On Understanding the Relation between Expert Annotations of Text Readability and Target Reader Comprehension

Sowmya Vajjala
National Research Council, Canada
sowmya.vajjala@nrc-cnrc.gc.ca

Ivana Lučić
Iowa State University, USA
ilucic@iastate.edu

Abstract

Automatic readability assessment aims to ensure that readers read texts that they can comprehend. However, computational models are typically trained on texts created from the perspective of the text writer, not the target reader. There is little experimental research on the relationship between expert annotations of readability, reader's language proficiency, and different levels of reading comprehension. To address this gap, we conducted a user study in which over a 100 participants read texts of different reading levels and answered questions created to test three forms of comprehension. Our results indicate that more than readability annotation or reader proficiency, it is the type of comprehension question asked that shows differences between reader responses - inferential questions were difficult for users of all levels of proficiency across reading levels. The data collected from this study is released with this paper[1], which will, for the first time, provide a collection of 45 reader bench marked texts to evaluate readability assessment systems developed for adult learners of English. It can also potentially be useful for the development of question generation approaches in intelligent tutoring systems research.

1 Introduction

Readability assessment refers to the task of predicting the reading difficulty of a text and its suitability to a target user's reading abilities. However, a typical computational approach relies on standard corpora that are created based on the writer's perception of what is difficult for a reader, and not on the target readers' comprehension data. While it is difficult to create such validated corpora in large samples sufficient to build automated models, lack of such data also raises a question about the validity of such models (Valencia et al., 2014; Williamson et al., 2014; Cunningham and Mesmer, 2014). A reasonably sized corpus of readers' comprehension scores for texts of varying reading levels can be a starting point in this direction, as it can enable evaluating the suitability of an existing readability assessment system for that target group as well as look for the validity of the labeled dataset.

This issue then raises a question of how we should evaluate comprehension. There is a significant body of research on forming questions to assess different levels of comprehension in educational and tutoring systems research (e.g., Day and Park, 2005; Adamson et al., 2013; Mazidi and Nielsen, 2015). Readability is not considered as a factor in such studies. In the few user studies that do consider readability (Rayner et al., 2006; Crossley et al., 2014; Vajjala et al., 2016), differences between different levels of comprehension were not considered.

In this paper, we take first steps towards understanding the relation between expert annotations, reader proficiency and comprehension for automatic readability assessment research by conducting a web-based reading study with over 100 participants in a natural reading environment. Participants read six newspaper texts, and answered six questions on each text, covering three levels of comprehension. We analyzed our results by using methods from educational assessment research. We are releasing the data from this study, which for the first time, creates a freely available reader response based dataset for evaluating readability assessment systems. While it is not a large dataset and we cannot claim to have solved the problem of validating the readability annotations against target user groups, we believe this study is a first step in a much needed direction.

Our paper's contributions can be summarized as

[1] https://github.com/nishkalavallabhi/BEA19UserstudyData

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 349–359
Florence, Italy, August 2, 2019.

follows: we conducted a user study with over 100 participants by,

- asking questions of different forms (short answer, T/F) that target three levels of comprehension (literal, re-organization, inference) for the first time,
- using a web-based reading setup where the readers read the full text in a normal computer based interaction setting, which can make the results potentially more relevant to practical, non-lab scenarios.
- using methods from educational assessment to show the differences in user responses for different levels of comprehension.

The rest of this paper is organized as follows: Section 2 summarizes related research. Sections 3 and 4 describe the study and results. Section 5 summarizes the insights gained from this study.

2 Related Work

Reading is the primary means of learning and knowing. Thus, readability or complexity of a text affects the comprehension process. Considering its important role in learning and assessment, text complexity has been extensively studied in the form of user studies, theories of comprehension, and computational approaches.

User studies on the impact of text complexity on reading comprehension have been done in Cognitive Psychology research since the 70s (Evans, 1972; Kintsch and van Dijk, 1978; Walmsley et al., 1981; Green and Olsen, 1988; Smith, 1988; Britton and Gülgöz, 1991). Eye-tracking was also used in the past to understand reading processes and comprehension difficulties (Just and Carpenter, 1980; Rayner, 1998; Jr et al., 2007). Attempting to study the problem from a second language reading perspective, Crossley et al. (2014) conducted a sliding-window based reading study where participants read texts word by word, using a collection of news articles written at three reading levels by language teachers. Comprehension was assessed by means of yes/no questions. More recently, Vajjala et al. (2016) combined both eye-tracking and second language reading perspectives by doing an eye-tracking study using texts from the same source (but not full text), asking readers to respond to two types of questions - factual and yes/no questions. They concluded that developing questions that address different forms of comprehension may lead to a better understanding of the text-reader interaction.

Though there has been some work on creating questions that aim at testing different levels of comprehension (Day and Park, 2005, e.g.,), it was not utilized in these studies. Further, eye-tracking and sliding window approaches are closer to a lab environment than real-world reading, which makes it difficult to conduct larger-scale studies which can yield more reader response data, which is needed for evaluating computational approaches.

Unrelated to such user studies, there is a large body of research on readability assessment in the past century. Some of the early research on assessing readability relied on asking readers comprehension questions to evaluate text difficulty (Dale and Tyler, 1934). Such approaches were also criticized in terms of what is the right way to assess comprehension and how the nature of questions asked may influence readers' performance (Lorge, 1939). However, modern day research on readability assessment over the past decade largely ignored this aspect in creating and evaluating readability models. Since we don't have access to the data from such older studies, there is a need for the creation of new reader response based corpora to evaluate modern computational models.

Computational models of automatic readability assessment (ARA) (Collins-Thompson, 2014) and automatic text simplification (ATS) (Siddharthan, 2014) were proposed in the past 15 years. Unlike early research in this direction, such approaches generally rely on the presence of corpora that are either manually annotated for grade level/readability score. These are typically written by teachers or other experts, without a direct input from target readers. Evaluation of ARA and ATS systems is also typically done either automatically by splitting the data into train-test set or, occasionally, by asking a small group of human raters to evaluate the texts in terms of their grammaticality, and simplicity - not by actually testing for comprehension with target population. Except for some systems specifically developed for addressing certain intellectual disabilities (Carroll et al., 1998; Canning et al., 2003), there is very little research in this direction. Considering this background, to our knowledge, this is the first study in the recent past which conducted a user study with a goal

of supporting the development and validation of computational models of readability assessment.

3 Methods and Experiment Procedure

Texts: We randomly selected 15 texts from the OneStopEnglish corpus (Vajjala and Lucic, 2018), consisting of manually simplified news articles from The Guardian, by English teachers, to suit beginner, intermediate, and advanced readers of English as Second Language (ESL). This corpus was also used in past user studies related to readability assessment (Crossley et al., 2014; Vajjala et al., 2016).[2]

Participants: 112 non-native English speaking participants were recruited for this study from among the student population of an American university by means of an internal email advertisement. Participants were compensated for their participation with Amazon.com gift coupons.

Questions: The onestopenglish.com news lessons included comprehension questions at the end of each article. However, these questions were primarily fill-in-the-blank and multiple choice questions, and they were not the same across all the reading levels for the same article. Further, they did not cover different forms of comprehension we wanted to check. Hence, the questions (and appropriate responses) for this study were created by an experienced language instructor following the guidelines of (Day and Park, 2005), and manually checked by the authors.

Questions covered three levels of comprehension: **literal**, **re-organization**, and **inferential**. Literal comprehension questions require learner's understanding of the straightforward meaning of the text. Therefore, the answers to such questions can be found directly and explicitly in the text. Reorganization questions require similar understanding, but learners are required to combine information from various portions of the text in order to provide a correct answer. Inference questions require a deeper understanding of the text, as the answer to such questions is not explicitly stated. The correct answer requires a combination of literal understanding of the text, learner's background knowledge and the ability to infer from what is written.

[2]An example of the degree of simplification and summary statistics about the texts we used can be found in Appendix in Table 8 and Table 9 respectively, and all the used texts are provided in the supplementary material.

Questions were created such that answers are the same for all three reading level versions of a given text (i.e., content deleted or added between versions will not affect answering these questions). Six questions were created per text, covering three levels of comprehension, and two question forms (True-False, short answer).[3]

Proficiency Test: All the participants completed a free English language proficiency test provided by the British Council[4] after they completed reading all the texts and answering all the questions. The test gave a percentage score, and hence was on a scale of 0–100.

Study Procedure: After IRB approval, the first step involved developing a web-based application for setting up the reading study. We developed a Python and MySQL based web application that allowed users to log in and read the displayed texts and their responses were stored. Each reader read 6 of the 15 texts randomly chosen balancing for reading level i.e., each user read two texts per reading level, and without reading the same text in multiple versions. After reading each text, they first saw two questions dealing with factual comprehension. The text was not visible while answering these questions. The next page had the text along with reorganization questions and the third page had the text along with inference questions. Reading time was calculated based on the time taken to click on the next page but was not used in our analysis. After finishing reading all texts and answering questions, the participants did the proficiency test. [5].

3.1 Data Analysis:

In order to test whether the reported comprehension scores (total and across levels) can be predicted from learner's reading proficiency and text readability, a variety of regression analyses were performed using SPSS (Corp, 2013).

To compare comprehension question types and the two question forms (T/F, short-answer) and find possible difficulty levels among them, Multi-

[3]The texts, questions and participant responses will be released with this paper and are provided as supplemental material for the submission.

[4]https://learnenglish.britishcouncil.org/en/content

[5]The code for this web-study will be released with the paper for reproducibility. It can potentially be re-used and enhanced to create a framework for testing larger, future studies in this direction.

Facet Rasch Measurement (MFRM) models were employed using the software package FACETS (Linacre, 2012). MFRM is typically used in psychometric and educational assessment research to examine different facets of question-answer data and their inter-relationships (Eckes, 2011). These relationships can include differences between participants, texts, question difficulty etc. In our case, the primary MFRM model used was a three facet model. The facets of measurement included the participants, three question types (based on comprehension) and two question forms (T/F and short answer questions). Assumptions required for all statistical analyses used were confirmed for both the analyses.

The data collected through this study is available on github at: https://github.com/nishkalavallabhi/BEA19UserstudyData.

4 Results

Post-study, we analyzed the responses from all the readers, and scored them manually using the question-answer key created while forming the questions.[6] The proficiency score was obtained automatically from the British Council test. Since we asked two questions per level of comprehension, each individual comprehension category had a score between 0–2 and total comprehension had a score between 0–6. Proficiency was on a score range 0–100. Table 1 shows descriptive statistics about the range of scores for our data set.

Score	**Mean**	**S.D**
Proficiency	76.6	10.25
Literal comp.	1.47	0.61
Reorganization comp.	1.45	0.67
Inferential comp.	1.33	0.65
Total comp.	4.26	1.19

Table 1: Summary of participant responses

Purely in terms of mean scores, readers generally seemed to do poorer on inferential comprehension than on the other two question types. The proficiency score was in the range of $[52 - 100]$ with a mean of 76.6. Table 2 shows the correlation between the scores for different comprehension question types and the overall comprehension score.

[6] provided in the supplementary material.

	Lit.	Reorg.	Inf.	Total
Lit.	1	-0.28	0.101	0.553
Reorg.		1	0.134	0.622
Inf.			1	0.672
Total				1

Table 2: Correlations between participant scores for different comprehension types

Clearly, while different comprehension scores had very low correlation among each other, they (as expected) had a higher correlation with the total comprehension score. This shows that the questions were indeed different in terms of what they are testing.

4.1 Regression Analyses

We estimated regression models to predict the different reading comprehension scores based on proficiency, reading level, and an interaction between proficiency and reading level. Table 3 shows the summary of a multiple regression model to predict the total comprehension score, in terms of the co-efficient, standard error, and the significance of the predictor variables. The results show that this model has a low R^2 of 5.3%. This indicates that proficiency and reading level can explain only 5.3% of the variance in participants' reading comprehension scores. Also, only proficiency was a significant predictor, albeit with a low unstandardized coefficient (B). This is clearly not useful information in a practical scenario to use as a basis to build predictive models to recommend appropriate texts for language learners based on their proficiency and text's complexity.

Table 3: Regression model with full data

	B	**S.E.**	**t**	**Sig.**
total comp.	1.902	.890	2.137	0.33
proficiency	.031	.012	2.715	0.007
reading level	.164	.412	.399	.690
prof. and reading level interaction	-.002	.005	-.444	.657
R^2= .053				

Vajjala et al. (2016) in their eye-tracking study with texts of two reading levels concluded that low proficiency readers fixate more for difficult texts compared to easy texts. To verify if that cognitive effort is also reflected in their comprehen-

sion performance, another multiple linear regression was calculated only with those participants who scored less than the median score (75) on the proficiency test. This additional analysis was conducted to investigate possible specific relationships of the same variables for low proficiency participants. The results are summarized in Table 4 and do not result in a different conclusion compared to the model with full data above in Table 3.

	B	**S.E.**	**t**	**Sig.**
total comp.	3.321	1.760	1.887	0.60
proficiency	.010	.026	.375	0.708
reading level	-.309	.815	-.379	.705
prof. and reading level interaction	.005	.012	.433	.665
R^2= .014				

Table 4: Regression model with low proficiency data

Regression models with these variables turned out to be poor fits for predicting scores for the three levels of comprehension separately as well, explaining less than 3% of the variance for all the three models (where literal, reorganization, and inference scores were the predictor variables respectively instead of total comprehension score). Proficiency had a statistically significant relation with only the literal comprehension score, and reading level was not significant in any of the models. Therefore, we are not discussing these analyses in further detail.

From what we see so far, it appears that we cannot predict reader comprehension based on an expert annotated measure of text readability, and/or a test of language proficiency. However, considering that the labels are given from the perspective of an instructor/writer and not the actual target reader, and considering that the texts did result in different scores from users, it is possible that there could be some other linguistic characteristics of text beyond the manually assigned readability label which relate to different forms of comprehension. One approach to explore this could be using the feature extraction modules from existing ARA systems. These methods extract a wide range of language features from texts, and there is evidence that, based on these features, texts can be successfully divided into different levels (Nelson et al., 2012). We would leave this exploration for future research.

4.2 MFRM Models:

To compare reader responses to different types and forms of questions, we constructed three three-facet MFRM models. The first model used participants, reading levels, and comprehension types (literal, reorganization, and inferential) as the three facets, and it was conducted to reveal possible difficulty levels of the three question types. The three facets for the second model were participants, reading levels, and question form (true/false and short answer), and this analysis was used to evaluate the comparability of the question form. Finally, the third model combined literal and reorganization comprehension into one group factual comprehension, and had the facets as participants, reading levels, and comprehension-question type combinations.

MFRM model calibrations from FACETS can be visualized by a vertical ruler known as **Wright map** or **variable map**. The first column in this map is the measurement scale with logits as measurement units. The second column shows the estimates of the participants' scores on the reading comprehension questions. This facet is positively oriented, so higher scoring participants appear on the upper portion of the ruler, while lower scoring participants appear at the bottom. The third column compares the reading levels in terms of difficulty. This facet is negatively oriented, and that means that more difficult levels would appear on the top of the ruler, while less difficult ones would appear at the bottom. The fourth column is discussed in more detail in paragraphs below for each model, and the fifth column maps the rating scale to the logit scale (first column).

In the associated summary tables for these three models (Tables 5–7), the reliability statistic is the ratio of true to observed variance for the elements of a given facet. It has a value between 0 to 1 (values closer to 1 are preferred) and this shows how reproducible is the ordering, and how reliably different are the values on the scale.

MFRM for levels of comprehension: Figure 1 shows the MFRM summary showing the distance between reading levels and comprehension levels.

As seen in the figure, the "levels" column does not show any differences between the reading levels. All the three levels are placed horizontally,

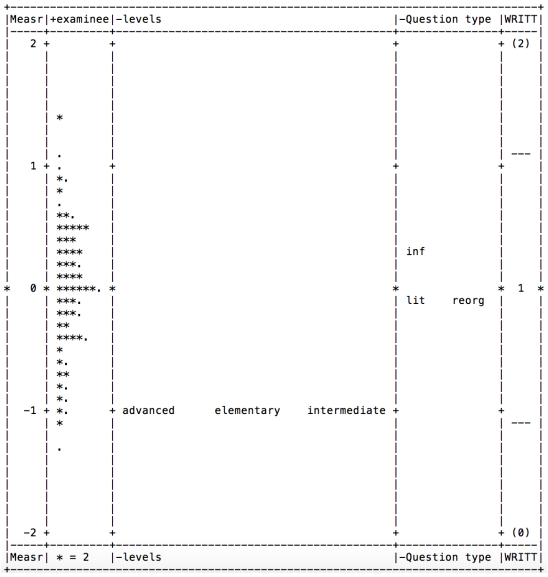


Figure 1: MFRM variable map comparing different types of comprehension

indicating there are no differences in terms of comprehension difficulty per reading level. The "Question Type" column in the ruler displays difficulty information about question type, and it is negatively oriented. This means that inferential comprehension questions were somewhat harder than literal and reorganization comprehension questions which seem to be the same level of difficulty.

The difference in logits (Table 5) is at 0.38, which is about 10% of the logit spread observed for the participants' reading comprehension. Small standard errors associated with the logit values indicate less variation from the mean. These values, along with other score averages for each comprehension type, are presented in Table 5.

comp. type	**observed raw score average**	**average proficiency measure (logits)**	**S.E.**
lit.	1.48	-0.13	0.07
reorg.	1.48	-0.12	0.07
inf.	1.33	0.25	0.06
Mean	1.43	0.00	0.06
S.D.	0.09	0.22	0.00
Separation = 3.26, Reliability = 0.91			

Table 5: MFRM for types of comprehension

The reliability statistic (given by $= separation^2/(1 + separation^2)$ for this model is 0.91, which indicates that there is a good separation between the two levels of questions - literal and reorganization grouped at one level, and inferential comprehension at the other level. Since literal and re-organization are two forms of factual comprehension, we can interpret this result as supporting evidence for inference questions being more difficult than factual questions.

MFRM for question types: The second MFRM model checked for the differences in comprehension scores between the two forms of questions - T/F and short answer. Figure 2 shows the summary of this model. The "question form" column in this summary indicates that short answer questions are more difficult to answer than true/false questions.

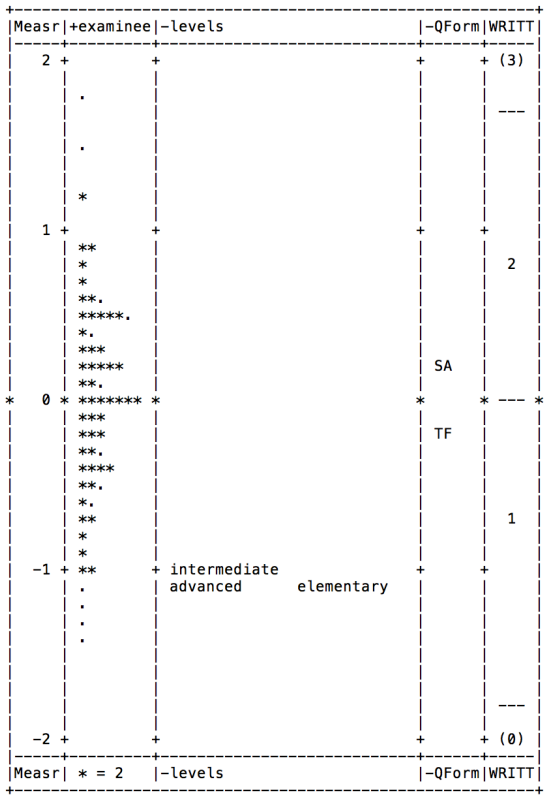
The difference in logits, presented in Table 6, is about 10% (0.42 logits) of the logit scale spread observed for the participants' reading comprehension. Again, small standard errors are observed. In terms of the reading level, as seen in column 3 of Figure 2, there are negligible differences between the three reading levels, with intermediate level being slightly more difficult than the other two. Table 6 shows the reliability statistic as 0.97, indicating that the participants showed a good degree of differences between the two forms of questions.

ans. type	**observed raw score average**	**average proficiency measure (logits)**	**S.E.**
short	2.03	0.21	0.05
T/F	2.25	-0.21	0.05
Mean	2.14	0.00	0.05
S.D.	0.16	0.30	0.00
Separation = 5.41, Reliability = 0.97			

Table 6: MFRM for forms of questions

MFRM for question types and forms: Since we saw two levels of difficulty among three levels of comprehension in the first MFRM model (Figure 1) and two levels of difficulty between question types, we evaluated a third MFRM model to understand the interaction between question form and the level of comprehension. Figure 3 shows the vertical ruler for this model, where literal and reorganization comprehension are grouped to-

Figure 2: MFRM variable map comparing the two forms of questions



gether into one (i.e., factual comprehension). The "Tasks" column in this summary shows that both types short answer and T/F inference questions were more difficult than Literal/Reorganization questions of both forms. This is further demonstrated by a pretty large difference in logit values which is presented in Table 10 7.

ques. type/form	observed raw score average	average proficiency measure (logits)	S.E.
T/F, fact.	1.56	-1.83	0.08
short, fact.	1.39	-1.16	0.08
T/F, inf.	0.69	1.42	0.07
short, inf.	0.65	1.57	0.07
Mean	1.43	0.00	0.06
S.D.	0.09	0.22	0.00
Separation = 23.08, Reliability = 1.00			

Table 7: MFRM for forms of questions and factual versus inferential comprehension

The difference between the lowest point of inference (T/F) and the highest point of literal and reorganization comprehension (short-answer), as displayed on the ruler (Figure 3), is 2.58 logits,

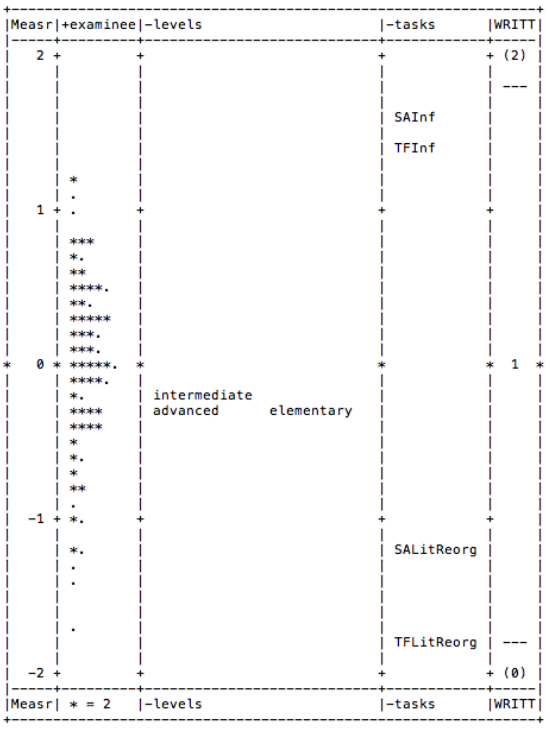


Figure 3: MFRM variable map comparing 2 forms of questions with respect to factual and inferential comprehension

which is 64.5% of the total logit spread. This means that the difference in difficulty is compelling. Additionally, within a given level of comprehension, short answer questions are more difficult to answer compared to T/F questions. For factual comprehension the difference in logits is 0.67, and for inferential comprehension, this difference is much lower at 0.15 logits. This accounts for 16.75% and 3.75% of the logit spread, respectively. This model shows that short answer questions are more difficult than T/F questions within a given level of comprehension.

5 Conclusions and Discussion

In this paper, we started out with the goal of understanding the relationship between expert annotations of readability, reader's language proficiency, and their reading comprehension, while also aiming to create a dataset which is useful for benchmarking computational models of readability assessment. To achieve this, we conducted a user study, and built a range of models on the data from this study.

The initial regression models were built to understand how much of reader comprehension scores can be explained by text's reading level,

and the reader's language proficiency. These resulted in a poor fit for the data with neither reading level, nor proficiency score contributing much to predicting reader comprehension. As it was seen in related studies, reader proficiency had a statistically significant correlation with the comprehension score. However, it did not convert into actual predictive power despite the fact that we had over 100 users and almost 700 data points for the regression model. This result, while questioning the validity of expert annotations to target population's comprehension, also leads us to speculate that single measures of readability level and user proficiency by themselves may not be sufficient to match texts to readers in terms of predicted comprehension. We may have consider a broader set of linguistic features, and go beyond a single proficiency measure.

The results of first MFRM model (Figure 1 and Table 5) lead us to a conclusion that the participants had difficulty answering inference questions compared to literal and re-organization questions, irrespective of the reading level. There are no differences between the scores for responses to literal and reorganization questions though, indicating that the separation is between factual (which includes both literal and reorganization) and inferential comprehension, rather than the three levels. The results from the second MFRM model (Figure 2 and Table 6) show that there are differences between question types with or without considering comprehension levels separately. Short answer questions were generally difficult to answer correctly compared to T/F questions. As the results from the third MFRM model (Figure 3 and Table 7) showed, even within a given level of comprehension, short answer questions remained more difficult than T/F questions.

5.1 Discussion

Overall, the results from our study are mixed. It did not provide any evidence in the direction of using expert annotation of text readability and reader's language proficiency information to be able to predict reader comprehension and recommend linguistically appropriate texts to language learners. On one hand, this may indicate that the level of simplification performed in the texts is not substantial enough to merit differences in comprehension, and such an experiment may hold more value in scenarios that aim at content simplification, along with form. On the other hand, it may also question the validity of expert annotations of text readability.

However, we know it is possible to automatically distinguish between these levels in this corpus using machine learning models (Ambati et al., 2016; Vajjala and Meurers, 2016; Vajjala and Lucic, 2018). Whether the variation between texts of any specific linguistic property (e.g., lexical richness, syntactic complexity, coherence) can be correlated with the differences in comprehension scores instead of "reading level" assigned by the teachers should be explored as a part of future work.

The MFRM results provide evidence in the direction of different questions resulting in different responses, and hence, call for the need to focus on methods to automatically generate questions that target multiple levels of comprehension. Asking the right kind of questions is important in various scenarios that relate to the application of Artificial Intelligence in education such as - learning support in tutoring systems, and the assessment of comprehension in both self-learning and test taking scenarios.

Limitations: The study has been conducted in a relatively less-controlled manner compared to, say, an eye-tracking study, so there is no way to know whether the participants actually read the texts. Additionally, the study did not consider how much the readers' background knowledge helped them in answering the questions. While these factors may have affected the outcome of this study (as they will for most studies of this nature), it would not also be possible to conduct a study with over 100 participants while controlling for both these aspects. One aspect that was not considered in this analysis was the variation within different texts used in the study (random variation). This can perhaps be addressed in future considering it as another facet that affects the outcome.

Finally, the results of this study could be specific to the texts or the proficiency test or the questions used. Consequently, we believe more such studies are needed in future to establish the relation between expert annotations and reader comprehension in the context of readability assessment. Conducting such studies with texts from different sources, and with texts that are validated more thoroughly (e.g., pedagogical texts, which are perhaps created with increasing levels of com-

prehension in mind) will be a useful direction to pursue to overcome this limitation.

Outlook: As mentioned earlier, an immediate extension to this work would be to study what linguistic properties that differ across reading levels (if any) correlate with reader comprehension of the text. Additionally, expanding the study to other forms of comprehension, collecting information about more than one form of proficiency as was done in (Crossley et al., 2014), and evaluating different readability assessment systems using this data could lead us in the right direction in terms of matching texts to target readers in future.

Acknowledgments

We thank the three anonymous reviewers for their useful comments. This study was funded through an Iowa State University's internal grant, when the first author was employed there.

References

David Adamson, Divyanshu Bhartiya, Biman Gujral, Radhika Kedia, Ashudeep Singh, and Carolyn P Rosé. 2013. Automatically generating discussion questions. In *International Conference on Artificial Intelligence in Education*, pages 81–90. Springer.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *HLT-NAACL*, pages 1051–1057.

Bruce K. Britton and Sami Gülgöz. 1991. Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83:329–345.

Yvonne Canning, John Tait, Jackie Archibald, and Ros Crawley. 2003. Cohesive generation of syntactically simplified newspaper text. In *Text, Speech and Dialogue: Third International Workshop, TSD 2000 Brno, Czech Republic, September 13-16, 2000 Proceedings*, page 145. Springer.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

IBM Corp. 2013. Ibm spss statistics for macintosh.

Scott A. Crossley, Hae Sung Yang, and Danielle S. McNamara. 2014. Whats so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

James W. Cunningham and Heidi Anne Mesmer. 2014. Quantitative measurement of text difficulty: Whats the use? *The Elementary School Journal*, 115(2):pp. 255–269.

Edgar Dale and Ralph W. Tyler. 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4:384–412.

Richard R. Day and Jeong-Suk Park. 2005. Developing reading comprehension questions. *Reading in a Foreign Language*, 17(1):60–73.

Thomas Eckes. 2011. Introduction to many-facet rasch measurement. *Frankfurt: Peter Lang*.

Ronald V. Evans. 1972. The effect of transformational simplification on the reading comprehension of selected high school students. In *Journal of Literacy Research*, pages 273–281.

Georgia M. Green and Margaret S. Olsen. 1988. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, chapter 5. Preferences for and Comprehension of Original and Readability Adapted Materials. Lawrence Erlbaum Associates.

Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. *Eye movement research: A window on mind and brain*, chapter Eye movements in reading words and sentences. Oxford:Elsevier Ltd.

M.A. Just and P.A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–355.

Walter Kintsch and Teun A van Dijk. 1978. Toward a model of text comprehension and productions. *Pschological Review*, 85(5):363–394.

John M Linacre. 2012. Facets computer program for many-facet rasch measurement, version 3.70. 0. *Beaverton, Oregon: Winsteps. com.*

Irving Lorge. 1939. Predicting reading difficulty of selections for children. *The Elementary English Review*, 16(6):229–233.

Karen Mazidi and Rodney Nielsen. 2015. Leveraging multiple views of text for automatic question generation. In *Proceedings of the 17th International Conference on Artificial Intelligence in Education (AIED 2015)*, pages 257–266. Springer International Publishing.

J. Nelson, C. Perfetti, D. Liben, and M. Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

K. Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422.

Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3):241–255.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Carlota S. Smith. 1988. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, chapter Chapter 10: Factors of Linguistic Complexity and Performance. Lawrence Erlbaum Associates.

Sowmya Vajjala and Ivana Lucic. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Sowmya Vajjala, Detmar Meurers, Alexander Eitel, and Katharina Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 38–48.

Sheila W. Valencia, Karen K. Wixson, and P. David Pearson. 2014. Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, 115(2):pp. 270–289.

Sean A. Walmsley, Kathleen M. Scott, and Richard Lehrer. 1981. Effects of document simplification on the reading comprehension of the elderly. *Journal of Literacy Research*, 13(3):237–248.

Gary L. Williamson, Jill Fitzgerald, and A. Jackson Stenner. 2014. Student reading growth illuminates the common core text-complexity standard: Raising both bars. *The Elementary School Journal*, 115(2):pp. 230–254.

A Supplemental Material

Table 8: Example texts for three reading levels

Reading Level	**Example**
Advanced	*Amsterdam still looks liberal to tourists, who were recently assured by the Labour Mayor that the city's marijuana-selling coffee shops would stay open despite a new national law tackling drug tourism. But the Dutch capital may lose its reputation for tolerance over plans to dispatch nuisance neighbours to scum villages made from shipping containers.*
Intermediate	*To tourists, Amsterdam still seems very liberal. Recently the city's Mayor assured them that the city's marijuana-selling coffee shops would stay open despite a new national law to prevent drug tourism. But the Dutch capitals plans to send nuisance neighbours to scum villages made from shipping containers may damage its reputation for tolerance.*
Elementary	*To tourists, Amsterdam still seems very liberal. Recently the city's Mayor told them that the coffee shops that sell marijuana would stay open, although there is a new national law to stop drug tourism. But the Dutch capital has a plan to send antisocial neighbours to scum villages made from shipping containers, and so maybe now people wont think it is a liberal city any more.*

Table 9: Summary Statistics for Texts

	Elementary			**Intermediate**			**Advanced**		
	WC	AWL	ASL	WC	AWL	ASL	WC	AWL	ASL
Text 1	474	4.57	20.22	482	4.80	22.55	484	4.88	23.74
Text 2	607	4.11	16.46	660	4.13	20.32	661	4.19	21.03
Text 3	589	4.19	16.88	641	4.21	19.47	657	4.27	20.03
Text 4	662	4.73	16.44	681	4.8	18.72	736	4.89	20.37
Text 5	527	4.5	17.53	561	4.6	19.31	599	4.72	20.68
Text 6	691	4.51	17.08	707	4.66	19	769	4.8	19.5
Text 7	627	4.65	20.7	646	4.82	22.07	714	4.9	23.9
Text 8	273	4.21	18.71	327	4.28	22.57	376	4.33	26.77
Text 9	596	4.58	17.26	658	4.71	19.63	703	4.74	21.71
Text 10	500	4.65	21.57	580	4.76	25.13	609	4.75	25.65
Text 11	445	4.40	18.27	536	4.54	20.28	554	4.61	22.86
Text 12	578	4.46	16.03	655	4.67	24.97	690	4.76	27.35
Text 13	472	4.55	19.13	552	4.73	22.54	586	4.77	23.17
Text 14	535	4.38	12.71	610	4.52	14.24	673	4.57	16.38
Text 15	437	3.97	17.52	547	4.01	19.83	599	4.09	20.9