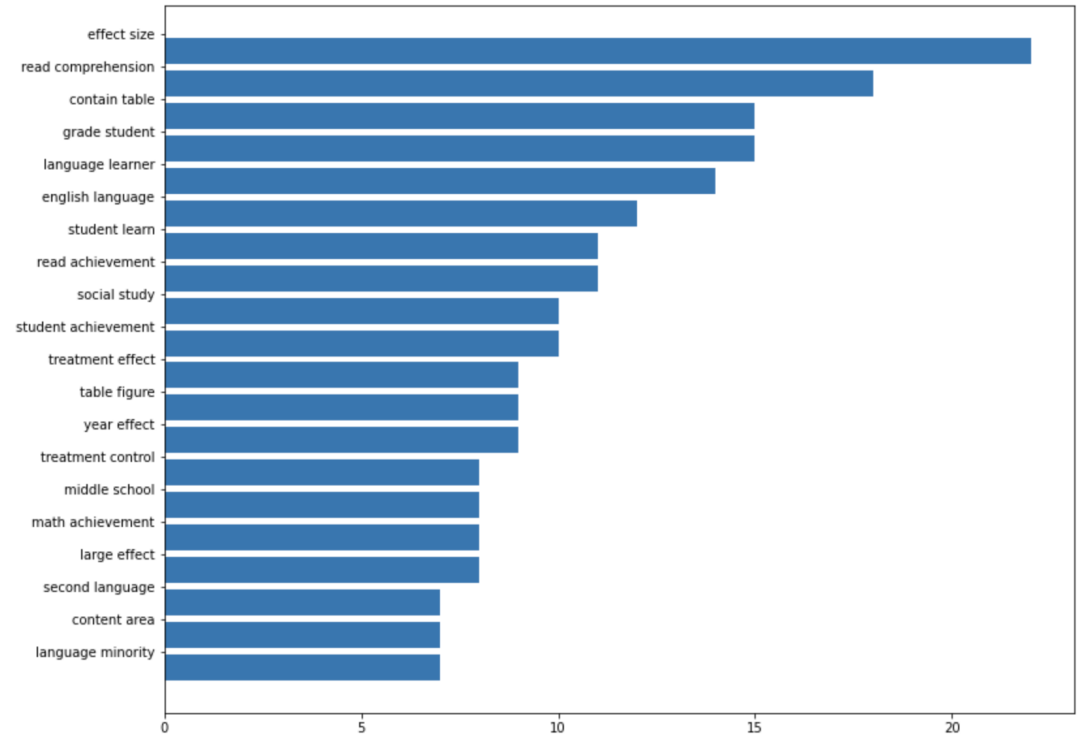

Fig. 13. Top 20 Bigrams in Topic 9 of the WWC Corpus

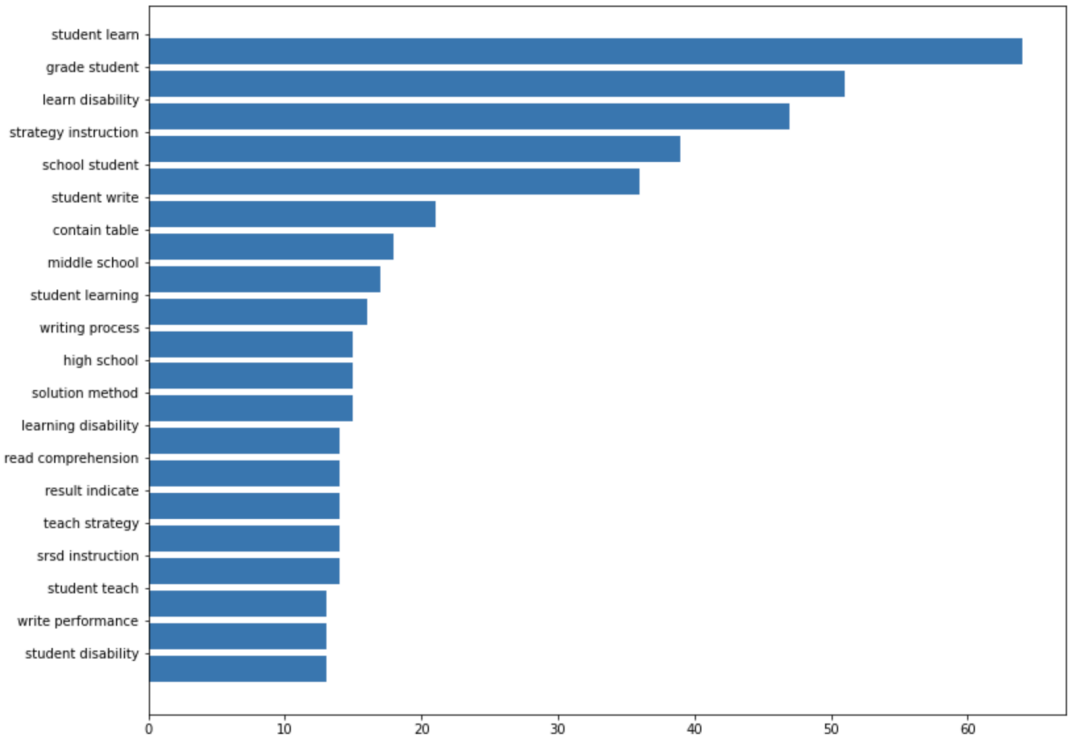

Fig. 14. Top 20 Bigrams in Topic 10 of the WWC Corpus

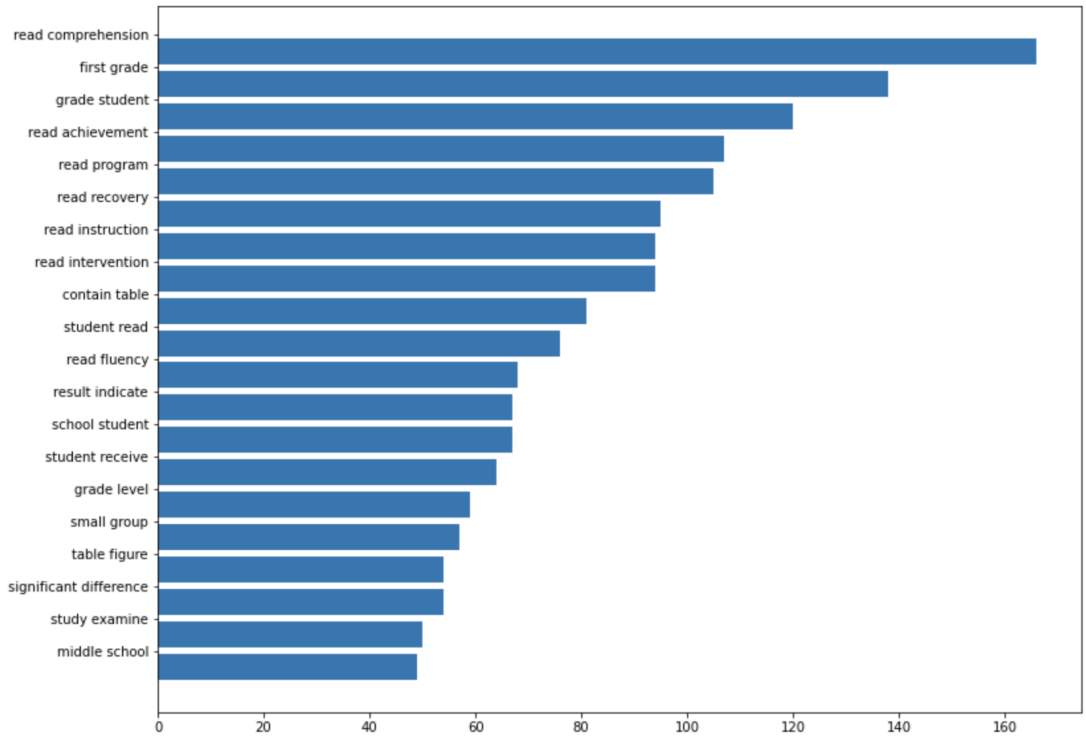

Fig. 15. Top 20 Bigrams in Topic 11 of the WWC Corpus

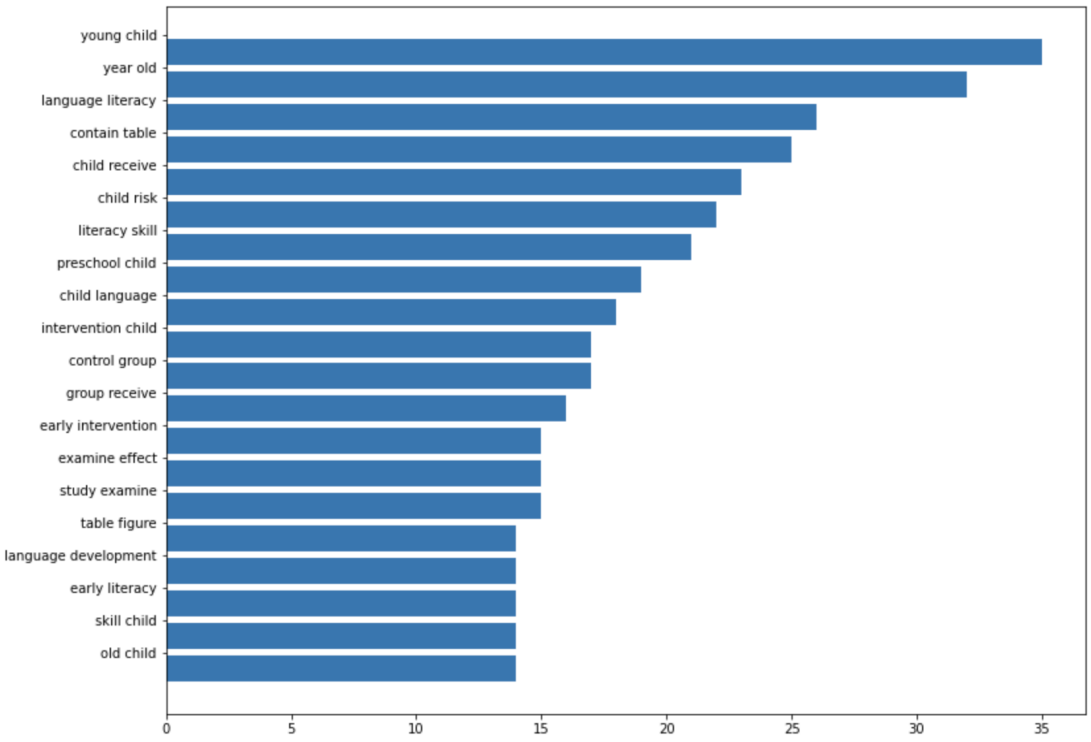

Fig. 16. Distribution of Number of Documents per Topic