TECHNICAL BRIEF

Using Artificial Intelligence (AI) to improve math accessibility for students with visual impairments

May 2022

Kang Xue and Elizabeth Barker

nwea

Suggested citation: Xue, K. and Barker, E. (2022). *Using Artificial Intelligence (AI) to improve math accessibility for students with visual impairments.* NWEA.

This work was supported by an AI for Accessibility grant from Microsoft.

Table of Contents

List of Tables

List of Figures

1. Abstract

This study, which is part of a larger project that aims to make online math more accessible to students with visual impairments (VI), examines the text quality of math assessment items for students with VI who use screen readers. Using data from about 29.5 million students taking standard versions of the MAP Growth math assessment, and 48,845 students taking accessible versions, we identified high-quality items, those that measured achievement for both students with and without VI equally well, and low-quality items, which showed differences between the two groups of students. The researchers introduced three word embedding methods and three classifiers to predict item quality for accessible assessments. This work advanced our understanding of barriers, and used cutting-edge technologies to develop a new way to better present math content online to improve accessibility and increase the opportunity to learn math for students for students with VI.

2. Introduction

While the use of screen readers, refreshable braille, and other technologies has made mathematics materials more accessible for students with visual impairments (VI), math content and assessment for students with VI, especially online, remains challenging, due to various barriers that cause many students with VI to lag behind their sighted peers in math achievement. The goal of this project is to create a more accessible math assessment for middle school students with visual impairments, and specifically, to create equation protypes that are accessible for students with VI who use screen readers. We used statistical methods and AI to lay the groundwork for an item-level analysis of MAP Growth math assessment data from students with VI. NWEA is uniquely positioned for this work because we have access to a large sample of assessment data both from students with and without visual impairments. Our team used the analysis point-biserial correlation coefficient (or point measurement) between students' achievement scores (i.e., RIT scores) and item response to determine whether items were quality or non-quality. This work advanced our understanding of barriers, and used cutting-edge technologies to develop a new way to better present math content online to improve accessibility and increase the opportunity to learn math for students for students with VI.

In order to accomplish our goal and create improved equation prototypes, our work included multiple phases: identifying quality items, conducting a literature review on barriers in math for students with visual impairments and blindness, training of AI with our quality and non-quality items, and creation of two equation prototypes informed by both prior research and the results of the AI analyses.

The Accessible MAP Growth assessment offers a partially-accessible online mathematics solution for grades 2-12 that includes the use of alternative text descriptions and excludes inaccessible item types such as items design with drag and drop functionality. While this test design provides some access, alternate text can also increase cognitive load. Empirical evidence reveals that students, especially those with VI using assistive technology, struggle with items that have more complex text, long descriptions, design barriers, and multiple parts to answer. To improve accessible assessments, we needed to understand which math skills are critical and which item designs present the fewest barriers. Phase one analyzed items from the Accessible MAP Growth 2-12 mathematics assessment to identify quality versus non-quality items. Quality items were considered those where both students with VI and blindness scored well. Non quality items were those on which students with VI and blindness did not do well,

while students without VI and/or blindness did well. Once quality versus non-quality items were identified, our team of experts analyzed the items for patterns of potential difficulties and barriers in non-quality items. The results of our analysis showed that math items that aligned to Common Core Standards, that are more complex (for example, those that include 2- or 4- step equations), on topics including place value or number value, and with a large amount of item information, were included in non-quality items, and so potentially pose barriers for students with VI.

Domain: Expressions and Equations
7.EE.B: Solve real-life and mathematical problems using numerical and algebraic expressions and equations.

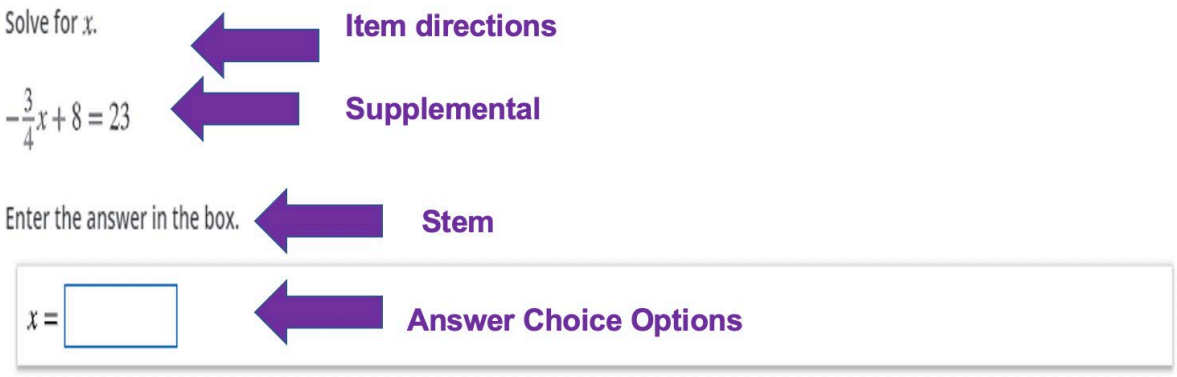

Figure 1. Sample item. Math items on an assessment may include a stem as well as equations, figures, or other content, and for multiple choice tests, include answers from which students select. In this project, we focused on equations on assessments for students who use screen readers to access the items' content.

In the second phase of our work,[i] we reviewed evidence from previous research on this subject. This literature review revealed a potential concern that equations given auditorily increase cognitive load. Da Paixão Silva et al. (2017) explored the amount of time and effort it took BVI users to comprehend information about a quadratic equation.[ii] Students were tasked to find out if this equation had two identical real roots, if it opened up or down, and if it was complete. The fastest time recorded was 132 seconds, and the slowest 235 seconds. It is important to note that the participants in this study were chosen for their familiarity with the math concepts, and so the amount of time taken was not correlated to a lack of understanding. The length of time a student takes to answer an item can indicate complexity, and a long response time can demonstrate that the item itself may create too much cognitive load. The Logan Project[iii] created a method known as process driven math (PDM) that specifically addressed cognitive load difficulties for students with visual impairments and for students who access math primarily through their sense of hearing. Instead of voicing a math problem in one long chunk of characters, the human reader/scribe would break down a math problem into pieces according to its mathematical vocabulary. Pairing these key insights from the literature with our AI analysis, we created two, 4-step equations that allow students to use their screen readers and keyboard navigation to dive deep into both the left and the right side of the equations using regions. Like

PDM, this approach permitted students to access any part of the equation, rather than listening to an entire, long, equation read aloud to pick out the part they wanted to reference.

Further AI analysis of our items will help to support our final steps of productionizing our prototypes. The following analysis describes the steps we have and are taking to utilize AI and natural language processing (NLP) to support math access and learning for students with visual disabilities and blindness.

3. Methods

3.1. Data and item labeling

To train an NLP-based classifier, the text quality of a math assessment item needs to be determined. Given a student with visual impairments has sufficient mathematical knowledge to solve the mathematical problem embedded in the item, the text quality of the item is the degree to which the text is accessible to the student. If an item text is high quality, students have similar performance (or correct response rate) on this item in the accessible tests as students with a similar academic level in non-accessible tests. If an item text is low quality, students' performance in accessible and non-accessible assessments is not consistent.[1]

To determine text quality for screen readers in the accessible test, we used two datasets in this research: accessible test data and non-accessible test data from students in grades 6 through 8. The data included results from around 29.5 million students taking standard versions of the MAP Growth math assessment, and 48,845 students taking accessible versions.

3.2. Item labeling using item diagnostic quality

To label the quality of an item text for screen readers, we compared each item's diagnostic qualities in the accessible and non-accessible tests. The point-biserial correlation coefficient is widely used in psychometrics to evaluate an item's diagnostic quality. The point-biserial correlation coefficient[iv] is used to determine the relationship between two variables when one of them is dichotomous. Two point-biserial correlation coefficients for each item were calculated between students' RIT score and item responses for a given item. A high point-biserial value reflects that the item distinguishes students with high RIT scores from students with lower RIT scores. Items with fewer than 50 responses in the accessible test were dropped to make the calculated correlations robust.

From the histogram of the items' point-biserial correlations (or point measurement, shown in Figure 2), we observed that the distribution of items' point-biserial correlation in the accessible and non-accessible tests are similar, and most items' coefficient values fall between 0.2 to 0.4, the reasonable range of the point-biserial correlation in real assessments. The observation indicates that most items are doing well to distinguish between students with high and low math achievement levels.

In assessment, high item difficulty impacts the point-biserial correlation (in that very hard items decrease the value of correlation coefficients), so we also introduced the item difficulties to

evaluate the item diagnostic quality: for the difficult items (difficulty over 250 RIT, the 75% quantile of item difficulty in the item pool), if the correlation coefficient value is less than 0.15, the item was classified as low diagnostic quality; for other items (difficulty less than or equal to 250 RIT), if the correlation coefficient value is less than 0.25, the item was classified as low diagnostic quality. Based on this criterion, in non-accessible tests, items can be divided into a high diagnostic quality group and a low diagnostic quality group. In our research, we only kept the items belonging to the high diagnostic quality group.

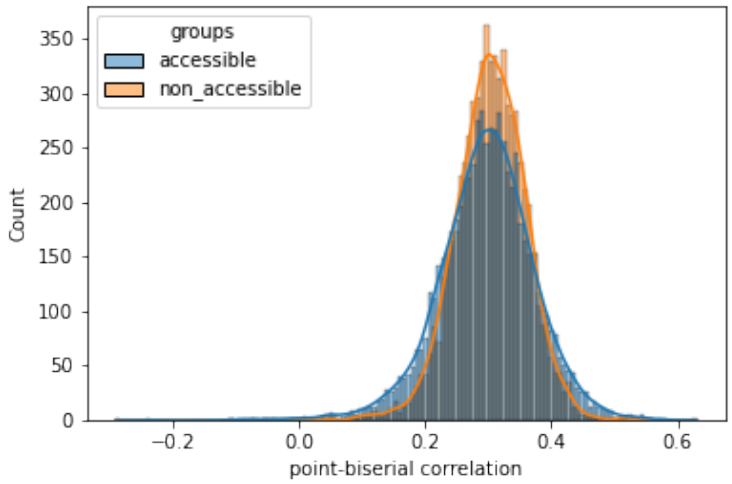

Figure 2. Point-biserial correlations for items from accessible and non-accessible math assessments. This shows a global view that the accessible test contains items that have a similar correlation distribution as the non-accessible test.

An item's performance is consistent if its two-point measurement in accessible assessment and non-accessible assessment are close in value. We used Z-scores to detect the outliers of the absolute differences between the two-point measurements of each item. The outliers indicate items on which the point measurement differences between non-accessible and accessible assessments are more significant than most items. The scatter plot of the point measurements in the accessible and non-accessible assessments is shown in Figure 3. If the Z-score is over 3, the item is an outlier (shown as orange points in Figure 3). The figure shows that the point measurements in the accessible assessments and non-accessible assessments hold a linear relationship. The x-axis is the point measurement in accessible assessment, the y-axis indicates the point measurement in non-accessible assessment, and the orange points are the outliers.

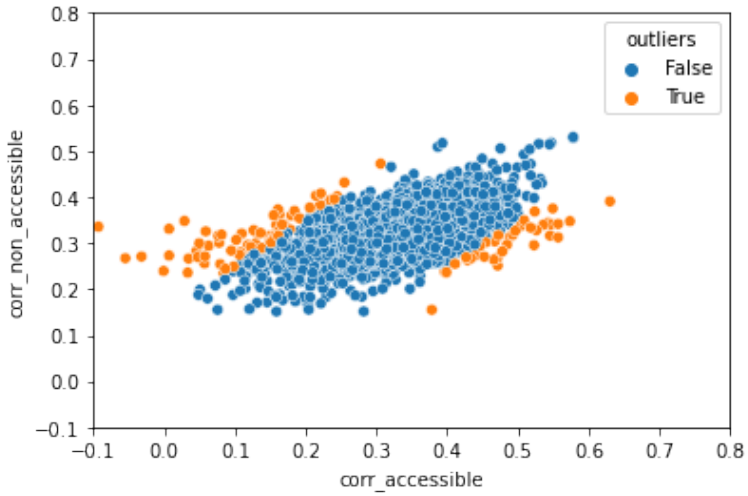

Figure 3. Scatter plot of point-biserial correlations for items from accessible and non-accessible math assessments.

After the outlier detection, we finally classified the items into two groups: a *consistent* item group and *inconsistent* item group. The items in the first group have consistent point measurement in both accessible assessments and non-accessible assessments, while the items in the second group do not have consistent point measurement in the two assessments. We designed a machine learning method to detect if a new item belongs to the consistent group or inconsistent group using the item's content information. Of the 3,557 items in our sample, the number of items in the first group (consistent items) is 3,492, with 65 items in the second group (inconsistent items) Based on our criteria, most of the items (98.2%) with high quality in non-accessible assessments hold a consistent quality in the accessible assessments.

3.3. NLP-based classifiers

Natural language processing (NLP) is a technique to automatically manipulate natural language, like speech and text, using computer programs. This research used the NLP-based method to build a classifier to determine if item text is high quality in the accessible tests. The framework diagram (Figure 4) shows the three steps in our method: pre-processing, feature extraction (or word embedding), and classifiers.

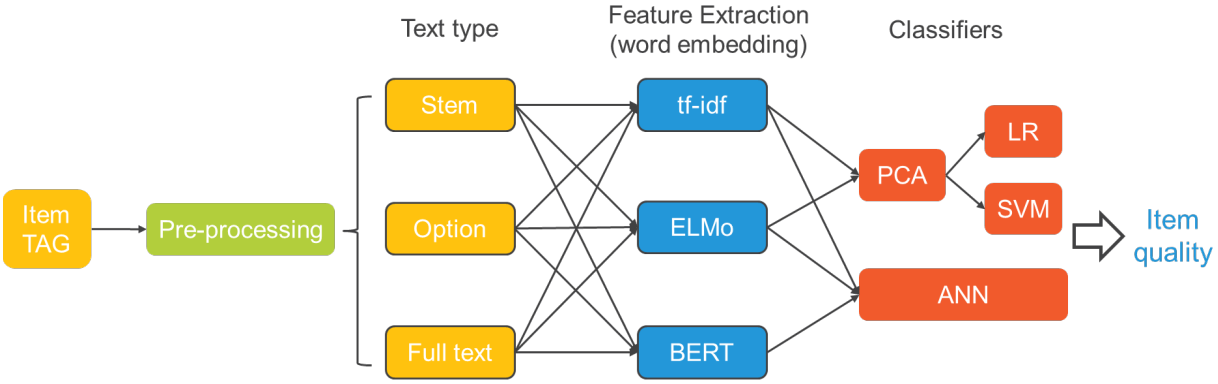

Figure 4. Natural language processing framework for math item quality classification.

We first applied some preprocessing procedures to the item text (the text TAG for the screen reader) by removing punctuation and stopping words (e.g., a, an, the). Preprocessing is widely used in NLP to remove redundant information from the raw text data. Then we stratified the whole dataset into the training dataset and testing dataset based on consistency and inconsistency, with 70% of the sample in the training dataset, and the testing dataset containing the remaining 30% of the sample.

After that, we used word embedding, one essential technique in NLP, to convert string values in the item text to numeric values. In this step, we used three different word embedding methods:

1. Term frequency-inverse document frequency[v] (tf-idf): a statistical measure that evaluates how relevant a word is to a document in a collection of documents. Tf-idf is easy to compute and compares the similarity between two documents. Since tf-idf is based on bag-of-word, it does not capture position in text, semantics, co-occurrences in different documents.

2. ELMo word embedding[vi]: a new deep learning-based technique for embedding words into real vector space using bidirectional LSTMs trained on a language modeling objective. The advantage of ELMo is concatenating the left-to-right and right-to-left information in the text; however, it limits the ability of the representation to take advantage of both contexts simultaneously.

3. BERT word embedding[vii]: a deep learning-based technique for a “masked language modeling” object. To achieve this object, during the training procedure, BERT randomly masks a small proportion of the sentences (e.g., 15%) and predicts the masked text using the words surrounding them. In contrast to ELMo, BERT uses transformers instead of Bidirectional LSTM to achieve a better understanding of the context.

In feature extraction, we applied these three word embedding methods to three text types for each item: item stem, item options, and full item text. There are nine types of word embedding features for each item in our research.

Then, we started to train machine learning classifiers using the numeric vectors of the training dataset from word embedding. This step used three classifiers in machine learning:

1. Principal Component Analysis[viii] (PCA; [5]) plus Logistic Regression (LR): the high-dimensional word embedding vectors are first transformed into a low-dimensional vector using PCA to avoid the curse of dimensionality, then using the low-dimensional vector and labels to train the LR model;

2. Principal Component Analysis[viii] (PCA) plus Support Vector Machine (SVM): in contrast with LR, the SVM is a clear and more powerful way of learning complex non-linear functions;

3. Artificial Neural Networks[ix] (ANN): ANN combines dimensional reduction and classification in a single framework in which all parameters can be updated simultaneously during training; in addition, ANN approximates statistical distribution without a specific mathematic equation.

4. Experimental results

To test the performance of the proposed methods, we first, divided the data into three parts: a training dataset, a validating dataset, and a testing dataset. The training dataset was used to train the designed classifier, while the testing dataset was used to evaluate the trained models. To avoid overfitting in training classifiers, for logistic regression, the 5-folded cross-validation was used to choose the best intercepts and coefficients; for ANN, validation data was used to stop training and output model with the best performance on prediction in the validation test. The testing dataset contained 30% of the whole dataset. The training dataset contained 70% of the whole dataset. In logistic regression and SVM, the validating dataset was 20% of the training dataset in cross-validation, and in ANN, 10% of the training dataset was used for validating in early stopping.

As previously mentioned, the proportion of the low-quality items in the whole item pool is very low (less than 2%). The extremely imbalanced dataset results in models that have poor predictive performance, specifically for the minority class (i.e., the low-quality items). To deal with this issue, we introduced two methods. For PCA plus LR and PCA plus SVM, we used the oversampling method to increase the sample size of the low text quality group by randomly duplicating the observations in the low text quality group; for ANN, we weighted the loss for the two groups in the training procedure based on the sample size of each group.

Due to the imbalanced data, we also introduced two metrics, Area under the ROC Curve (AUC) and Geometric mean (G-mean), to evaluate the proposed methods. An ROC curve plots True Positive Rate (TPR) versus False Positive Rate (FPR) at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. AUC is the probability that the model ranks a random positive example more highly than a random negative example. G-mean is the root of the product of precision and recall. It measures the balance between classification performances on both the majority and minority classes. A low G-Mean indicates a poor performance in the classification of the minority group (positive cases) even if the majority group (negative cases) are correctly classified.

Both AUC and G-mean lay between 0 and 1, in which 1 means all items are correctly classified, and 0 means all items are incorrectly classified. The AUC is sensitive to classifying imbalanced data, and a low G-Mean indicates poor performance in classifying the minority cases even if the majority cases are correctly classified. If a classification method’s AUC and G-mean are closer to 1, the method has a better performance.

In the proposed methods, there are some hyperparameters that needed to be tuned. For PCA+LR and PCA+SVM, we tuned the number of components in PCA. For BERT+ANN, we tuned the structure of ANN and the max text length of BERT tokenizer. The tables in this paper only show the best performance of each proposed method.

Table 1. The best AUC and G-mean of the proposed methods using different word embedding methods

Text type	Word Embedding	Classifier	AUC	G-mean
Full text	TF-IDF	PCA+LR	0.670	0.669
stem	ELMo	PCA+SVM	0.710	0.708
Full text	BERT	ANN	0.730	0.727

From the general performance table above, the BERT+ANN method achieved the highest AUC and G-mean compared with other methods. We also compared the performances when inputting different text types for each classifier. The summary of performance is shown in Table 2.

Table 2. Summary of classification performance of different word embedding methods

Word Embedding	Text type	AUC	G-mean
TF-IDF	stem	0.616	0.605
	option	0.638	0.623
	Full-text	**0.670**	**0.669**
ELMo	stem	**0.710**	**0.708**
	option	0.643	0.642
	Full-text	0.702	0.700
BERT	stem	0.652	0.650
	option	0.622	0.610
	Full-text	**0.730**	**0.727**

From Table 2, we observe:

1. When inputting full text, the tf-idf achieved the best performance (i.e., AUC=.670 and G-mean=.669). This is because tf-idf uses the whole words in the items to generate the bag-of-words and create the numeric vector for each item. Full text can help to increase the number of words and distinguish the differences between items.

2. For ELMo, using the stem as input resulted in the best performance. The reason is that ELMo's bi-directional LSTM is trained to predict the next words in the content. This means that the structure of ELMo is good at having sequential information of a context. In contrast to options or full text of an item, the stems are constructed as a whole piece of paragraph for students to understand.

3. For BERT, like tfi-idf, the inputs of full text leads to the best performance. This is because BERT is trained through masked language modeling, which helps BERT generate more accurate word representation. In addition, BERT can handle larger and more complex context compared with ELMo and tf-idf.

5. Conclusion and discussion

In this research, we introduced three word embedding methods and three classifiers to predict item quality for accessible assessments. To solve the extremely imbalanced dataset, we used two methods to increase the robustness of the proposed methods. Generally, the experimental results show that the two deep learning-based word embedding methods (i.e., ELMo and BERT) achieved better classification performance than the classic frequency-based word embedding method (i.e., tf-idf). Meanwhile, we also observed that the same word embedding methods' performance on different item text differed. ELMo works better on item stem, but tf-idf and BERT performed better using the full text of an item.

It is important to note that this study used an accessible test, in which items had already been examined by content experts to remove items that were deemed inaccessible for students who were visually impaired. This curation of items is one likely reason that such a large number of items on the accessible test were high quality for these students. However, such curation can be effort intensive and expensive. NLP methods may provide a useful way to test new items as they are developed to identify quality items. Additionally, in this study, we examined item content and TAG, but we did not examine other item metadata (for example, instructional area, item type, or item response time). Analyses using these other data may also provide important insight into potential barriers in how math content is presented to and accessed by students with visual impairments, so that we may better design supports to make access more equitable and support math learning for students with VI.

6. References

[i] Steinbach, S. (2022). Looking beyond vision: Supports for students who are blind or visually impaired in mathematics. NWEA.

[ii] da Paixão Silva, L. F., de Faria Oliveira, O., Freire, E. R. C. G., Mendes, R. M., & Freire, A. P. (2017). How much effort is necessary for blind users to read web-based mathematical formulae? A comparison using task models with different screen readers. *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, 1–10.

[iii] Gulley, A. P., Smith, L. A., Price, J. A., Prickett, L. C., & Ragland, M. F. (2017). Process-driven math: An auditory method of mathematics instruction and assessment for students who are blind or have low vision. *Journal of Visual Impairment & Blindness*, *111*(5), 465–471.

[iv] Linacre, J. M., & Rasch, G. (2008). The expected value of a point-biserial (or similar) correlation. *Rasch Measurement Transactions*, *22*(1), 1154.

[v] Linacre, J. M., & Rasch, G. (2008). The expected value of a point-biserial (or similar) correlation. *Rasch Measurement Transactions*, *22*(1), 1154.

[vi] Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019, June). Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* (pp. 15-18).

[vii] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[viii] Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37-52.

[ix] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.