Better student assessing by finding difficulty factors in a fully automated comprehension measure

Brooke Soden Hensler[1] and Joseph Beck[2]

Robotics Institute[1], Machine Learning Department[2]
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213
joseph.beck@gmail.com

Abstract. The multiple choice cloze (MCC) question format is commonly used to assess students' comprehension. It is an especially useful format for ITS because it is fully automatable and can be used on any text. Unfortunately, very little is known about the factors that influence MCC question difficulty and student performance on such questions. In order to better understand student performance on MCC questions, we developed a model of MCC questions. Our model shows that the difficulty of the answer and the student's response time are the most important predictors of student performance. In addition to showing the relative impact of the terms in our model, our model provides evidence of a developmental trend in syntactic awareness beginning around the 2nd grade. Our model also accounts for 10% more variance in students' external test scores compared to the standard scoring method for MCC questions.

1 Introduction

The goal of intelligent tutoring systems (ITS) is to assist students in learning. But, the effectiveness of the tutoring can be difficult to determine as it is often difficult to assess how much the student is actually learning. Our goal is to better understand and to better score multiple choice cloze questions, and in doing so improve the accuracy and efficiency with which we assess students. We accomplish this by creating a generic, widely usable model of multiple choice cloze question assessment[1].

We chose to model multiple choice cloze (MCC) question assessment because it is a format that is commonly used in assessing students' comprehension of text, is amenable to ITS, can be used in any domain utilizing text, and is economical in that it uses very little time by humans to initiate and can then be created and scored by computer. An MCC question is created by deleting a word from text and asking the

[1] **Acknowledgements:** This work was supported by the National Science Foundation, ITR/IERI Grant No. REC-0326153. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation

reader to select the deleted word from a set of response choices. MCC questions can be generated and presented in an ITS automatically. Such a format is used in the Project LISTEN Reading Tutor [1]. The Reading Tutor presents MCC questions to students with the goal of assessing reading comprehension.

Unfortunately, as with many assessments, especially newly constructed ones, it is very difficult to precisely infer students' knowledge based on question performance due to questions ranging widely in difficulty. For example, it would be inaccurate to equate raw score performance on ten easy questions to the raw score performance on ten difficult questions. The variance in question difficulty is especially acute within the Reading Tutor since questions are (semi) randomly generated in such a way that it is very rare for the same cloze question to appear multiple times. Therefore, we are working with a unique set of MCC questions to assess each student.

Given the disparity in question difficulty across students, our goal is to better extract information from the questions in order to assess the difficulty of the question and then use question difficulty information to better interpret the raw performance scores. By doing so, we hope to obtain more accurate student assessments. In addition to question difficulty, another factor influencing performance is students' motivation and engagement. Students' motivation and engagement vary tremendously, but past research has used MCC questions to model engagement [2].

ITS are prime candidates to develop and implement MCC assessments as they have a number of features which put them at a clear advantage over traditional pencil and paper-administered MCC assessments. First, ITS provide automatic question generation and scoring. Second, ITS enable us to consider precise response times and automatic part of speech identification for each question. These features allow us to consider factors which have not previously been part of MCC question assessment.

2 Data Description

Participants were 496 students in grades 1 through 6 (ages 5-12) in urban and suburban public schools in Pennsylvania and represented varying socio-economic statuses and ethnicities. Although there were participants from all six of these grades, MCC questions were only administered to students reading stories designated as 3rd grade reading level or above. Over the course of a school year, each student answered an average of 38 MCC questions for a total of 18,654 questions.

The Reading Tutor generated MCC questions by deleting a word (semi) randomly from the next sentence in the story the student was reading. The deleted word will be referred to throughout the paper as the "target word." The distractors were selected from the story being read and chosen to be words of similar frequency in English as the deleted word (see Fig. 1). The Reading Tutor read the sentence aloud (skipping over the deleted word) to the student and then read each response choice. The student's task was to select the word that had been deleted from the sentence. See [1] for additional details about how the cloze question intervention was instantiated in the Reading Tutor.

3 Modeling Approach

We developed our model of MCC question assessment by predicting whether a student would answer a *particular* MCC question correctly. Since our outcome measure is binary, we used multinomial logistic regression to calculate the relative impact of a number of terms. Our model includes *Part of Speech*, *Level of Difficulty*, *Response Time*, and *Student Identity* as factors. As covariates, we used *Tag-Primary POS Match*, *POS Confusability*, *Question Length*, *Deletion Location*, and *Syntactic Guess Rate* (see Table 1). We chose to model terms as covariates if we felt effects were generally linear and the clarity and interpretability benefited from a cleaner linear representation. Terms that were likely non-linear we treated as factors.

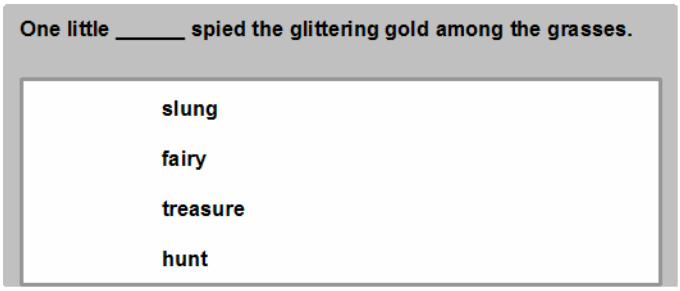

Fig. 1. Example of Multiple Choice Cloze Question presented on Reading Tutor.

Although better student assessment was the impetus for creating this model, *Part of Speech* was the driving force behind it because it is amenable to computer interpretation and past research [3] has used part of speech (POS) as an indication of word knowledge. We felt that we could glean some of the students' response strategies and possibly students' knowledge from *Part of Speech*. Syntactic awareness, the idea that individuals are aware of the parts of speech of words and sensitive to the ordering of words in a sentence, has been demonstrated to be a developmental trend [4, p.275-276]. Further, studies have indicated that some parts of speech are learned before others [5, 6]. We hypothesized that more proficient readers would use syntactic cues while less proficient readers would not. For example, we suspected that more proficient readers would cue on the fact that a verb should fill the blank given the following MCC question: "We can _____ the stars in the sky." Given the response choices: *at*, *with*, *most*, *see*, a more proficient reader would likely choose "see" since it is the only verb.

To add syntax to the model, we utilized the Moby Part of Speech Database (available at http://www.dcs.shef.ac.uk/research/ilash/Moby/). The Moby POS Database identifies all possible parts of speech for words in its database and arranges these POS in decreasing order based upon frequency of use (much like the ordering of entries for each word in a standard dictionary). This POS information was useful, but could not tell us the specific POS of the target word as it was used in the cloze sentence. For example, the Moby POS Database indicates that the target word in Fig. 1, "fairy," could be a Noun or an Adjective. In order to determine the POS of the

target word as it is used in the cloze sentence, we first replaced the blank with the target word. This process restored the original sentence so that the word was in context. Then, we used the TreeTagger [7] part of speech tagger to determine the target's part of speech in the sentence. The TreeTagger has demonstrated accuracy rates as high as 96.36% [7]. While we did not find the TreeTagger to be quite as accurate on our texts, spot checks of TreeTagger output on Reading Tutor data were performed by two humans and found to be acceptably accurate. Despite the TreeTagger's usefulness in its more accurate annotation of the target word as used in the cloze sentence, it is neither appropriate nor useful in determining the POS of the distractors. Therefore, we used the Moby POS Database to determine POS information for all of the response choices.

Table 1. Terms in Model

Factors	**Description of Term**	**Example Based on Fig. 1**
Part of Speech	Simplified part of speech classification of the target word as Noun, Verb, Adjective, Adverb, or Function Word.	Noun
Level of Difficulty	4 Levels of Difficulty based on frequency in English or special annotation.	Hard
Response Time	Response time rounded to the nearest second and capped at 9 seconds.	8 sec.
Student Identity	Unique Identification for each student.	Sally Student
Covariates		
Tag-Primary POS Match	Whether or not the tagged POS of the target word matched the most common POS the word could take on.	Yes
POS Confusability	The number of POS the target word can take on.	2
Question Length	Number of characters of the cloze question and the corresponding response choices.	95 characters
Deletion Location	Proportion of the sentence that is before the blank (location of word deletion).	0.19
Syntactic Guess Rate	Probability that the student could have answered the question using only part of speech information.	0.33

The first part of speech term in our model, *Part of Speech*, is the POS of the target word as annotated by the TreeTagger. To create the *Part of Speech* term, we converted the very specific part of speech tags returned by the TreeTagger (e.g., "WP$" is the tag returned for "Possessive wh-pronoun") to reflect a simplified classification system designating target words as Nouns, Verbs, Adjectives, Adverbs, or Function Words. We chose this granularity for POS classification since it is appropriate given our population, and it is consistent with other research (e.g.,[3]).

Our second part of speech term, *Tag-Primary POS Match*, is a covariate and could take on two values. *Tag-Primary POS Match* tests whether the target word's tagged POS (as computed by TreeTagger) is the same as its primary POS in the Moby POS Database. The motivation for the creation of this term was that the most common POS would probably be best known to the student, and that if a more obscure form of the word was used, the question would be more difficult.

Many words in English have multiple parts of speech (e.g., the word "pop" can, depending on contextual use, be all of our POS classifications), and we suspected that words that had multiple POS would be more difficult in that it is possible that the student may not have experienced the word used as a particular POS. Therefore, we added our third POS term, *POS Confusability*, as a covariate to account for the ambiguity of POS of the target word. *POS Confusability* is simply the number of POS the target word can take on.

The *Level of Difficulty* of the target word (and thereby all of the distractors) is classified in the Reading Tutor based on frequency in the English language [1]:
- "sight" words (the most frequent 225 words in a corpus of children's stories)
- "easy" words (the top 3,000 except for sight words)
- "hard" words (the next 22,000 words)
- "defined" words (words explicitly annotated with explanations)

Level of Difficulty was included as a factor because less proficient readers may not know the meanings of rare words. Word frequency has been used in other studies to select appropriate distractors for automatically generated MCC questions [8].

Student Identity was used as a factor to account for the overall individual performance on the questions to which she responded. Inclusion of *Student Identity* allowed us to more accurately estimate the relative impact of the other terms in our model by holding constant the impact of individual differences. This approach also accounts for the correlation among trials of a particular student, and properly calculates "N" for computing statistical significance [9].

Response Time was included as a factor because it has been demonstrated to account for engagement and performance in past studies [2]. The Reading Tutor recorded *Response Time* in milliseconds and it was later rounded to the nearest second in order to bin response times for a cleaner and more comprehensible analysis. Also, we truncated *Response Time* to nine seconds because we found that response times of greater than nine seconds were approximately the same for considering engagement.

We included *Question Length* as a covariate because we suspected that longer questions would be more difficult than shorter ones. *Question Length* was calculated as the number of characters of the sentence in addition to all of the response choices. The location of the deleted word in the sentence, *Deletion Location*, was included as a covariate because it was hypothesized that the cognitive load would be greater when the deleted word appeared early in the sentence, thereby making the question more difficult. *Deletion Location* was calculated as a percentage where the location of the blank (as measured by a count of the characters that preceded the beginning of the blank) was divided by the length of the question (as measured by the total number of characters of the sentence).

Syntactic Guess Rate, our final covariate, accounts for the probability that a student could have answered the question using only part of speech

information. The ability to use part of speech information may not necessarily be an explicit strategy the student uses, but rather a skill that develops and is evidenced by response selection. For example, if the deleted target word were a noun, the blank within the sentence could syntactically take on a noun; if the student were solely using syntactic knowledge, she might consider only response choices that were nouns (even if she cannot explicitly identify these words as nouns). Presumably questions with many distractors able to take on the same part of speech as the answer, that is, words that might "fit," would be harder. *Syntactic Guess Rate* was calculated by counting the number of response choices that could take on the part of speech of the deleted target word, and then taking its inverse (e.g., In Table 1, since three response choices can take on the POS "noun," the *Syntactic Guess Rate* is one divided by three, or 0.33).

4 Results & Discussion

After training the model using multinomial logistic regression, we were able to get a relatively good fit (Nagelkerke $R^2 = 0.22$). We found that *Part of Speech*, *Level of Difficulty*, *Response Time*, *Student Identity*, *Question Length*, *Deletion Location*, and *Syntactic Guess Rate* have a statistically reliable impact on students' MCC question performance (see Table 2). The impact of each of these terms is reflected by the β coefficient, which is the impact each term is having on student performance when all of the other terms in the model are held constant. A positive β value indicates that as the corresponding feature increases, the student is more likely to answer the MCC question correctly. Note that since β coefficients are not normalized, it is inappropriate to compare the coefficients from different factors and covariates with each other (although it is appropriate to compare the β coefficients for various levels of a factor).

For *Part of Speech*, there are five β coefficients, one for each POS classification (e.g., Noun, Verb, etc.). Each *Part of Speech* β coefficient reflects the relative impact on student MCC performance when the target word was the particular POS. For *Part of Speech*, a positive β coefficient indicates that the particular POS was easier for the students, while a negative *Part of Speech* β coefficient indicates that particular POS was more difficult for students. Nouns were the easiest POS for students ($\beta = 0.100$), while Function Words were most difficult ($\beta = -0.191$).

Question Length affected student performance such that the longer the question (and its response choices), the more difficult the question was ($\beta = -0.006$, $p < 0.001$). It is very likely that longer questions were more difficult because students had more information to process which resulted in a higher cognitive load.

An increased cognitive load is also the most likely explanation for the impact of *Deletion Location*. Recall that *Deletion Location* was the percentage of the question that appeared before the blank (the deleted target word) in the cloze sentence. Students were more likely to answer a question correctly (i.e., the question was easier) when the *Deletion Location* score was high (i.e. when the blank appeared late

in the sentence). The earlier in the sentence the blank appeared, the less likely the student would get the question correct ($\beta = 0.394$, $p < 0.001$).

Syntactic Guess Rate accounts for the chance that a question could have been answered correctly by relying on POS information. A high *Syntactic Guess Rate* score indicates fewer answer choices with a POS that would correctly fit in the blank, and thereby a higher chance that the student may have been relying on POS to answer correctly. The higher the *Syntactic Guess Rate* score, the more likely a student was to answer the question correctly ($\beta = 0.234$, $p = 0.003$).

Only two of the terms in our model, *Tag-Primary POS Match* and *POS Confusability*, did not have overall significance. However, when our population was divided based upon reading proficiency, *Tag-Primary POS Match* did have a varying effect depending on students' reading proficiency level. We will discuss two main findings in greater detail in the subsections below: a developmental trend of syntactic awareness, and using our model for more accurate student assessment.

Table 2. Impact of Terms in Model

Terms in Model	**Relative Impact of each Term**	**Significance of Overall Effect of Term**
Factors	**β**	**χ^2 p-value***
Part of Speech	-0.19 … 0.10	0.0001
Level of Difficulty	-0.96 … 0.17	1.01×10^{-46}
Response Time	-1.64 … 0.60	6.30×10^{-85}
Student Identity	-1.40 … 4.08**	4.50×10^{-171}
Covariates		
Tag-Primary POS Match	0.03	0.598
POS Confusability	-0.02	0.305
Question Length	-0.01	4.34×10^{-15}
Deletion Location	0.39	1.84×10^{-9}
Syntactic Guess Rate	0.23	0.003

*Significance of χ^2 is similar to the significance of β. χ^2 indicates the relative significance of the overall term, while, for factors, β p-values are only available for specific levels.

**For accuracy of relative impact of student performance, only students who answered 20 or more questions were included in this analysis (N=281).

4.1 Syntactic Awareness

In order to investigate students development of syntactic awareness, we used the students' Woodcock reading comprehension composite [10] test score to divide students into two groups. The Woodcock Reading Mastery Test is a standardized paper-based reading test administered by human testers that has several subtests which will be discussed in the next section. We had test scores for 373 students, and defined Low proficiency students as those who scored at the 2nd grade level or lower on the reading comprehension composite, and High proficiency as those who scored

higher than the 2nd grade level. This split divided our population approximately in half.

Investigation of *Syntactic Guess Rate* revealed striking differences between High and Low proficiency readers. Our model found that students in the High group were sensitive to how many of the possible responses could take on the same part of speech as the correct answer for the cloze sentence ($\beta = 0.393$, $p = 0.002$), while students in the Low group were insensitive to this term ($\beta = 0.080$, $p = 0.467$). This result suggests that students' syntactic awareness, at least within the context of MCC questions, begins around the second grade.

Further evidence of High proficiency readers' greater awareness of syntax over that of Low proficiency readers is shown in *Tag-Primary POS Match*. *Tag-Primary POS Match* shows that High proficiency readers are affected very little by whether the POS of the target word as used in the cloze sentence is also the target word's most common POS ($\beta = -0.030$, $p = 0.709$). This could be indicative of higher proficiency readers' familiarity with multiple senses of some words. On the other hand, Low proficiency readers are possibly affected by *Tag-Primary POS Match*, and may do better when the target word is used in its most common, and likely most familiar, sense ($\beta = 0.104$, $p = 0.122$).

4.2 Using difficulty model for student assessment

We now discuss using our MCC question assessment model to estimate student reading proficiency. The approach we used was to consider the β parameter associated with each student in the logistic regression model. This parameter represents how student identity influences the probability she will correctly answer an MCC question. Therefore, β can be considered an estimate of how well the student has done answering MCC questions (holding other aspects of each question constant) and a possible estimate of the student's reading proficiency. Another, simpler commonly used, approach (e.g. [11]) to estimating student proficiency is to consider the percentage of cloze questions she answered correctly. This approach is a commonly used method of scoring cloze questions.

We compared these two approaches of estimating student MCC performance and determined how well they related to external tests of reading. Since it is difficult to assess students who have only answered few MCC questions, we restricted the analyses in this Section to students who answered at least 20 MCC questions and for whom we had Woodcock test scores. This restriction reduced our sample to 281 students.

To compare these two methods of assessing students, we computed the (square of the) correlation coefficient between each of those measures and various reading tests (see Table 3). For this comparison, we used the students' scores on relevant subtests of the Woodcock [10]. The subtests we used were decoding (ability to read words), vocabulary, and passage comprehension. We also used the reading comprehension score, a composite composed of vocabulary and passage comprehension, and total reading score, a composite of all of the Woodcock subtests. For every test, the student-specific β parameter extracted from our model as an assessment outperformed

simply taking the percent of MCC questions that student answered correctly. Generally, β accounted for about 10% more variance (8% to 11%) in the test scores than did the percent correct. Therefore, β is a stronger assessment of student reading proficiency.

Table 3. Variance accounted for by logistic regression model and average percent correct

Test	r^2		Improvement in r^2
	Student β	Student % correct	
Decoding	0.34	0.23	0.11
Vocabulary	0.44	0.34	0.10
Passage comprehension	0.41	0.33	0.08
Reading comprehension composite	0.47	0.37	0.10
Total reading composite	0.43	0.32	0.10

5 Future Work

Our model uses word frequency to determine appropriate distractors for MCC questions where distractors are selected from the same word frequency classification as the target word. Our word frequency classification system breaks words into three possible categories, which can allow for words to be chosen as distractors from the same word frequency category that are actually relatively different in actual word frequency (e.g., since "Hard Words" includes 22,000 words, the 24,000th most frequent word in English could be matched with the 4,000th most frequent word). It may prove beneficial to use a more precise word frequency in selecting appropriate distractors and thereby provide a more accurate model of question difficulty.

The current model would benefit from testing on different populations. Our model relies on data of elementary school students from one geographic area. It would be interesting to test whether our model would be robust across populations. Further testing on a more diverse set of populations, especially those with a greater range of reading proficiency, could also reveal differences in the relative impacts of each term in the model, perhaps even necessitating additional terms. For example, a more extensive MCC assessment model could extrapolate Low versus High proficiency differences in older populations, such as college students. An extended MCC assessment model would likely reveal that college students are insensitive to some of the terms in our current model, such as POS, but are sensitive to other factors which we do not currently take into consideration.

6 Contributions & Conclusions

The initial goal of our model was to better understand student performance on MCC questions. In the past MCC questions have been interpreted in a crude way by

looking at mean score performance (e.g.,[11]), or by using a very complex linear regression model with 54 terms [1]. Our model provides a more accurate assessment of students (by providing a β score for each student) than the standard interpretation of MCC scores, which is simply the mean score. A short-coming of our assessment is interpretability. While mean scores are easily interpreted as percentages and mapped onto familiar letter grades, our measure is more complicated and does not currently have an easily translated score, but is more accurate.

Another contribution is better understanding of the process of answering MCC questions by using our model to estimate direct effects and developmental trends. Such tasks are impossible looking at mean MCC scores as they offer no additional information beyond the score. Past models, such as the 54-term linear regression model [1], included so many factors and information external to the cloze question that it was not possible to determine just what MCC responses load on. Our model has few enough features that the effect of each term can be interpreted.

In conclusion, we have shown a domain independent MCC question assessment model that is broadly applicable as it can be on any text. (e.g., a web page). Further, we have presented a model of MCC performance. Our model enables us to determine what makes some questions hard, to examine developmental trends of students, and to more accurately assess students.

References

1. Mostow, J., J. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri, *Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions.* Technology, Instruction, Cognition and Learning, 2004. **2**: p. 97-134.
2. Beck, J.E. *Engagement tracing: using response times to model student disengagement*. in *International Conference on Artificial Intelligence and Education*. 2005. p. *to appear*.
3. Schwanenflugel, P.J., S.A. Stahl, and E.L. McFalls, *Partial word knowledge and vocabulary growth during reading comprehension.* Journal of Literacy Research, 1997. **29**(4): p. 531-553.
4. Kamil, M.L., P.B. Mosenthal, P.D. Pearson, and R. Barr, eds. *Handbook of Reading Research, Volume III*. 2000, Lawrence Erlbaum Associates: Mahwah, New Jersey.
5. Gentner, D., *Some interesting differences between verbs and nouns.* Cognition and Brain Theory, 1981. **4**(2): p. 161-178.
6. Golinkoff, R.M., K. Hirsh-Pasek, L. Bloom, et al., *Becoming a word learner: A debate on lexical acquisition*. 2000, New York: Oxford University Press.
7. Schmid, H. *Probabilistic Part-of-Speech Tagging Using Decision Trees.* in *International Conference on New Methods in Language Processing*. 1994. p. 44-49.
8. Coniam, D., *A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests.* CALICO Journal, 1997. **14**(2-4): p. 15-33.
9. Menard, S., *Applied Logistic Regression Analysis.* Quantitative Applications in the Social Sciences, 1995. **106**.
10. Woodcock, R.W., *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. 1998, Circle Pines, Minnesota: American Guidance Service.
11. Abraham, R.G. and C.A. Chapelle, *The meaning of cloze test scores: an item difficulty perspective.* The Modern Language Journal, 1992. **76**: p. 468-479.