Is It Fair? Automated Open Response Grading

John A Erickson
Worcester Polytechnic Institute
jaerickson@wpi.edu

Anthony F Botelho
Worcester Polytechnic Institute
abotelho@wpi.edu

Zonglin Peng
Independent Researcher
zpeng@wpi.edu

Rui Huang
Independent Researcher
rhuang2@wpi.edu

Meghana V. Kasal
Independent Researcher
mkasalvinayakuma@wpi.edu

Neil T Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

ABSTRACT

Online education technologies, such as intelligent tutoring systems, have garnered popularity for their automation. Whether it be automated support systems for teachers (grading, feedback, summary statistics, etc.) or support systems for students (hints, common wrong answer messages, scaffolding), these systems have built a well rounded support system for both students and teachers alike. The automation of these online educational technologies, such as intelligent tutoring systems, have often been limited to questions with well structured answers such as multiple choice or fill in the blank. Recently, these systems have begun adopting support for a more diverse set of question types. More specifically, open response questions. A common tool for developing automated open response tools, such as automated grading or automated feedback, are pre-trained word embeddings. Recent studies have shown that there is an underlying bias within the text these were trained on. This research aims to identify what level of unfairness may lie within machine learned algorithms which utilize pre-trained word embeddings. We attempt to identify if our ability to predict scores for open response questions vary for different groups of student answers. For instance, whether a student who uses fractions as opposed to decimals. By performing a simulated study, we are able to identify the potential unfairness within our machine learned models with pre-trained word embeddings.

Keywords

Natural Language Processing, Unfairness, Deep Learning, Word Embeddings, Pre-Trained Word Embeddings, Simulated Study

1. INTRODUCTION

In recent years, natural language processing (NLP) has been at the forefront of machine learning in multiple fields. Linguistics provides another source of information outside the standard data from user logs. Instead of relying on correlational assumptions from this data, inferences can be deduced directly from the users linguistics. While utilizing linguistics in education isn't genuine, modern machine learning and natural language processing has helped to automate the analysis and provides effective tools for learning.

The development of more advanced deep learning has brought a deeper semantic understanding of words within these linguistical models. The emergence of word embeddings were an important development in machine learning and NLP, but the publishing of publicly available pre-trained word embeddings, such as such as GloVe [21] or Wikipedia or Word2Vec [19], provided researchers with a powerful tool for optimizing algorithms with linguistics. While word embeddings were powerful for studies within areas such as MOOCS (i.e [14] [20]), smaller studies, with less robust linguistic data, were unable to utilize this modern approach for semantic relationship of words.

Since research has shown that some of the semantic meanings inferred from pre-trained word embeddings can elicit undesirable biases [2], the major question then becomes, does this underlying bias suggest the algorithm or predictive model will make unfair decisions? For instance, if an algorithm utilizes linguistics and NLP with pre-trained word embeddings will the predictions be unfairly made from those underlying biases. Our research attempts to explore:

1. Whether, through 3 simulated studies, the format a student writes an answer (i.e. fractions vs. decimals) effect the scoring model and potentially elicit unfair scoring?
2. What effect, through 3 simulated studies, if any, do *'distractor'* words have on the unfairness?
3. Whether or not underlying bias in pre-trained word embeddings can lead to unfairness in open response scoring models in middle school mathematics?

2. BACKGROUND

2.1 Intelligent Tutoring Systems

In recent years, online educational technologies have been on the forefront of learning for students. A common online educational technology, intelligent tutoring systems (ITS) [4],

has been prevalent in education for many years. Some of the most common ITS are ASSISTments[11], McGraw Hill's ALEKS™ and/or Carnegie Learning's Cognitive Tutor™. Through the use of both machine learning and software engineering, these systems have been shown to be effective at increasing the scores of students with end of the year standardized math exams[25] and the effects of their intelligent tutoring closely resembles the effect face to face tutoring has on students[31]. Other ITS, such as AutoTutor[9],have attempted to resemble the face to face tutoring more directly by developing automated conversations and dialogues between students and ITS [9]. However, most of the support and benefits of these ITS have been limited to questions with structured answers (i.e. multiple choice or fill in the blank questions).

2.2 Automation of Intelligent Tutoring Systems

Automated support of ITS is a draw for many teachers; one study noted that many utilize multiple choice questions for the efficiency and accuracy of grading [26]. However, since most of the automation is limited to questions with structured answers, the content which teachers provide is limited. Studies have looked to utilize NLP to automatically evaluate work or questions which require a student's unique linguistics (i.e. open response questions, or essays) including [29][28][24][1][7][30][17]. While most of this research has been primarily focused on content outside of mathematics, our previous research, [6], looked to help teachers diversify the content which they provide students in middle school mathematics by utilizing traditional and modern NLP to develop an automated scoring model for open response middle school mathematics questions. A more diverse set of question types can be beneficial to students and can elicit differing levels of cognition, as studies [18][15] note.

2.3 Natural Language Processing

Towards the goal of automating open response questions, or any linguistical/NLP prediction task, the major task is in how to numerically represent words thus that a machine learned algorithm can generate an accurate prediction. One of the more simplistic approaches utilizes the frequencies of each unique word within the corpus, whats commonly known as a *Bag of Words* approach. While undoubtedly easy to interpret and not computationally intensive, this approach has been utilized in studies such as [13] and is the foundation of more advanced approaches such as[27][10][22]. This has evolved into utilizing deep learning to generate word embeddings such as GloVe[21] and Word2Vec[19].

2.4 Pre-Trained Models

Embeddings are only as powerful as the data their train on. Not all researchers have robust corpuses, thus embeddings can be misleading. Pre-trained embeddings, such as GloVe or Word2Vec, publish their own embeddings generated from Wikipedia and GoogleNews. As these pre-trained word embeddings have grown in popularity, word embeddings have expanded to utilize bidirectional encoder representations from transformers (referred to as BERT[5]) to create pre-trained word embeddings, as well. Similarly, this has evolved from word level embeddings to pre-trained sentence and document level embeddings[23][3][16].

2.5 Fairness

When it comes to linguistics, the way someone speaks, the way someone articulates can be unique to themselves. Similarly, the way someone writes is personal to themselves and specific to their topic. So when algorithms are being pre-trained on data which isn't the researchers own data, there are questions to be asked. Research[2], has been able to identify some potentially harmful semantic relationships present in common pre-trained word embeddings. For instance, [2] was able to identify that Google's pre-trained Word2Vec on GoogleNews elicited some harmful stereotypes. There in lies the important question, if we omit variables which could cause unfairness in the automated scoring, are we continuing to avoid unfairness if we utilize pre-trained word embeddings. To identify this, we utilized Absolute Between-ROC Area (ABROCA) [8].

3. STUDY 1: SIMULATION STUDY

This research developed a simulated study to attempt to identify if pre-trained word embeddings are utilized within an automated scoring model for open response answers, do they influence the model to make unfair predictions. An example of this would be if a group of students state their answer with a fraction and surrounding text, does the predictive model generate scores similarly for those students that use decimals along with surrounding text? Through this simulated study, we are able to gain a deeper insight into what/if any unfair scoring occurs when utilizing the pre-trained GloVe word embeddings trained on Wikipedia between groups.

There are 3 studies within this simulated study to help achieve this goal. First, we develop answers which contain differing distributions of answers which contain fractions and decimals and generate the ABROCA value at the differing distributions. Second, we attempt to see if decimals and fractions alone generate differing ABROCA values. Third, we attempt to see if we replace decimals in the text with unknown tokens (more reliance on distractor words), do the ABROCA values differ at differing distributions? These studies will help provide deeper insight into the potential unfairness an automated scoring model can be producing when utilizing pre-trained word embeddings

3.1 Data Generation

At the foundation of this simulated study is the generation of the *student* dataset.The generation was split into two facets, the training dataset student answers and the test set student answers. This was performed such that the model would not be able to have any identical answers between the training set and the test set. Essentially, that the predictions aren't being made because the model has already seen that exact series of embeddings associated with a certain score.

Towards simulating authentic student answers, the generation of the corpus was founded on the goal of utilizing random selection. For the training set, as Table 2 shows (see Appendix A), there are 4 different length student answers in this corpus. There are answers which are 6, 5, 4 and 3 word length answers. The generation of the student answers can be surmised into 4 steps and visualized with Table 2 (see Appendix A). First, select whether it will be a student answer which uses decimals or fractions. Then,

randomly select what length the answer is. Once a length is randomly selected, another random selection is made between the two structures (i.e. 'Structure' within Table 2 in Appendix A). Finally, randomly select text from **Fill "1"** and **Fill "2" Fractions** or **Fill "2" Decimal** to fill the identifiers **'1'** and **'2'**.

This is the same approach that is utilized within the test set corpus generation as well. Table 3 (see Appendix A) shows that however there are different structures and phrases to select from than the training. Allowing for variability between test and training; thus, guaranteeing that a answer structure used in training isn't the same as in the test. From this, we can ascertain the model's predictions were not only based on identical phrases it has seen.

3.2 Methodology

This study sets out to identify if an automated scoring model for open response questions, which utilizes pre-trained word embeddings, elicit unfair scoring. With the answers generated, we set out to sample these simulated answers such that there is a balance of student answers which utilize fractions and decimals. The training set is comprised of student answers drawn from the pool of simulated answers containing fractions (considered as *Group A*), as well as answers sampled from similar student answers containing fractions and decimals as determined by a defined proportion threshold (considered as *Group B*).

A threshold was set for selecting decimals and fractions to control the balance of answers. This lends itself to our goal of being able to identify whether or not the format a student writes an answer, i.e. using factions vs. decimals, effects our ability to score student open response answers. Thus, with ABROCA, fairness can be identified at each threshold.

For the test set, a similar approach is taken. So the training and test set have both Group A answers which are distractor words and fractions, and Group B answers which have distractor words and a proportion of fractions/decimals (based on the same threshold for both training and test data).

To improve the reliability of the results, we re-sample/re-select the test dataset 10 times and evaluate the model's ability to score an open response answer. This form of cross validation allows us to see if the ability to predict the score was only for that unique set of words, or was the performance consistent across multiple iterations.

All of the studies will incorporate a Long Short Term Memory (LSTM) [12] model which utilizes the pre-trained word embeddings to automatically score open response answers and ABROCA is calculated. This is then used to run 3 studies. First, when incrementally increasing decimals being used in Group B, does the LSTM scoring model become more unfair? Second, whether or not fractions or decimals are the culprit of the potential unfairness in the automated scoring model by having answers just be fractions or decimals without distractor words? Third, when incrementally increasing unrecognized words (referred to as gibberish) being used in Group B, does the LSTM scoring model become more unfair? So does an imbalance in recognized words cause more unfairness between groups?

3.3 Results

In the end, Figure 1a showed that with the simulated study, when there is an increase in the proportion of decimals in Group B, there does not appear to be unfairness in the way Group A and Group B are evaluated. This is evident from the scale of the y-axis of Figure 1a, the ABROCA falls betweem 0.02 and around 0.04.

The second study, which there were only decimals and fractions in the test set (no distractor words), stayed constant at a ABROCA value of 0. The model was not unfair, it was able to equally evaluate both groups even when just isolated fractions and decimals.

The final simulation study, managed to show that increasing the imbalance between recognized and unrecognized tokens between groups increased the unfairness (ABROCA near 0.18). Figure 1b shows that the ABROCA score does indeed increase with more unrecognizable words within GloVe's pre-trained word embeddings. It should be noted that Table 1 shows some of the phrases used in the generated student answers were commonly associated with more correct answers

Table 1: Sample of Phrases and Their Associated Avg. Score

Generated Phrases	Avg. Score
my answer is	0.718750
i picked	0.622222
i guess the answer is	0.600000
i think it is	0.600000
i think the answer is	0.590909
i worked out	0.585366

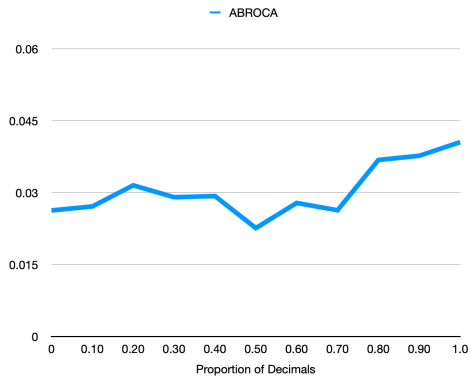
In the end, these simulated studies proved the largest risk for unfairness exists when there is differential coverage of answer-related tokens within applied methods utilizing pre-trained NLP embedding methods. So when answers consist of equally recognizable words within GloVe's pre-trained word embeddings, there's unlikely to be unfairness in the grading. There wasn't evidence that the inherent bias built into the pre-trained word embeddings elicited more unfair scoring of student answers in, in terms of this simulated study. But if there are unbalanced recognizable words and tokens in the student answers, attention needs to be paid to potential unfairness in the automated scoring.

4. STUDY 2: FAIRNESS IN REAL CONTEXTS

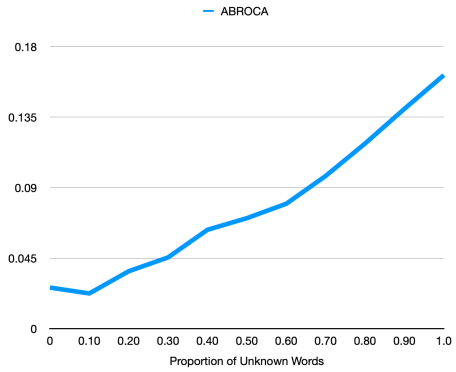
While a simulation study is powerful on its own, it is difficult to recreate authentic student data. For the final overall study of this research, we look to once again utilize ABROCA to identify if our own algorithm, trained on genuine student open response answers within ASSISTments, is unfair in its grading of women and men.

4.1 Data

The data consists of 141,612 graded authentic student open response answers across 2,042 unique problems. There were a total of 25,069 unique students who answered and 891 teachers graded those answers. Lastly, the scoring. This was performed on a 5 point scale, where students receiving a 4 is a perfect score.



(a) Study 1: ABROCA Values at Incremental Fraction/Decimal Thresholds



(b) Study 3: ABROCA Values at Incremental Fraction/Unknown Words Thresholds

It should be noted, to be able to perform the fairness analysis using ABROCA, gender was inferred. This performed by cross checking names with the census data. If the name was found only on the women or only on the men's list, it was labeled as such. In any names fell into multiple genders, it was labeled as unknown and excluded from this analysis.

4.2 Methodology and Results

Towards developing our predictions, we utilized another pre-trained algorithm, mentioned earlier, called SBERT. This is a pre-trained sentence embedding algorithm which allowed us to generate a single vector representation of each student answer. We then utilize a Canberra distance to identify which student answers are the most similar. Whichever was the most similar, that was the score we would assign. This approach managed to out do our previous models [6].

While utilizing, once again, ABROCA to identify potential unfairness, we apply this to our algorithm. We were able to show that our SBERT model with Canberra distance manages to fairly score both Male and Female student open response answers. Our model managed an ABROCA of 0.007, which is quite small, suggesting that our algorithm is indeed scoring fairly across these groups.

5. LIMITATIONS AND FUTURE WORK

While there were indications of unfairness in cases where there were unbalanced identifiable tokens within the student open response answers, this analysis is strictly middle school mathematics. This type of analysis would need to be applied to additional datasets to get a broader understanding of the potential unfairness in other subjects and age ranges. In terms of our analysis of our SBERT model for scoring student open response answers, while there wasn't unfairness identified, more work needs to be done to explore the embeddings themselves. Pre-trained word embeddings have been shown to have bias built in, but what bias is present in the pre-trained sentence embeddings? This is a question we look to explore further.

6. CONCLUSION

Overall, this study set out to run a simulated study to help identify potential unfairness within models utilizing pre-trained word embeddings. While there is bias present in the embeddings themselves, our simulated study didn't show this bias causing unfair scoring. However, our analysis did show that when developing models with pre-trained embeddings, unfairness can begin to occur when there is an imbalance of recognized tokens in the student answers. More specifically, our simulated study showed that when groups within the data use differing levels of recognized tokens, it increases the chance for unfair scoring.

While our simulated study showed how unfairness can present itself within a scoring model, our model on authentic student data did not show this unfairness. We were able to conduct an analysis of our model with ABROCA to compare our performance scoring identified male and female students.

In the end, we were able to utilize a simulated study to help identify potential unfairness in automated scoring models which utilize pre-trained word embeddings. Its been widely noted that those embeddings have bias built in, but our simulated study couldn't show an unfairness in the scoring of differing groups of simulated student answers. However, this study did suggest that there is a notable risk to fairness in cases where there are differences in the number of words that are unrecognized by pre-trained models across populations.

7. ACKNOWLEDGMENTS

We thank multiple NSF grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), as well as the US Department of Education for three different funding lines; the Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024), the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), and the EIR. We also thank the Office of Naval Research (N00014-18-1-2768) and finally Schmidt Futures we well as a second anonymous philanthropy.

8. REFERENCES

[1] Y. Attali and J. Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[2] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and

A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*, 2016.

[3] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[4] A. T. Corbett, K. R. Koedinger, and J. R. Anderson. Intelligent tutoring systems. In *Handbook of human-computer interaction*, pages 849–874. Elsevier, 1997.

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

[6] J. A. Erickson, A. F. Botelho, S. McAteer, A. Varatharaj, and N. T. Heffernan. The automated grading of student open responses in mathematics. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 615–624, 2020.

[7] P. W. Foltz, D. Laham, and T. K. Landauer. Automated essay scoring: Applications to educational technology. In *EdMedia+ innovate learning*, pages 939–944. Association for the Advancement of Computing in Education (AACE), 1999.

[8] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234, 2019.

[9] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter. Intelligent tutoring systems with conversational dialogue. *AI magazine*, 22(4):39–39, 2001.

[10] A. C. Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, D. Harter, T. R. G. Tutoring Research Group, and N. Person. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive learning environments*, 8(2):129–147, 2000.

[11] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[14] Z. Kastrati, A. S. Imran, and A. Kurti. Weakly supervised framework for aspect-based sentiment analysis on students' reviews of moocs. *IEEE Access*, 8:106799–106810, 2020.

[15] K. Y. Ku. Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking skills and creativity*, 4(1):70–76, 2009.

[16] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[17] J. Liu, Y. Xu, and Y. Zhu. Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*, 2019.

[18] M. E. Martinez. Cognition and the question of test item format. *Educational Psychologist*, 34(4):207–218, 1999.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[20] A. Onan and M. A. Toçoğlu. Weighted word embeddings and clustering-based identification of question topics in mooc discussion forum posts. *Computer Applications in Engineering Education*, 2020.

[21] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] J. Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[23] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[24] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168, 2017.

[25] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason. Online mathematics homework increases student achievement. *AERA Open*, 2(4):2332858416673968, 2016.

[26] M. G. Simkin and W. L. Kuechler. Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3(1):73–98, 2005.

[27] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.

[28] J. Z. Sukkarieh and J. Blackmore. c-rater: Automatic content scoring for short constructed responses. In *Twenty-Second International FLAIRS Conference*, 2009.

[29] J. Z. Sukkarieh, S. G. Pulman, and N. Raikes. Automarking: using computational linguistics to score short ,free- text responses. 2003.

[30] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891, 2016.

[31] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

APPENDIX

A. SIMULATED ANSWER STRUCTURES

Table 2: Training Set Corpus Generation

Length	Structure	Answer Content	Fill "1"	Fill "2" - Fractions	Fill "2" - Decimals
6	6 - A	i **1** the answer is **2**	'think', 'believe', 'feel', 'suppose', 'guess'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
6	6 - B	**1** the answer **2**	'i arrived at', ' i worked out', 'ended up with'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
5	5 - A	i **1** it is **2**	'think', 'believe', 'feel', 'suppose', 'guess'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
5	5 - B	i **1** the answer **2**	'chose', 'thought', 'picked	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
4	4 - A	**1 2**	'i arrived at', ' i worked out', 'ended up with'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
4	4 - B	my **1 2**	'guess was', 'answer was', 'belief was', 'answer is'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
3	3 - A	i **1 2**	'chose', 'thought', 'picked	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
3	3 - B	it **1 2**	'was', 'is'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6

Table 3: Test Set Corpus Generation

Answer Length	Answer Structure	Answer Content	Fill "1"	Fill "2" - Fractions	Fill "2" - Decimals
6	6 - A	i 1 2	'arrived at the answer', 'thought the answer was' 'calculated the answer was', 'guessed the answer was'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
6	6 - B	2 1	'is the answer i thought', 'was the correct choice here', 'is what i guessed right', 'was what I arrived at'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
5	5 - A	1 2	'im guessing it was', 'my work arrived at', 'the answer is clearly', 'clearly the answer is'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
5	5 - B	2 1	'was what i guessed', 'is what i calculated' 'was the clear answer', 'is the correct answer'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
4	4 - A	1 2	'my guess is', 'my answer is', 'my work showed', 'my thought is'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
4	4 - B	2 1	'is my choice', 'from my work', 'is my answer', 'is the answer'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
3	3 - A	2 1	'is right', 'is correct', 'i found', 'i thought'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6
3	3 - B	1 2	'answer is', 'choice was', 'i guessed', 'i thought'	3/4, 1/2, 1/3, 2/5, 3/5	0.75, 0.5, 0.33, 0.4, 0.6