Natural Language Engineering **23** (2): 245–294. © Cambridge University Press 2016
doi:10.1017/S1351324916000024

*Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading**

JACK MOSTOW[1], YI-TING HUANG[2†], HYEJU JANG[3], ANDERS WEINSTEIN[4], JOE VALERI[5], and DONNA GATES[6]

[1]*Project LISTEN, School of Computer Science, Carnegie Mellon University, RI-NSH 4103, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA e-mail:* mostow@cs.cmu.edu
[2]*Information Management, National Taiwan University No. 1, Sec. 4, Roosevelt Road, 10617 Taipei, Taiwan e-mail:* d97008@im.ntu.edu.tw
[3]*Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA e-mail:* hyejuj@cs.cmu.edu
[4]*School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA e-mail:* andersw@cs.cmu.edu
[5]*e-mail:* joevaleri@gmail.com
[6]*e-mail:* donnamgates7123@gmail.com

(*Received 14 April 2015; revised 10 January 2016; accepted 11 January 2016; first published online 14 April 2016*)

Abstract

We describe the development, pilot-testing, refinement, and four evaluations of Diagnostic Question Generator (DQGen), which automatically generates multiple choice cloze (fill-in-the-blank) questions to test children's comprehension while reading a given text. Unlike previous methods, DQGen tests comprehension not only of an individual sentence but of the context preceding it. To test different aspects of comprehension, DQGen generates three types of distractors: ungrammatical distractors test syntax; nonsensical distractors test semantics; and locally plausible distractors test inter-sentential processing.

*This paper combines material from Mostow and Jang (2012), our AIED2015 paper (Huang and Mostow 2015) on a comparison to human performance, and substantial new content including improvements to DQGen and the evaluations reported in Sections 4.1 and 4.2. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080157, the National Science Foundation through Grant IIS1124240, and by the Taiwan National Science Council through the Graduate Students Study Abroad Program. We thank the other LISTENers who contributed to this work; everyone who categorized and wrote distractors; the reviewers of our BEA2012 and AIED2015 papers and this article for their helpful comments; and Prof. Y. S. Sun at National Taiwan University and Dr. M. C. Chen at Academia Sinica for enabling the first author to participate in this program. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute, the U.S. Department of Education, the National Science Foundation, or the National Science Council.
†This work was performed while all the authors were at Carnegie Mellon University.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

(1) A pilot study of DQGen 2012 evaluated its overall questions and individual distractors, guiding its refinement into DQGen 2014.
(2) Twenty-four elementary students generated 200 responses to multiple choice cloze questions that DQGen 2014 generated from forty-eight stories. In 130 of the responses, the child chose the correct answer. We define the *distractiveness* of a distractor as the frequency with which students choose it over the correct answer. The incorrect responses were consistent with expected distractiveness: twenty-seven were plausible, twenty-two were nonsensical, fourteen were ungrammatical, and seven were null.
(3) To compare DQGen 2014 against DQGen 2012, five human judges categorized candidate choices without knowing their intended type or whether they were the correct answer or a distractor generated by DQGen 2012 or DQGen 2014. The percentage of distractors categorized as their intended type was significantly higher for DQGen 2014.
(4) We evaluated DQGen 2014 against human performance based on 1,486 similarly blind categorizations by twenty-seven judges of sixteen correct answers, forty-eight distractors generated by DQGen 2014, and 504 distractors authored by twenty-one humans. Surprisingly, DQGen 2014 did significantly better than humans at generating ungrammatical distractors and marginally better than humans at generating nonsensical distractors, albeit slightly worse at generating plausible distractors. Moreover, vetting DQGen 2014's output and writing distractors only when necessary would halve the time to write them all, and produce higher quality distractors.

1 Introduction

Traditionally, generation of questions to assess reading comprehension has relied on humans – whether teachers (and students) who generate questions during instruction, or materials developers who generate questions beforehand. Although open-ended questions can elicit informative responses, they are challenging (and typically labor intensive) to score objectively. In comparison, multiple choice tests offer such compelling advantages as psychometric reliability and ease of scoring. 'The real value of multiple-choice items, however, is their applicability in measuring higher-level objectives, such as those based in comprehension, application, and analysis' (Haladyna, Downing and Rodriguez 2002). Unfortunately, 'Good multiple-choice test items are generally more difficult and time-consuming to write than other types of test items. Coming up with plausible distractors requires a certain amount of skill. This skill, however, may be increased through study, practice, and experience' (Burton *et al.* 1991). But such skill is time consuming for humans to acquire, and labor intensive to apply individually to any given text.

In contrast, automated generation based on the cognitive and linguistic principles underlying such skills can be systematic, inexpensive, scalable, and even superior. In recent years, automated question generation has been used for such varied tasks as inserting comprehension checks in a reading tutor (Mostow *et al.* 2004), generating comprehension instruction (Mostow and Chen 2009), testing vocabulary (Gates *et al.* 2011), recognizing children's spoken questions (Chen, Mostow and Aist 2013), evaluating language proficiency (Sumita, Sugaya and Yamamoto 2005; Lee and Seneff 2007; Lin, Sung and Chen 2007; Huang, Chen and Sun 2012), assisting academic writing (Ming *et al.* 2012), and identifying the main concepts in text about

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Some of those cells patrol your body. They are hungry, and they eat germs! Some stop the trouble germs make. Others make antibodies. They stick to germs. That helps your body find and kill _____.

a) are – **ungrammatical**
b) intestines – **nonsensical** (but grammatical)
c) terrorists – **plausible** (meaningful by itself but incorrect given the preceding text)
d) germs – **correct**

Fig. 1. Annotated example of a multiple-choice cloze question generated by DQGen 2012.

a specific domain such as biology (Agarwal and Mannem 2011a) or linguistics (Mitkov, Ha and Karamanis 2006).

One type of question especially conducive to automated generation is the multiple choice cloze (fill-in-the-blank) question, in which one word in a sentence is replaced with a blank. Answering without guessing requires having relevant background knowledge and understanding the context in order to select the best word from a list of choices for completing the sentence. Cloze questions are used in many standardized tests, such as Scholastic Aptitude Test, Test of English as a Foreign Language, and Test of English for International Communication.

Research has explored automated generation of cloze questions for various purposes, for example, to test comprehension of important concepts in textbooks (Mitkov *et al.* 2006). In the domain of language learning, a growing number of studies explain how to generate such questions to test English language proficiency with verbs (Sumita, Sugaya and Yamamoto 2005), prepositions (Lee, Sung and Chen 2007), adjectives (Lin *et al.* 2007), and grammar patterns (Huang *et al.* 2012). Especially in language learning, cloze questions can test the ability to decide which word is consistent with the surrounding context. Thus, they tap comprehension processes that judge various types of consistency, such as syntactic, semantic, and inter-sentential, in the course of constructing a situation model that represents 'the content or microworld that the text is about' (Graesser and Bertus 1998). In brief, these processes encode sentences, integrate them into an overall representation of meaning, notice gaps and inconsistencies, and repair them (van den Broek *et al.* 2002; Kintsch 2005).

DQGen (Diagnostic Question Generator) generates cloze questions for diagnostic assessment of a child's comprehension while reading a text. Throughout this paper, 'DQGen 2012' refers to the initial version described by Mostow and Jang (2012), and 'DQGen 2014' refers to the improved version reported here. Figure 1 shows an example.

As Figure 1 illustrates, DQGen's questions have four components:

- The **stem** is the clozed sentence: 'That helps your body find and kill _____.'
- The **context** consists of sentences that precede the stem: 'Some . . . germs.'
- The **correct answer** is by definition the deleted original word: 'germs.'
- The **distractors** are the other choices: 'are,' 'intestines,' 'terrorists.'

To generate questions for a given text, DQGen must decide which sentences to turn into cloze stems, which words to delete, and which distractors to use. To reduce disruption to the flow of reading, DQGen uses the natural break at the end of

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

each paragraph as an opportunity to insert a cloze question. To test the reader's comprehension, DQGen replaces the last word of the paragraph with a blank and selects three types of distractors. If the last sentence is too short (fewer than four words) or DQGen fails to find an acceptable distractor of each type, it simply leaves the last sentence unchanged rather than turn it into a bad cloze question. According to a review of comprehension assessments (Pearson and Hamm 2005), end-of-sentence multiple choice cloze questions are widely used, for example in the Stanford Diagnostic Reading Test and the Degrees of Reading Power.

DQGen uses different types of distractors to detect failures in different cognitive processes involved in reading comprehension. **Ungrammatical** distractors test the syntactic processing that recognizes grammatical structure. **Nonsensical** distractors test the semantic processing that interprets the meaning of a phrase or sentence. **Plausible** distractors test the inter-sentential processing that integrates information from previous context into comprehension of the current sentence. Aggregating children's performance over questions with these three types of distractors should not only measure their overall comprehension, but profile the difficulties encountered by a given child or posed by a given text. For instance, a child who processes syntax and semantics but not the relation of a sentence to the context that precedes it would reject ungrammatical and nonsensical distractors, but be as likely to pick the plausible distractor as to pick the correct answer.

The first and presumably easiest type of distractor renders the completed sentence ungrammatical. Choosing an ungrammatical distractor indicates failure to detect a syntactic inconsistency. Syntactic processing is part of comprehension but not necessarily well developed in children. Analysis of children's responses to 69,000 multiple cloze questions automatically generated, presented, and scored by the Reading Tutor (Mostow *et al.* 2004) found that children's performance decreased as the number of distractors with the same part of speech (POS) as the correct answer increased. However, this effect was weaker for lower level readers, indicating less sensitivity to syntax (Hensler and Beck 2006). For an ungrammatical distractor (e.g., '*are*' in Figure 1), DQGen picks a word from the context whose POS differs from the correct answer's (*germs*).

The second type of distractor makes the completed sentence grammatical but nonsensical. Choosing a nonsensical distractor indicates failure to detect a local semantic inconsistency with the rest of the sentence. For example, the nonsensical distractor in Figure 1 is '*intestines*.' A nonsensical distractor has the same POS as the correct answer. As a heuristic test of nonsensicality, DQGen checks that plugging the distractor into the sentence forms a context rare or non-existent in normal language. To operationalize this constraint, it checks that the 5-gram '*find and kill intestines*.' does not occur in the Google N-grams corpus.[1]

The third type of distractor, e.g., '*terrorists*' in Figure 1, makes the completed sentence meaningful in isolation but globally inconsistent with the preceding context. Plausible distractors are essential in testing inter-sentential processing, i.e., 'understanding that reaches across sentences in a passage,' because otherwise 'an

[1] The Google N-grams corpus represents punctuation symbols such as '.' as separate tokens.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 1. *Chronological overview of the experiments*

Section	Experiment	Purpose	When
2	Pilot study on DQGen 2012 questions	Intended versus perceived types	2012
4.1	Children's data on DQGen 2014 questions	Descriptive statistics on distractiveness	May–Jun. 2014
4.2	DQGen 2014 versus DQGen 2012 distractors	Improvement across two versions of DQGen	Jun.–Aug. 2014
4.3	DQGen 2014 versus human distractors	Quality and efficiency of machine versus human	Dec. 2014

individual's ability to fill in cloze blanks does not depend on passage context' – a frequent criticism of cloze questions (Pearson and Hamm 2005). A plausible distractor has the same POS as the correct answer, like a nonsensical distractor, but the sentence it forms when plugged into the blank makes sense – in isolation. DQGen uses the same heuristic test to qualify sentences as locally plausible that it uses to disqualify them as nonsensical – namely, whether it ends with a 4- or 5-gram that occurs in the Google N-grams corpus, e.g., '*find and kill terrorists.*' Thus, it decides that '*That helps your body to find and kill terrorists.*' is plausible (even though it's not normal English). However, '*terrorists*' doesn't make sense in the context of the preceding sentences, because it is semantically unrelated to the words in those sentences. As a heuristic test to ensure that the distractor is plausible only locally, not in the context of the preceding sentences, DQGen excludes distractors topically related to that context.

The rest of this paper is organized as follows. Section 2 reports a pilot study (Mostow and Jang 2012) to evaluate DQGen 2012. Section 3 describes DQGen 2014, including improvements motivated by the pilot study. Section 4 evaluates DQGen 2014 on children's data, and compares it to DQGen 2012 and human performance. Section 5 relates the work reported here to prior research. Section 6 concludes with contributions, limitations, future work, and potential applications. Table 1 summarizes these studies in chronological order, including the data analyzed, the purpose of the study, and when the data was collected.

2 Pilot study

How good are the questions generated by DQGen? To evaluate DQGen 2012, we asked human judges to score them. Section 2.1 explains how we evaluated questions, Section 2.2 reports inter-rater reliability, and Section 2.3 presents results.

2.1 Methodology

For the evaluation, we used DQGen 2012 to insert sample questions in an informational text for children (*Tiny Invaders* (2006)) that explains the concept of

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

germs and their danger. Of the eighteen paragraphs in this text, we excluded one paragraph because it was only two sentences long, and another because our grammar checker rejected its final sentence, '*Read on to find out how*.' For each of the other sixteen paragraphs, DQGen 2012 generated a cloze question with ungrammatical and nonsensical distractors, but it found plausible distractors for only thirteen of the questions, which we evaluated as follows.

We recruited eight human judges, members of our research team but unfamiliar with DQGen. We asked them to evaluate each question at two levels. At the high level, we evaluated the overall quality of each question by asking judges to categorize it as *Good*, *OK*, or *Bad*. We computed the percentage of generated questions categorized by human judges as acceptable, defined as *Good* or *OK*. We used a 3-point scale rather than a finer-grained scale both to get higher inter-rater reliability, and because we were interested more in how many of the questions were acceptable than in precise categorizations of quality. At the low level, we defined the quality of a distractor by how often it was perceived as its intended type. Thus, we asked the judges to categorize each of the multiple choices (correct answer plus three distractors) as *Ungrammatical*, *Nonsensical but grammatical*, *Meaningful but incorrect given the preceding text*, or *Correct*. To avoid biasing their responses, we did not tell them that each question was supposed to have one choice in each category. To elicit additional feedback, the form (similar to the form in Appendix A.1) invited judges to comment on the questions and distractors.

2.2 Inter-rater agreement

It is important to measure inter-rater reliability among human judges, especially on experimenter-designed measures such as the form we used. We used two inter-rater reliability metrics for measuring inter-rater agreement among more than two judges, Kendall's Coefficient of Concordance (Kendall and Smith 1939) and Fleiss' Kappa (Shrout and Fleiss 1979). Fleiss' Kappa is a statistical measure of inter-rater reliability for unranked data, while Kendall's Coefficient of Concordance is a non-parametric statistic making no assumptions regarding the nature of the probability distribution for ranked data. The overall quality categorizations involved ranked data from more than two judges, so to measure their inter-rater reliability we used Kendall's Coefficient of Concordance. Kendall's Coefficient of Concordance for overall quality was 0.40 on a scale from 0 to 1. This low value reflects the considerable variation between the judges, whose average categorizations of overall quality ranged from 1.3 to 2.6.

Categorization of each answer choice involved unranked data from more than two judges, so we used Fleiss' Kappa to measure its inter-rater reliability. Kappa was 0.58; a value of 0.4–0.6 is considered moderate, 0.6–0.8 substantial, and 0.8–1 outstanding (Landis and Koch 1977). Figure 2 shows the Kappa values for each label by the judges.

The low values of inter-rater reliability measures revealed the judges' lack of consensus, presumably due to differing interpretations of the instructions. For instance, one judge commented that instruction for rating the overall quality did not

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

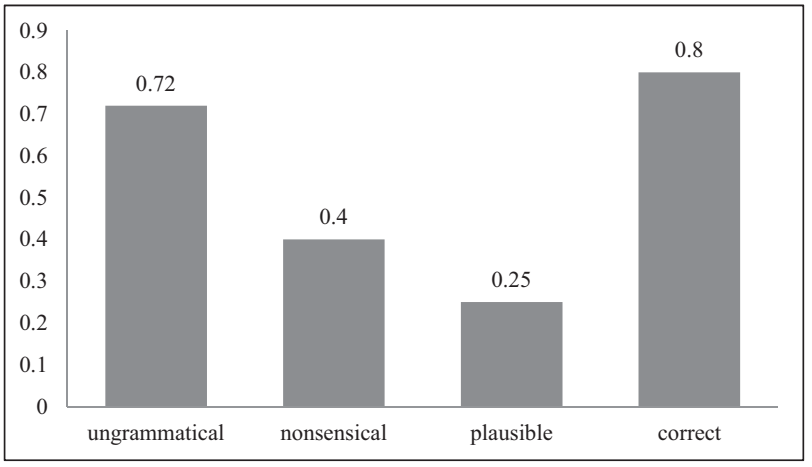


Fig. 2. Fleiss's Kappa for inter-rater reliability of each choice type in 2012 pilot study.

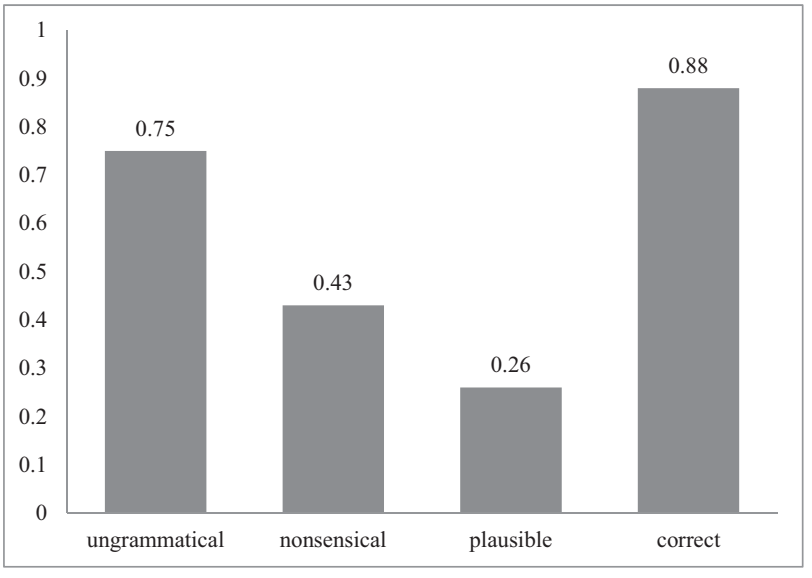


Fig. 3. Cohen's Kappa for agreement with the intended choice type in 2012 pilot study.

indicate whether a good question requires reading the preceding text. Another issue was missing and multiple categorical responses.

2.3 Results

We computed average categorizations of overall quality and agreement with the intended category of each answer choice.

To evaluate the overall quality of each question, we asked judges to rate the questions as *Bad*, *OK*, or *Good*, which we converted into numerical scores of 1, 2, and 3, respectively. Next, we averaged all these ratings to get the overall quality score 2.04, which corresponds to *OK* (Mostow and Jang 2012).

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Last year, a Pittsburgh elementary school did a pilot test of the Reading Tutor. Six children who started out almost three years below grade level made two years of progress in only eight _____.
a) tutors – **ungrammatical**
b) cauliflower – **nonsensical**
c) cities – **plausible**
d) months – **correct**

Fig. 4. An example of an unacceptable question generated by DQGen 2012.

Cohen's Kappa for agreement of judges with the intended category of each answer choice was 0.60. Note that in contrast to Section 2.2, where we used Fleiss' Kappa to measure inter-rater reliability, i.e., how well the judges agreed with each other on overall question quality, here we use Cohen's Kappa to measure distractor quality, i.e., how well the judges agreed with DQGen 2012 on the intended type of answer choices. Individual judges ranged from sixty-three per cent to seventy-nine per cent agreement with intended type (Cohen's Kappa 0.51 to 0.72). As Figure 3 shows, agreement was stronger for correct answers and ungrammatical distractors than for nonsensical and plausible distractors. On average, judges categorized ninety-four per cent of the correct answers as correct and agreed with DQGen 2012's intended type for ninety-one per cent of the ungrammatical distractors, sixty-three per cent of the nonsensical distractors, and only thirty-two per cent of the plausible distractors. Apparently judges found correct answers obviously right and ungrammatical answers obviously wrong, but nonsensical and plausible distractors harder to classify.

2.4 Error analysis

The pilot evaluation helped us refine DQGen 2012 into DQGen 2014 by fixing observed deficiencies. For example, some ostensibly 'ungrammatical' distractors were actually grammatical. Figure 4 shows an unacceptable question generated by DQGen 2012. First, the ungrammatical distractor '*tutors*' can be either a verb or a noun. If students interpret '*tutors*' as a noun, using it to complete the cloze sentence will be grammatical. Second, the nonsensical distractor '*cauliflower*,' like the answer '*months*,' is a noun, but must be plural and countable to fit grammatically into '*eight* ___' – and *'cauliflower*' is singular. Thus, the cloze sentence with '*cauliflower*' is ungrammatical. Finally, completing the cloze sentence with '*cities*' makes it grammatical but nonsensical (unless some of the six children moved between cities). Thus, students might well be able to select the intended answer over any of the distractors without actually reading the preceding context.

Answer choices, whether correct answers or distractors, are also problematic when they form idioms such as '*twisted in knots*' or '*make do*.' For instance, one pilot cloze question ended with '*twisted in* ___,' where the correct answer was '*knots*.' Another question ended with '*get your body to make* ___,' with '*do*' as a supposedly ungrammatical distractor.

Idioms pose multiple problems, although we found only two cases in our small pilot study. First, we want to test comprehension of the text, not just knowledge of specific idioms. Second, the word that completes an idiom can be far likelier than any other choice, making it too easy to guess based solely on local context, whether correct or not. Third, because idioms have non-componential semantics, the missing

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

word is liable to be semantically unrelated to other sentence words, causing DQGen 2012 to badly underestimate its local relevance.

Detecting idioms automatically is a research problem in its own right (Li and Sporleder 2009; Li, Roth and Sporleder 2010). We might be able to recognize an idiom by using the fact that its N-gram frequency is much higher than expected based on the frequency of its individual words. A simpler approach is to consult a dictionary of common phrases. Either approach would require extension to handle parameterized idioms such as '*a chip on [someone's] shoulder*,' or non-contiguous forms such as '*Actions do in fact speak louder than words*.'

3 How DQGen 2014 works

How does DQGen 2014 work? To produce each type of distractor, DQGen 2014 uses a generate-and-test approach. It chooses a candidate at random from a source of candidates for that type of distractor, and rejects the candidate if it does not satisfy the constraints for that type. In a few cases, DQGen 2014 treats a constraint as a preference: that is, if none of the candidates that satisfy it survive subsequent tests, DQGen 2014 drops the constraint and considers candidates that violate it.

Each type of distractor (ungrammatical, nonsensical, plausible) has a different source. Ungrammatical distractors come from the preceding sentences. Nonsensical distractors come from words categorized by Biemiller (2009) at or below grade 4. Plausible distractors come from Google N-grams that match the end of the sentence other than the answer.

We treat distractor generation as a search process. Each type of distractor requires certain properties. To reduce the amount of search required, we use a source of candidates likely to satisfy some of those properties. However, a side effect of choosing candidates from a particular source is that they may tend to have other properties as well. For example, DQGen uses the rest of the paragraph as a source of ungrammatical distractors, which is likely to contain enough different parts of speech to find some that are ungrammatical. But words drawn from the same paragraph are also likely to be at a similar reading level, semantically related to the target word, and topically related to the stem. In contrast, DQGen uses a graded word list as a source of nonsensical distractors, to make it likely that the student understands them. But the fact that this list is independent of the story makes it likely that they are semantically unrelated to the target word, and topically unrelated to the stem. Finally, DQGen uses Google N-grams as a source of plausible distractors to ensure that they at least occur in the immediate context, and hence are likely to make sense in the completed sentence. The fact that the Google N-grams were sampled from web text makes the candidate distractors likely to include words more frequent on the web than in other types of text.

Sections 3.1, 3.2, and 3.3 respectively describe DQGen 2014's constraints on individual words, on the completed sentence, and on relevance. Each section describes how DQGen 2012 worked, and then the improvements motivated by analysis of its errors and incorporated in DQGen 2014. Table 2 summarizes the source and constraints for each distractor type, and the order they are applied.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 2. *Constraints on each distractor type: (Parenthesized constraints are assumed true without testing). Preferences are constraints dropped if no candidates survive otherwise. *Starred constraints ought to be tested but are not.*

Scope:	Purpose:	Constraint:	Ungrammatical distractor chosen from other words in paragraph	nonsensical distractor chosen from a list of words up to grade 4	Plausible distractor chosen from matching Google N-grams
Lexical	Distinct	2012: Same root as answer	No	No	No
		Same root as another choice	No	No	No
	Familiar	2012: Unigram frequency at least 5 million	Yes	Yes	Yes
		2012: Single word	(No)	No	(No)
		In WordNet	Yes	Yes	Yes
	Meaningful	Stop word	prefer No	No	No
		Modal verb	No	No	*
		Homograph	No	No	*
		Proper name	No		*
Sentence	Grammatical	Same as some POS of answer	No	Yes	(Yes)
		2012: Link grammar parser succeeds	No	Yes	prefer Yes
		Same syntactic structure as original			Yes
	Normal	Last four (preferably) five words of sentence is in a Google N-gram table	No	No	Yes
Context	Relevant to stem	Related to words earlier in sentence			Yes (highest score)
	Irrelevant to preceding context	Related to words in preceding two sentences			No (less than the average candidate or the correct answer)

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

3.1 Lexical constraints

We now specify the lexical constraints used in DQGen 2012 or added in DQGen 2014. These constraints apply to answers as well as to all three distractor types: choices should be distinct, familiar, and meaningful.

3.1.1 Distinct

The choices in a well-formed multiple choice question should be mutually exclusive, and only one choice should be correct (Haladyna *et al.* 2002). DQGen 2012 constrained all the distractors to differ both from the answer and from each other. However, one of the questions in the pilot study included both '*throat*' and '*throats*,' which differ but share the same root, demonstrating that this constraint was too weak. Therefore, we strengthened it in DQGen 2014 to ensure that the choices are not only distinct but dissimilar, that is, have different roots. Allowing more than one choice with the same root would make sense only for questions that focus on minor differences between them, for example to test knowledge of inflectional morphology.

3.1.2 Familiar

Distractors must be familiar to children; otherwise they may test vocabulary rather than reading comprehension (Cassels and Johnstone 1984). Educators use various metrics to quantify the difficulty of text and individual words. A common way to express these metrics is grade level, i.e., the grade at which students are expected to understand the text or know the word meaning. DQGen 2012 satisfied this constraint for ungrammatical and nonsensical distractors by generating them from the paragraph and a grade-leveled word list (Biemiller, 2009), respectively. However, neither of these sources is large enough to provide enough candidate plausible distractors, because they cover too few of the Google N-grams used to test local coherence. Therefore, to exclude words likely to be unfamiliar to children, DQGen 2012 filtered out candidates whose unigram frequency falls below 5,000,000. We tuned this threshold by informal trial and error; higher thresholds proved too stringent to yield any plausible distractors.

In addition, DQGen 2012 constrained all four choices to be single words rather than phrases. Therefore, in using Biemiller's table to generate candidate nonsensical distractors, it filtered out the multi-word entries, such as '*barbeque sauce*.'

The pilot study and subsequent testing exposed shortcomings of previous strategies. In particular, the fact that a word has a frequency of at least 5,000,000 does not guarantee its familiarity to children. To bolster enforcement of this constraint, DQGen 2014 additionally requires all four choices to occur in the WordNet (Fellbaum 2012) database of 155,287 nouns, verbs, adjectives, and adverbs.

Moreover, one judge in the pilot study considered some words too difficult for children, such as '*gauge*.' In fact, Biemiller (2009) rates the noun sense of this word at grade 2, but its verb sense (to estimate) at grade 10. These examples illustrate a limitation of DQGen 2012's methods to pick familiar words as distractors. It picked ungrammatical distractors from the words in the paragraph, nonsensical distractors

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

from Biemiller's word list, and plausible distractors from Google N-grams, filtered by unigram frequency to avoid rare words. In all three cases, DQGen 2014 constrained words rather than word senses.

A more sophisticated approach would determine a distractor's word sense, or at least POS, when used to complete the sentence, and categorize the familiarity of its specific sense or POS. Tagging the distractor POS is easier than determining its word sense(s) when inserted in the sentence. Rating the familiarity of different word senses would require either a grade-leveled list of them like Biemiller's (2009), or a resource with information about the frequency of different word senses, such as the ordering of synsets in WordNet. Either way we'd need reliable POS tagging and word sense disambiguation to identify the distractor's POS and sense when inserted in the stem.

3.1.3 Meaningful

DQGen 2014 added lexical constraints to exclude stop words, homographs, and proper nouns to ensure that the answer and distractors are meaningful, that is, carry unambiguous semantic content and test comprehension.

Some words carry too little meaning to use as choices in cloze questions. In particular, stop words, modal verbs such as *can*, *cannot*, and *will*, and common verbs such as any form of *be*, *do*, *have*, and *get*, lack sufficiently specific semantic content to decide their relevance to the context. A related reason to avoid them as distractors, as a judge in the pilot study commented, is lest children notice that they are seldom the correct answer, and therefore eliminate them without considering them. DQGen 2014 implements these constraints by excluding the modal verbs *are*, *be*, *can*, *cannot*, *could*, *did*, *do*, *does*, *get*, *got*, *gotten*, *had*, *has*, *have*, *is*, *may*, *might*, *must*, *shall*, *should*, *was*, *were*, *will*, and *would,* as well as a list of common verbs and stop words (too many to list them all here).

Similarly, DQGen 2014 rejects homographs – words with the same spelling but different meanings and sometimes pronunciations – because their meaning is ambiguous. For instance, *address* can be a noun (location of a building) or verb (speak to someone). Thus, whether it is grammatical or makes sense depends on how it is read, rendering it unsuitable as a choice in a cloze question. To implement this constraint, DQGen 2014 excludes the homographs *abuse*, *address*, *appropriate*, *august*, *bow*, *close*, *combat*, *combine*, *conduct*, *content*, *convert*, *does*, *dove*, *herb*, *lead*, *live*, *lives*, *minute*, *moderate*, *pate*, *pervert*, *present*, *presents*, *produce*, *project*, *read*, *record*, *sewer*, *sow*, *subject*, *tear(s)*, *use*, *uses*, *wind*, and *winds.*

Conversely, a proper noun such as the name of a person, location, date, or time, is so specific that including it as a choice in a cloze question is liable to test particular world knowledge rather than reading comprehension ability. To avoid this risk, DQGen 2014 excludes any word that starts with a capital letter.

3.2 Constraints on completed cloze sentences

Sections 3.2.1 and 3.2.2 respectively specify how DQGen 2014 tests grammaticality and nonsensicality.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

3.2.1 Grammatical

As Table 2 shows, all three types of distractors involve grammaticality constraints. Ungrammatical distractors must make the completed sentence ungrammatical, e.g., '*That helps your body find and kill are.*' In contrast, nonsensical and plausible distractors must make the completed sentence grammatical, e.g., '*That helps your body find and kill terrorists.*' To check the grammaticality of a completed sentence, DQGen 2014 uses three natural language processing tools: the Stanford POS Tagger (Toutanova, Klein, Manning and Singer 2003), the Stanford Parser (Klein and Manning 2003), and the Link Grammar Parser (Sleator and Temperley 1993).

DQGen 2012 used the Stanford POS Tagger (Toutanova *et al.* 2003) to categorize an answer or distractor candidate as a noun, verb, adjective, or adverb. It required a plausible or nonsensical distractor to have the same POS as the answer, and an ungrammatical distractor to have a different POS. However, the pilot test revealed two problems with this scheme.

First, some words in English can have more than one POS. For example, the ungrammatical distractor '*tutors*' in Figure 4 can be either a noun or a verb. DQGen 2014 therefore rejects an ungrammatical distractor if it has any POS that the answer can have, not just its tagged POS in the sentence. DQGen 2014 finds every possible syntactic category (e.g., noun, verb, etc.) of a candidate in WordNet; if the candidate is not in WordNet, it finds these categories in the British National Corpus.

Second, this categorization is too coarse to ensure grammaticality when desired, e.g., for nonsensical distractors. For instance, it fails to distinguish singular from plural nouns, and therefore misclassifies '*in eight cauliflower'* as grammatical. To fix this problem, DQGen 2014 uses the Stanford POS Tagger to identify the correct answer's finer-grained POS, so it tags '*cauliflower*' as a singular noun (NN), as opposed to a plural noun (NNS) like the correct answer '*months.'*

To be grammatical, a distractor must be appropriately inflected. However, all words listed in Biemiller's (2009) table are base forms. Therefore, a distractor generated from the table must be inflected into the appropriate form to match the POS of the correct answer, for instance, by converting a singular noun into plural form, or the base form of a verb into the appropriate number and tense. DQGen 2014 uses systematic transformation rules to inflect regular words, and table lookup to inflect irregular nouns and verbs.

Moreover, more than one POS can fit grammatically in the blank. For example, consider the sentence '*Solar panels will collect the energy necessary for the electricity in the space* ___.' The answer '*station*' is a noun, but the verb '*left*' fits grammatically too. To test if a completed sentence is grammatical, DQGen 2014 therefore also uses the Link Grammar Parser (Sleator and Temperley 1993), a syntactic dependency parser, as a heuristic grammaticality checker. It usually accepts grammatical sentences and rejects ungrammatical ones, especially for the short sentences typical of children's text. However, it sometimes fails to accept a grammatical sentence, as the last row of Table 3 illustrates.

As the '*space station*'/'*space left*' example illustrates, replacing an answer with a grammatical distractor may change the syntactic structure of the sentence, as Figure 5

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

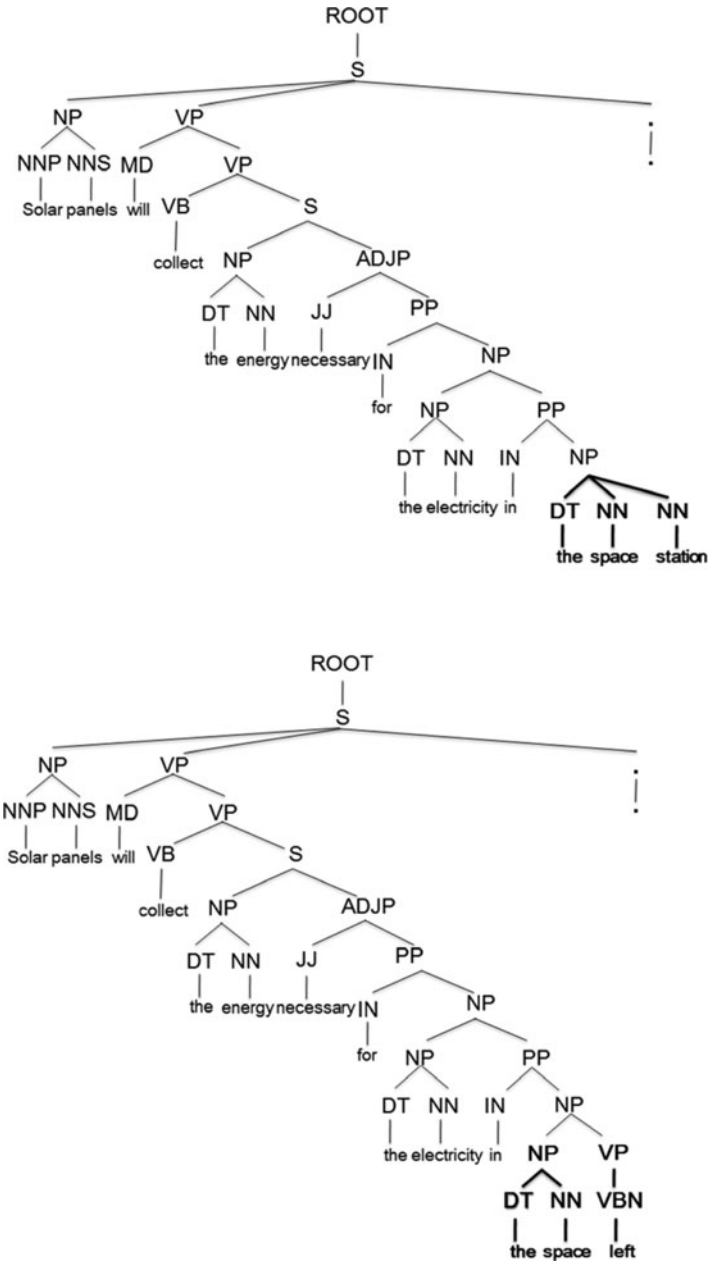


Fig. 5. An example syntactic structure changed by replacing the answer with a distractor.

shows. DQGen 2014 prohibits such changes lest they alter what the question actually tests. To enforce this prohibition, DQGen 2014 parses the completed sentence and compares its parse tree to the parse tree of the original sentence. If their syntactic structures differ, it rejects the candidate.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 3. *Examples of grammaticality checking by Link Grammar Parser*

Sentence	Grammaticality	Link Grammar Parser
The germs hide in food or people.	grammatical	accepts
The germs hide in food or world.	ungrammatical	rejects
So keep dirty hands away from cuts and your face.	grammatical	rejects

3.2.2 Normal

As a heuristic test of whether a completed sentence is plausible, DQGen 2012 checked whether its final 5-gram (including the end-of-sentence punctuation) occurs in the Google N-grams corpus (Brants and Franz 2006), implying that it is sufficiently normal English to appear at least forty times on the Web. For plausible distractors, the 5-gram consisting of the last four words of the sentence, followed by a period, must appear in this corpus. (For ungrammatical and nonsensical distractors, the last 4-gram of the completed sentence must not occur.) To enforce this constraint, DQGen 2012's source of candidate plausible distractors consisted of Google 5-grams of the form $W\ X\ Y\ \_$. Here, W, X, and Y are the three words preceding the correct answer in the original sentence, e.g., '*find and kill*.' If DQGen 2012 found fewer than five such 5-grams, it also allowed 4-grams of the form $X\ Y\ \_$., e.g., '*and kill* _.'

In the pilot study, however, judges categorized only thirty-two per cent of the intended plausible distractors as plausible, and forty-two per cent as nonsensical. To reduce the incidence of such cases, DQGen 2014 checks if the final 5-gram preceding the period appears in the Google N-gram corpus, in which case the completed sentence is likelier to make sense. For instance, DQGen 2012 erroneously allowed '*family*' as a plausible distractor for '*We nip if they stray too far from* __.' because the Google N-gram corpus includes the 5-gram '*too far from family*.' In contrast, DQGen 2014 rejects '*family*' as a plausible distractor because the corpus does not include the 5-gram '*stray too far from family*.'

The 5-gram constraint is often too strong, disqualifying all candidate plausible distractors. When this case occurs, DQGen 2014 relaxes the constraint to use just the final four words of the sentence, in an attempt to generate a cloze question anyway.

3.3 Constraints on global and local relevance

A plausible distractor should be relevant to the rest of the sentence, but not to the preceding context. Sections 3.3.1 and 3.3.2 respectively specify how DQGen 2014 tests local and global relevance.

3.3.1 Relevant to stem

To score local relevance, DQGen 2012 measured the relatedness of a distractor to each word in the stem only by how often they co-occur in the same thirty-word

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

window in British National Corpus. To aggregate this measure over the words in the stem, it used a Naïve Bayes formula:

$$\Pr(c|stem) \propto \Pr(c) \prod_{i=1}^{n} \Pr(w_i|c).$$

Here, c is a candidate plausible distractor and w_i is the ith content word in the stem. DQGen 2012 averaged these local coherence scores over the candidates, and allowed only candidates whose local coherence scores were above the mean, as a heuristic to increase the likelihood that they made sense locally.

This approach quantifies relevance as co-occurrence, thereby overlooking related words that seldom co-occur. To compute relevance more robustly, DQGen 2014 therefore instead averages the distributional similarity of a given word to the content words in the stem. DQGen 2014 uses DISCO (Kolb 2008, 2009) to calculate distributional similarity as the extent to which two words tend to co-occur in British National Corpus with the same other words, where co-occur means within three words of each other.

3.3.2 Irrelevant to preceding context

A good plausible distractor should not be too plausible. DQGen 2014 uses the same method to calculate relatedness to the context as to the stem, but constrains plausible distractors to be less relevant to the context than the answer is. DQGen 2014 ranks the candidates that satisfy this constraint by their global relevance (relatedness to the context), and chooses from the bottom half the candidate with the highest local relevance (relatedness to the stem). This heuristic gives local relevance priority over global (ir-)relevance; in comparison, DQGen 2012 chose the candidate with the lowest global relevance, which often turned out to be nonsensical because it was not locally coherent.

The purpose of DQGen 2014's heuristic is to pick a distractor that makes sense locally but not in the context of the preceding sentences, in order to detect failures of inter-sentential comprehension processes that monitor global consistency. To maximize the contrast in global relevance of the plausible distractor versus of the answer, how large a context should we consider? On the one hand, the longer the context, the more words related to the answer it may contain, driving up its relevance score. On the other hand, the further back in the text, the fewer such words are likely to occur. To quantify this tradeoff empirically, we computed the relevance of the answer to the context as a function of the number of sentences in the context. On average, answer relevance was greatest when the number of sentences was two. Therefore, DQGen 2014 first computes global relevance between a candidate and content words within a two-sentence context, and filters out any candidates whose global relevance is larger than the answer's global relevance. Next, DQGen 2014 ranks the rest of the candidates by their global relevance and considers the bottom half (the most unrelated to the context), from which it chooses the candidate with the highest local relevance (the most related to the stem). This heuristic maximizes the local relevance of plausible distractors subject to limiting their global relevance.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

4 Evaluations of DQGen 2014

Sections 4.1, 4.2, and 4.3 respectively describe how we evaluated DQGen 2014 on children's responses, compared it to DQGen 2012, and evaluated it against human performance. Section 4.4 analyzes errors in DQGen's performance.

4.1 Evaluation on children's data

What happens when children encounter questions generated by DQGen 2014? Do the three types of distractors indeed indicate different types of failure? To find out, we ran DQGen 2014 on texts (narrative fiction, informational texts, and other genres) in Project LISTEN's Reading Tutor (Mostow 2013), generating questions for 282 of them. We modified the Reading Tutor to administer the generated cloze questions (randomizing the order of choices each time), and analyzed children's responses to them. Twenty-four students participated in this experiment from May 29 to June 27, 2014. The Reading Tutor estimated their reading levels to range from grade 1–6. Children encountered a total of 200 questions generated from forty-eight stories, with 118 distinct cloze questions. Thus this evaluation is based on a large number of distinct questions but sparse data on each one, in contrast to the later evaluations with a small number of questions but much more data about each one.

4.1.1 Descriptive statistics

We define the *distractiveness* of a distractor as the frequency with which students choose it over the correct answer. To analyze distractiveness, we computed the frequency of each response type within the 200 responses: correct answers (sixty-five per cent), plausible distractors (fourteen per cent), nonsensical distractors (eleven per cent), ungrammatical distractors (seven per cent), and non-responses (four per cent) where the child did not answer. This order is consistent with our expected order of distractiveness, namely plausible > nonsensical > ungrammatical.

Selecting a plausible distractor as correct: Plausible distractors are designed to make sense locally, so as to detect failure by the reader to relate the clozed sentence to the preceding context. For instance, consider this stem:

- 'In the Winter I do not hear many ____.' [birds]

Some children chose '*voices*' as the answer. This choice makes sense in the sentence, but is unrelated to the preceding context: '*Some days the birds sing. In the Spring the birds get up early. They wake up long before I do.*'

Selecting a nonsensical distractor as correct: The child who answered '*lids*' to this question presumably did not understand the word, ignored the fact that lids can't hear, or chose randomly.

Selecting an ungrammatical distractor as correct: Similarly, the child who answered '*early*' didn't know the word, ignored grammar, or chose randomly.

Null responses: In seven cases, the Reading Tutor presented the stem, but before it could present the choices, the student clicked *Goodbye* to exit in four cases, timed out in two cases, or clicked *Back* to return to the story menu in one case.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

To examine further the relation of response type to distractiveness, we correlated the frequency of each response type against the child's reading level. For this analysis, $N = 24$ students, and frequency of a given type is the percentage of a student's responses of that type. E.g., the student who encountered the most questions (nineteen) had a reading level estimated by the Reading Tutor as second grade. This student had seven correct responses (thirty-seven per cent), three plausible responses (sixteen per cent), three nonsensical responses (sixteen per cent), four ungrammatical responses (twenty-one per cent), and two non-responses (eleven per cent). Correlation with reading level was negative for the percentage of ungrammatical responses ($r = -0.41$, $p = 0.05$), consistent with better readers being more sensitive to syntax (Hensler and Beck 2006). Correlations with reading level were not significant for other response types, unsurprisingly given the limited sample size ($N = 24$ students).

4.1.2 Analysis of response types

To account for variance due to questions and students in testing for significant differences, we first used the glmer function in *R* to perform mixed-effects logistic regression with response type as a fixed effect, and stem and student as random effects. The dependent variable was whether the student chose that response type for that question. Like logistic regression, this type of model predicts a binary outcome – in this case, whether a distractor type children chose was affected – as the log odds ratio of the probability of choosing over the probability of non-choosing, but takes the variance of questions and students into consideration. Thus, the (admittedly non-independent) input observations for the regression consisted of $N = 1{,}000$ tuples, five for each response type – one tuple with an outcome of TRUE for the actual response, and four tuples with outcomes of FALSE for the other four response types. The random effects for student and stem turned out not to be statistically significant. This lack of statistical significance may imply that the results are likely to generalize to similar students and questions. On the other hand, it might just reflect low statistical power caused by insufficient data.

Dropping the non-significant effects eliminated both the random effects, yielding the logistic regression model (glm in *R*) whose coefficients appear in Table 4. As Table 4 shows, the base reference response Correct was significantly likelier than all four other response types. Table 4 does not show which if any differences among these four types are significant, but the differences in their beta coefficients suggest that responses are substantially likelier to be ungrammatical or null than plausible or nonsensical.

4.1.3 Analysis of response times

As a potentially more sensitive indicator of student performance, we also analyzed students' response times to the different types of questions, which averaged about 30 sec overall. For analysis of response times, $N = 193$ non-null responses (to 116 distinct cloze questions), because response time is undefined for non-responses. To determine which differences were not only reliable but likely to generalize to unseen

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 4. *Logistic regression model of response selection; reference base for response_type is Correct*

Model: chose ~ response_type		
Fixed effects:	β coefficient	p-value
Intercept	0.619	<0.001
Response_type = ungrammatical	−3.206	<0.001
Response_type = nonsensical	−2.710	<0.001
Response_type = plausible	−2.476	<0.001
Response_type = non-response	−3.936	<0.001

Table 5. *Best-fitting model of response time*

Model: duration ~ 1 + (1 | student)		
Random effects:	Variance:	Standard deviation:
Student	46.73	6.84
Fixed effects:	β coefficient:	t-value:
Intercept	28.99	13.93

data from similar cloze stems and students, we used mixed-effects linear regression. To find the model that fit the data best, we used backward model selection, starting with the three available predictors that we expected could affect the outcome: response type as fixed effect, and stem and student as random effects (just their intercepts, not their slopes, which our data was too sparse to estimate). We kept removing the least significant predictor (the one with the highest p-value) until doing so stopped improving model fit in a Likelihood Ratio Test. Table 5 shows the resulting model.

Random effect of individual student: Response time differed reliably by student (SD = 6.84), i.e., the best-fitting model had a distribution of random per-student intercepts (1 | student) ~ $N(0, 6.84^2)$.

No main effect of response: Response time differed only slightly, not significantly, by students' responses:

- Correct answer: $M = 29.15$ sec, $N = 130$, SD = 20.23 sec
- Plausible distractor: $M = 28.56$ sec, $N = 27$, SD = 14.46 sec
- nonsensical distractor: $M = 33.50$ sec, $N = 22$, SD = 25.93 sec
- Ungrammatical distractor: $M = 26.50$ sec, $N = 14$, SD = 10.45 sec

No random effect of stem: We would have expected significantly different variance in response time among the 116 question stems. The absence of such an effect reassuringly suggests that the results are likely to generalize to future similar questions, but it might simply be an artifact of limited sample size.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

4.2 Comparison of distractor quality: 2012 versus 2014

Did the changes motivated by the 2012 pilot study actually improve DQGen 2014? We now report an experiment to find out.

4.2.1 Comparison methodology

To compare DQGen 2012 and DQGen 2014, we adapted the blind evaluation methodology used in the pilot study. This comparison differed from the pilot study in three ways.

First, all thirteen questions in the pilot study (including the question in Figure 1) came from a single story (2006). Due to its stronger constraints, DQGen 2014 selected only one question stem for '*Tiny Invaders*,' so we needed additional questions in order to compare it to DQGen 2012. Besides, we had tuned DQGen 2014 to this text insofar as we had modified it to address problems identified in the pilot study. For a fair comparison, we needed to evaluate them on previously unseen texts. Out of the 282 Reading Tutor texts for which DQGen 2014 had generated at least one question, it had generated more than one question for twenty-six of them. We randomly chose six of these twenty-six texts to add. We had not previously run DQGen 2014 on these texts, so they had not affected its development, and hence constituted a fair basis for comparison. Appendix A.2 shows the sixteen stems that DQGen 2014 extracted from the seven texts ("Tiny Invader" plus six texts).

Second, the pilot study had presented only four choices to categorize for each stem. To compare the quality of the distractors generated by DQGen 2012 and DQGen 2014, we evaluated how often they were perceived as their intended types. This evaluation presented seven choices in random order, consisting of the correct answer, three distractors generated by DQGen 2012, and three distractors generated by DQGen 2014, without disclosing which was which. DQGen 2012 and DQGen 2014 generated the same distractor in only one case; for that question, we presented only six choices.

Third, the pilot study had also scored the overall quality of each question. We dropped this part of the evaluation because presenting seven choices for each stem precluded comparison of overall question quality between DQGen 2012 and DQGen 2014. However, as in the pilot study, we invited judges to comment on the candidate choices. There were thirteen such comments (two comments on correct answers, six comments on five distractors generated by DQGen 2012, and five comments on three distractors generated by DQGen 2014), some of which contributed to the error analysis in Section 4.4.

The Appendix shows our evaluation form. Five judges participated. All five were native English speakers and members of our research team. Two of them were also judges in the pilot study. The other three were unfamiliar with this study.

4.2.2 Result: judges' agreement with each other and with intended distractor type

As in the pilot study, we used Fleiss' Kappa to assess inter-rater agreement between categorizations by the five judges. Kappa was 0.63 for the version comparison

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 6. *Individual judges' agreement with each other (Cohen's Kappa)*

Judge	A	B	C	D	E
A	...	0.47	0.57	0.48	0.62
B		...	0.63	0.66	0.73
C			...	0.61	0.61
D				...	0.85
E					...

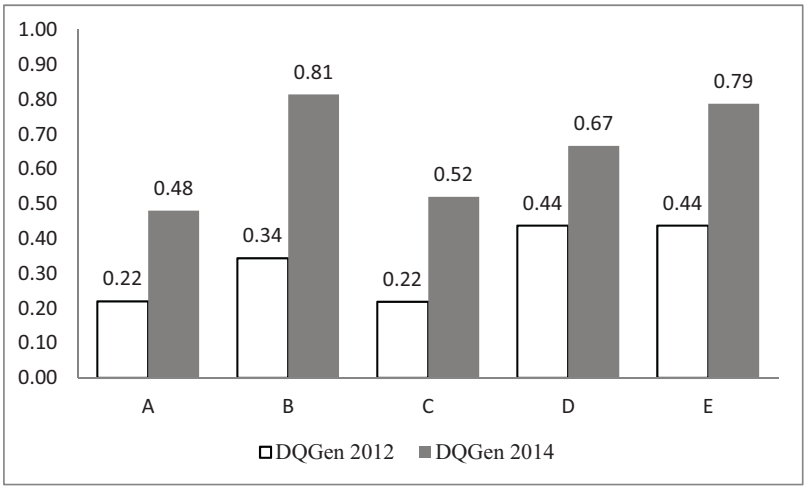


Fig. 6. Cohen's Kappa for each judge's agreement with intended distractor type in DQGen 2012 and DQGen 2014.

experiment in Section 4.2, versus 0.58 for the pilot study in Section 2.2, and indicates 'substantial agreement' according to Landis and Koch (1977).

Figure 6 shows each individual judge's agreement with intended distractor type. Cohen's Kappa measures the agreement between two judges. We compared the individual ratings with the intended types of the distractors generated by DQGen 2012 and DQGen 2014 in order to understand how often each judge agreed with them. Figure 6 makes clear that all five judges agreed more often with DQGen 2014 (with Cohen's Kappa from 0.48 to 0.81) than with DQGen 2012 (with Cohen's Kappa from 0.22 to 0.44). Figure 3 showed earlier that in the pilot study, judges agreed with intended type more often for correct answers and ungrammatical distractors than for nonsensical and plausible distractors. Figure 7 showed that the same pattern held true in the study comparing DQGen 2012 and DQGen 2014. This study used different data than the pilot study – namely the subset of stems for which both versions of DQGen generated questions, and ratings from a different set of judges. Consequently, Figures 3 and 7 show different values for DQGen 2012.

We compared judges' agreement with DQGen 2014 to their agreement with each other. As Table 6 shows, judges agreed slightly more often with DQGen 2014 (mean Cohen's Kappa of 0.65) than with each other (0.63).

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 7. *Confusion matrix comparing categorizations of DQGen 2012's and DQGen 2014's distractors*

Category:	Ungrammatical		nonsensical		Plausible		Correct	
Intended type:	DQGen 2012	DQGen 2014	DQGen 2012	DQGen 2014	DQGen 2012	DQGen 2014	DQGen 2012	DQGen 2014
Ungrammatical	**81%**	< **98%***	10%	3%	8%	0%	1%	0%
nonsensical	40%	24%	**46%**	< **69%**†	14%	8%	0%	0%
Plausible	13%	0%	46%	15%	**38%**	< **61%**‡	4%	24%

*Chi-square $p = 0.001$. †$p = 0.004$. ‡$p = 0.34$. Note: due to rounding, some rows do not sum to one hundred per cent.

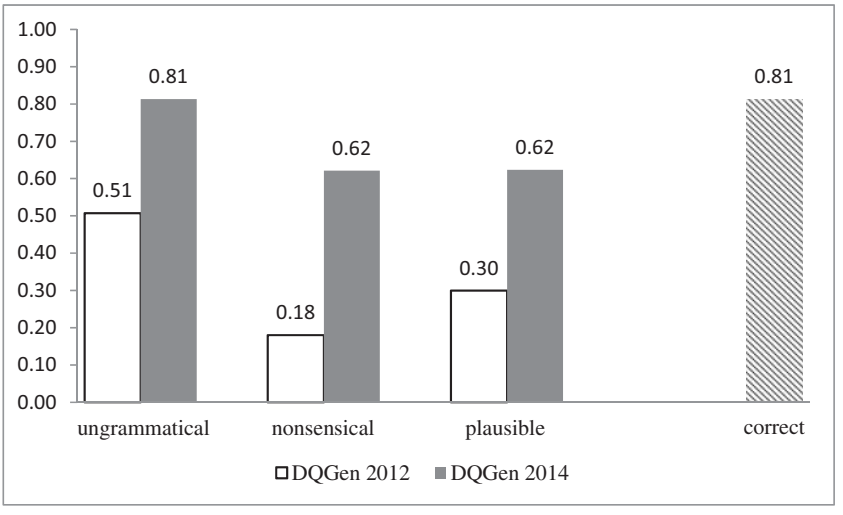


Fig. 7. Cohen's Kappa for agreement with the intended choice type in DQGen 2012 and DQGen 2014.

The confusion matrix in Table 7 shows the percentage of categorizations of distractors as ungrammatical, nonsensical, plausible, or correct for each intended type generated by DQGen 2012 and DQGen 2014. Table 7 has a row for each intended type, but no row for correct answers because they were the same for DQGen 2012 and DQGen 2014. Except for four categorizations as meaningful and one as nonsensical, the judges categorized every correct answer as Correct.

Table 7 has a pair of columns for each category, showing the percentage of categorizations that agreed with the intended type in that row. The **boldfaced** percentages along the diagonal of the matrix are significantly larger for DQGen 2014 than for DQGen 2012. As the '<' signs indicate, DQGen 2014 beat DQGen 2012 at generating distractors perceived as intended, for all three distractor types. Conversely, none of the off-diagonal percentages, representing disagreement with intended distractor type, are larger for DQGen 2014 than for DQGen 2012, except twenty-four per cent versus four per cent for the percentage of plausible distractors

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 8. *Best-fitting model of agreement; reference base for distractor_type is nonsensical*

Model: *agreement* ~ *version* + *distractor_type* + (*1*|*stem*) + (*1*|*judge*)		
Random effects:	Variance:	Standard deviation:
Stem	0.16	0.40
Judge	0.15	0.39
Fixed effects:	β coefficient:	p-value:
Intercept	0.92	0.002
Version = DQGen 2012	−1.14	<0.001
Distractor_type = ungrammatical	2.02	<0.001
Distractor_type = plausible	−0.40	0.091

categorized as Correct. That is, all of the DQGen 2014 distractors intended to be plausible were grammatical and almost all of them made sense locally, but some of them were too plausible. The error analysis in Section 4.4 discusses the most frequent errors.

4.2.3 Generalizability of results

Our comparison of distractor quality necessarily used a limited number of judges (five), texts (seven), and question stems (sixteen). Can we be confident that our results generalize to similar unseen judges, texts, and stems? To find out, we used the glmer function in *R* to fit a logistic mixed-effects model. Like the logistic mixed-effects model in Section 4.1.2, this model has a binary outcome variable – in this case, whether a judge agreed with the intended type. The model predicts the log odds ratio of the probability of agreement over the probability of disagreement. Its fixed effects are version (DQGen 2012 or DQGen 2014), distractor type (ungrammatical, nonsensical, or plausible), and their interaction. The model includes individual judge and stem as random effects in order to test whether they account for significant variance in the results. It does not include text, because stem uniquely determines text. The sample for this analysis consisted of $N = 2$ versions × 3 distractor types × 16 stems × 5 judges = 480 categorizations (including five duplicate categorizations of the lone distractor generated by both DQGen 2012 and DQGen 2014). We used the same backward model selection procedure as in Section 4.1.3 above, which resulted in eliminating the interaction term and yielded the best-fitting model shown in Table 8. This procedure left stem and judge as random effects because eliminating them produced a worse model according to a likelihood ratio test. Controlling for these random effects means that they did not account for the significant fixed effects discussed below, but the fact that the random effects were significant means that the results might differ for similar stems or judges, i.e., drawn from the same distributions.

We now relate the model in Table 8 to the agreement rates in Table 7.

Main effect of version: DQGen 2014 significantly beat DQGen 2012 overall.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Main effect of intended distractor type: Compared to the quality of their nonsensical distractors, both DQGen 2012 and DQGen 2014 generated significantly better ($p < 0.001$) ungrammatical distractors, with a trend ($p < 0.1$) toward worse plausible distractors.

No interaction of version with distractor type: That is, including such an interaction does not predict performance better than the combined fixed effects of version and distractor type.

Random effects: Performance differed reliably by stem (SD = 0.40) and judge (SD = 0.39). That is, agreement varied systematically by individual stems and judges. Consequently, we cannot assume that our results would generalize to similar stems or judges. To achieve more generalizable results, future research could improve judge selection and instructions to reduce inter-rater disagreement, and identify features of stems that cause disagreement with intended distractor types.

4.3 Evaluation of DQGen 2014 against human performance

How well did DQGen 2014 perform compared to humans? We now describe an experiment to find out.

4.3.1 Comparison methodology

To evaluate DQGen 2014 against human performance, we had to specify the task being performed and the criteria by which to evaluate it. Given a text with some sentences selected to turn into multiple choice cloze stems by deleting the last word, the task was to generate a distractor of each type – ungrammatical, nonsensical, and plausible. To enable controlled evaluation of distractors, we gave DQGen 2014 and humans the same seven texts and stems listed in Appendix A.2 and used in Section 4.2's comparison to DQGen 2012.

Our principal evaluation criterion was whether a generated distractor achieved its purpose according to human judges blind to its source (DQGen or human), its intended type (ungrammatical, nonsensical, or plausible), and the correct (original) answer. An additional evaluation criterion was time: we wanted to know how long it took humans to categorize or write each type of distractor. Besides quantifying the relative difficulty of categorizing versus writing the three types of distractors, the practical purpose of this information was to predict which would be faster – writing distractors by hand, or hand-vetting distractors generated by DQGen.

Apparatus: The pilot study and the comparison of DQGen 2014 to DQGen 2012 had presented questions in an Excel spreadsheet for judges to fill in their responses, but those studies had each used fewer than ten judges, and they did not capture timing information.

To overcome both limitations, we developed a website to run the experiment. We implemented the website in PHP and connected it to a MySQL database server that logged a timestamped event for each page entrance or exit, keyboard input, or menu selection. The database also kept track of each participant's position in the protocol

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

in order to continue at the same point after an interruption, and to avoid repeating any of the protocol.

Participants: To recruit human judges proficient in English and sufficiently knowledgeable about reading comprehension to categorize and write distractors, we posted a request to Carnegie Mellon's doctoral Program in Interdisciplinary Educational Research (www.cmu.edu/pier) and to the Society for the Scientific Study of Reading (triplesr.org). The request directed participants to the website for the experiment.

After data cleaning to filter out data from in-house software testing, failed attempts to log in, unfinished protocols, and two null categorizations, we had data for twenty-seven participants' completed experimental protocols.

Procedure: The experimental protocol consisted of logging into the experiment website, a brief introduction, the two main tasks (first categorizing, then writing), and finally a survey with a series of optional typed-input questions about various aspects of the experiment.

The introduction thanked participants for 'helping our research by doing two tasks: rating (the first task) and designing (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension.' It explained that in the first task, they would read texts containing a total of eight cloze stems, see different candidate completions of each stem, and categorize each completion as **correct**, **plausible**, **nonsensical**, or **ungrammatical.** It showed the annotated example in Figure 1 and said:

> Please classify each choice on its own merits, independently of the others.
> Your responses will be timed as a measure of the effort they require.
> Therefore you will not get an opportunity to revise them.
> Also, please try to avoid interruptions during a text.
> However, pausing between texts is fine.

In the categorization task, participants read 3–4 texts containing a total of eight cloze stems. Cloze stems appeared on a new screen with this note: 'If you need to reread the text first, please click on the *Previous* button above. Otherwise, click on one of the four buttons below to classify the following completion (independently of the others).' The button for each category included its description shown in Figure 1. Participants categorized seven candidate single-word completions, one at a time, for each cloze stem, e.g., '*The next morning, Silly Pilly was ready to go to* ___.' The seven candidates, reordered randomly for each participant, consisted of the correct answer (*school*), the three distractors generated by DQGen 2014 (*along*, *slang*, and *breakfast*), and three authored by humans (e.g., *blue*, *slip*, and *home*). The writing task was similar:

> In the second task, you will read texts that contain cloze stems. You will be prompted to type in four **1-word** completions of each cloze stem, one completion of each kind. **These words should be no harder for a child than the reading level of the text.**

To avoid problematic input such as null responses, typos, and non-words, we included code to reject them and prompt for a replacement, but these events,

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

averaging 25 sec, occurred for only eleven of the 504 human-written distractors in our data.

Assignment to conditions: All participants did categorization before writing, which we considered harder and in fact averaged about three times as long per completion. We wanted to give judges some experience in thinking about different types of distractors before writing them. The resulting order bias was mitigated by not explicitly labeling distractors by source or intended type. Although judges might conceivably have imitated the eight human- and eight DQGen-generated distractors they rated for each intended type, the fact that they classified so many of them as other than their intended types suggests that any such imitation was limited.

To avoid text-specific bias, we counter-balanced the study design so that half the participants (the 'AB' group) categorized completions for set A (stems 1–8) and then wrote completions for set B (stems 9–16), and the other half (the 'BA' group) categorized completions for the stems in set B and then wrote completions for the stems in set A.

Participants in each group categorized the same distractors generated by DQGen 2014, but they categorized different distractors authored by humans, so as to give us a more diverse sample. To limit the protocol duration, each participant categorized distractors authored by only one participant from the other group. Accordingly, we used the following algorithm to assign participants to categorize human-authored distractors.

The first participants saw distractors written by staff experienced with cloze questions. However, as soon as participants completed the protocol, the distractors written by these protocol completers became available for subsequent participants to categorize. Once a participant completed the protocol by writing distractors for set B, another participant was assigned to categorize them, and to write distractors for set A. Similarly, those distractors were eventually (if ever) categorized by some subsequent participant assigned to categorize set A and write distractors for set B, and so on.

This 'daisy-chaining' algorithm assigned each new participant to categorize cloze choices from whichever set (A or B) had been categorized so far by fewer participants who had finished the protocol. It chose human-authored distractors not yet categorized by anyone who had finished the protocol. Consequently, all twenty-seven participants categorized distractors generated by DQGen 2014 for either set A or set B. Twenty-one participants' distractors got categorized – sixteen participants with one judge per distractor, and the other five with two. Our data set contains no categorizations for the remaining six participants' distractors, either because nobody categorized them, or because we discarded data from other participants who may have categorized them but didn't finish the rest of the protocol.

4.3.2 Results: agreement with intended distractor type

Like Table 7 in Section 4.2.2, Table 9 shows the percentage of categorizations of each intended distractor type as Ungrammatical, nonsensical, Plausible, or Correct, except that it compares DQGen 2014 to humans instead of to DQGen 2012. Table 9 is based

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 9. *Confusion matrix for categorizations of DQGen 2014's and human distractors and correct answers*

Category: Intended type:	Ungrammatical			Nonsensical			Plausible			Correct	
	DQGen 2014		Human	DQGen 2014		Human	DQGen 2014		Human	DQGen 2014	Human
Ungrammatical	**93%**	>	**81%***	4%		16%	3%		1%	0%	2%
Nonsensical	14%		5%	**81%**	>	**74%**†	5%		20%	0%	1%
Plausible	2%		2%	23%		23%	**54%**	<	**63%**‡	21%	13%
Correct	0%			2%			18%			80%	

*Chi-square $p < 0.001$. †$p = 0.089$. ‡$p = 0.053$.

on 1,486 categorizations by twenty-seven judges of sixteen correct answers, forty-eight distractors generated by DQGen 2014, and 504 distractors written by twenty-one humans. Inter-rater reliability was substantial on DQGen 2014's distractors (Fleiss' Kappa = 0.66). Only forty human-authored distractors were categorized by more than one judge, namely five sets of eight distractors categorized by a pair of judges. Cohen's Kappa for each pair of judges averaged 0.46 ($N = 5$, SD 0.25), i.e., only moderate agreement, versus 0.60 (SD 0.09), close to substantial agreement, on the eight DQGen-generated distractors they both categorized, but the two means did not differ reliably on a paired T-test ($p = 0.31$).

The **boldfaced** diagonal entries in Table 9 compare the percentages of categorizations that agreed with the intended types of DQGen 2014- and human-generated distractors. To determine which differences were not only reliable but likely to generalize to unseen data from similar cloze stems and judges, we used a logistic mixed-effects model, just as in Section 4.2.3. To find the model that fit the data best, we used backward model selection, starting with five predictors we expected could affect the outcome. Three were fixed effects: distractor source, intended type, and their interaction. Two were random effects: stem and judge. We kept removing the least significant predictor (the one with the highest p-value) until doing so stopped improving model fit in a Likelihood Ratio Test. We now relate the resulting model in Table 10 to the agreement rates in Table 9.

Main effect of intended distractor type: Compared to their nonsensical distractors, both DQGen 2014 and humans generated significantly worse ($p < 0.02$) plausible distractors, with a trend ($p < 0.1$) toward better ungrammatical distractors.

No main effect of source: Surprisingly, DQGen 2014's distractor quality did not differ significantly overall from humans.'

Interaction of source with distractor type: Although DQGen 2014 and humans did not differ significantly overall, they differed for some distractor types after adjusting for the fixed effect of distractor type and the random effect of stem. DQGen 2014's ungrammatical distractors were significantly ($p < 0.001$) better than humans', its plausible distractors were probably ($p \sim 0.05$) worse than humans', and there was a trend ($p < 0.1$) for its nonsensical distractors to be better than humans'.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 10. *Best-fitting model of agreement; reference base for distractor_type is nonsensical*

Model: *agreement* ~ *distractor_type* + *source* × *distractor_type* + (*1* | *stem*)		
Random effects:	Variance:	Standard deviation:
Stem	0.12	0.35
Fixed effects:	β coefficient:	p-value:
Intercept	1.08	<0.001
Distractor_type = ungrammatical	0.40	0.098
Distractor_type = plausible	−0.53	0.014
Distractor_type = ungrammatical × source = DQGen 2014	1.11	<0.001
Distractor_type = nonsensical × source = DQGen 2014	0.41	0.082
Distractor_type = plausible × source = DQGen 2014	−0.39	0.053

No random effect of individual judge: We would have expected a judge effect if some judges were systematically worse, e.g., categorized at random. The absence of such an effect reassuringly suggests the results are likely to generalize to future similar judges – perhaps because the comparison to human used twenty-seven judges, versus the five judges used in the comparison to DQGen 2012.

Random effect of stem: Performance differed reliably by stem (SD = 0.35), i.e., the best-fitting model had a (1 | stem) ~ $N(0, 0.35^2)$ distribution of random per-stem intercepts. For some stems, judges could not tell correct answers from plausible distractors, as the error analysis in Section 4.4 below will discuss further.

4.3.3 Time analysis

We analyzed the time for humans to write and categorize distractors, both to measure the difficulty of the task, and the practicality of using DQGen 2014 to assist rather than replace human authors.

To see whether DQGen could speed up human authoring, we compared the time for humans to categorize versus write distractors. They averaged about 5 sec to categorize a choice and about 19 sec to write any type of distractor. Based on Table 1, vetting a distractor generated by DQGen and rewriting it only if unacceptable would average (5 sec) + (1 – agreement) × (19 sec) = about 10 sec, barely half of the 19 sec to write it by hand. Moreover, ninety-two per cent of distractors would match their intended type if vetting is perfect, i.e., categorizes DQGen-generated distractors properly by definition. Only seventy-three per cent of human-authored distractors do so.

To analyze the time to categorize a choice, we used mixed-effects linear regression starting with source, intended type, agreement, and their interactions as fixed effects, and stem and judge as random effects. Backward model selection led to the model in Table 11 and summarized below.

No main effects: Categorization time didn't differ significantly by source, intended type, or agreement.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 11. *Best-fitting model of time to categorize a choice*

Model: *duration* ~ *intended_type* × *agreement* + *(1 | stem)* + *(1 | judge)*

Random effects:	Variance:	Standard deviation:
Stem	3.48	1.87
Judge	1.92	1.38
Fixed effects:	β coefficient:	*t*-value:
Intercept	6.26	10.32
Intended_type = ungrammatical × agreement = agree	−1.89	−4.76
Intended_type = nonsensical × agreement = agree	−0.97	−2.37
Intended_type = plausible × agreement = agree	0.14	0.33
Intended_type = correct × agreement = agree	−1.61	−3.24

Random effects: Categorization time differed significantly by stem (SD = 1.75) and judge (SD = 1.36).

Interaction of agreement with intended type ($p < 0.001$) : Categorization was significantly faster when it agreed with choices intended to be correct (4.6 sec < 8.7 sec), ungrammatical (4.3 sec < 7.2 sec), or nonsensical (5.2 sec < 5.7 sec). For plausible distractors, categorization was slower (6.6 sec > 5.8 sec) (albeit not significantly) when it agreed with intended type, perhaps because confirming that a distractor is plausible requires additional thought.

4.4 Error analysis of DQGen 2014

We now analyze DQGen 2014's most frequent errors as measured by the percentage of categorizations that disagreed with intended distractor type. For this purpose, we pooled the 240 categorizations of DQGen 2014's distractors by the five judges in Section 4.2's comparison to DQGen 2012 with the 647 categorizations by the twenty-seven judges in Section 4.3's comparison to human performance.

DQGen 2014's ungrammatical distractors were categorized as their intended type ninety-four per cent of the time, versus seventy-eight per cent for nonsensical distractors and only fifty-six per cent for plausible distractors. To shed light on why, Section 4.4.1, 4.4.2, and 4.4.3 respectively examine the specific nonsensical, plausible, and human-authored distractors most miscategorized.

4.4.1 nonsensical distractors

nonsensical distractors were miscategorized most often (seventeen per cent) as Ungrammatical. The nonsensical distractor miscategorized as Ungrammatical the most often (eighteen of its nineteen categorizations) was '*share*' in:

- 'We nip if they stray too far from ____.' [home]

This example illustrates an issue that arises when a distractor can have more than one POS. DQGen 2014 chooses nonsensical distractors to have the same POS as the correct answer, in this case the noun '*home*.' The word '*share*' is often a verb, but

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

can also be used as a noun. Perhaps judges categorized it as Ungrammatical because they considered only its verb POS. Another possibility is that they thought of its noun sense, but considered it a countable noun and hence ungrammatical unless preceded by '*a*' or '*the*.'

More generally, the boundary between syntax and semantics may be blurry and vary by individual. To illustrate, consider the nonsensical distractor '*slang*' generated by DQGen 2014 for this sentence:

- 'the next morning, Silly Pilly was ready to go to ____.' [school]

The word '*slang*' is a noun, but the verb '*go to*' also imposes semantic constraints: the grammatical object of '*go to*' should be a location or event (at least when used literally rather than metaphorically, as in '*go to pot*'). This constraint may explain why seven of nineteen judges categorized '*slang*' as Ungrammatical.

The point illustrated by this example has a methodological implication for similar future studies: namely, instructions to judges may need to define 'ungrammatical' more precisely. It also has a technological implication: generating distractors to distinguish among sensitivity to surface-level syntax (e.g., POS and number agreement), semantic dependency relations (e.g., constraints on semantic role fillers, such as verb-object compatibility), and deeper types of coherence (e.g., based on inference from a situation model) would require further natural language processing such as semantic role labeling or inferring dependency relations.

Although nonsensical distractors were miscategorized most often as Ungrammatical, they were miscategorized six per cent of the time as Plausible. The nonsensical distractor miscategorized the most often (nine of its eighteen categorizations) as Plausible was '*pressure*' in:

- 'And now, each in our way, and with God's help, we must answer the ____.' [call]

Half the judges categorized this metaphorical use of '*pressure*' as Plausible. Why did DQGen 2014 consider it nonsensical? It assumes that every meaningful sequence of five words occurs often enough (at least forty times) in Google's trillion-word corpus to appear in Google N-grams, so a sequence that does not must be nonsensical. This example illustrates the fallibility of this heuristic assumption: even the 3-gram '*answer the pressure*' is not in Google N-grams. Realizing that '*answer the pressure*' is meaningful would require deeper natural language understanding, especially of metaphorical usage.

4.4.2 Plausible distractors

Plausible distractors were miscategorized twenty-one per cent of the time as nonsensical (i.e., not plausible enough) and twenty-two per cent of the time as Correct (i.e., too plausible).

The plausible distractor miscategorized most often as Correct (seventeen of its nineteen categorizations) was '*democracy*' in this line from *Bill Clinton's First Inaugural Address, Wednesday, January twenty first, 1993*:

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

- 'our people have always mustered the determination to construct from these crises the pillars of our ____.' [history]

The actual correct answer was '*history*,' but '*democracy*' really fits just as well. This example illustrates a drawback of defining the correct answer as the word used in the original sentence: this heuristic occasionally leads to a trick question, i.e., one with a distractor just as valid as the original word. DQGen 2014 assumes that the correct answer fits better than topically unrelated distractors. This assumption fails when the topicality constraint fails to disqualify a plausible distractor.

Similarly, sixteen and fifteen, respectively, of nineteen judges miscategorized '*globe*' as Correct in two sentences earlier in the same speech:

- 'Now, the sights and sounds of this ceremony are broadcast instantaneously to billions around the ____.' [world]
- 'We earn our livelihood in peaceful competition with people all across the ____.' [earth]

As two judges commented, '*globe*' fits in both cases. In both examples, the distractor is a synonym for the correct answer. An obvious improvement is to disqualify synonyms for the correct answer as plausible distractors.

Two plausible distractors were tied for being miscategorized most often as nonsensical (twelve of nineteen categorizations):

- 'ride' in 'Corky's had a long ride and she needs a ____.'[run]
- 'computer' in 'It took six great big strong guys to load it all into the ____.' [truck]

In both cases the final 5-gram preceding the period, but with the distractor filled in, occurs in Google N-grams, and the distractor is topically related to the stem. In the first case, '*ride*' sounds awkward because it also occurs earlier in the same sentence, and seems nonsensical because a dog who just had a long ride wouldn't need one. In the second case, '*computer*' seems nonsensical because loading something into a computer doesn't require physical strength. Both distractors are nonsensical because they contradict world knowledge, so detecting them as such would presumably take deeper interpretation.

4.4.3 Human-authored distractors

This paper focuses on automated generation of good distractors, but the human-authored distractors provide some useful insight because there were several of each type for each stem. As in Sections 4.4.1 and 4.4.2, we examine the cases with the most miscategorizations, except that they were distributed over multiple human-authored distractors of the same type, not just one distractor.

There were six cases of ungrammatical human-authored distractors miscategorized as Plausible in this sentence:

- 'Now, the sights and sounds of this ceremony are broadcast instantaneously to billions around the ____.'

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Six of thirteen judges classified '*brave*,' '*flows*,' '*light*,' '*politics*,' or '*run*' as nonsensical here. Evidently, the judges parsed them as nouns, but their authors did not (except *politics*). DQGen 2014 considers alternative parts of speech, which may help explain why it outperformed humans in generating ungrammatical distractors.

There were also six cases of human-authored plausible distractors miscategorized as nonsensical in this sentence:

- 'Their miserable shack was transformed into a luxurious ____.'

Six of thirteen judges categorized '*car*,' '*cushion*,' '*hovel*,' and '*train*' as nonsensical here. All of these distractors are countable concrete nouns and hence legitimately plausible in the local context, yet the judges considered them nonsensical.

Conversely, there were six cases of human-authored nonsensical distractors miscategorized as Plausible in:

- 'We nip if they stray too far from ____.'

Those distractors were '*England*,' '*heaven*, '*rivers*', '*sheep*,' and '*water*,' each of which makes sense out of context. Thus, here it was the authors who were apparently unable to ignore the preceding context in deciding if distractors are nonsensical.

DQGen 2014 by its very design disregards the context when generating nonsensical or ungrammatical distractors. In contrast, writing or judging nonsensical distractors is evidently a difficult task for humans who know the context, because they have trouble disregarding it. Another possibility is that they simply ignored or forgot the qualifier 'meaningful by itself' in the first place, and therefore did not even try to disregard the preceding context.

One implication is that depriving humans of the context would make both tasks easier for them. Another implication is that some of the sixteen distractors generated by DQGen 2014 to be plausible but classified by humans as nonsensical in sixty-two of 196 categorizations might in fact be plausible out of context. Thus, depriving humans of context might not only make it easier for them to judge plausibility, but increase their accuracy as well.

5 Relation to prior work

Sections 5.1, 5.2, and 5.3 respectively relate this research to purposes, methods, and evaluations of prior work.

5.1 Purposes

What a question tests depends on when it is asked, e.g.,:

- Before reading a text: tests prior knowledge
- Just before a sentence: tests inference from context
- While reading a sentence: tests inference from the clozed sentence itself
- Just after a sentence: tests comprehension of the sentence, not necessarily integrated with context

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

- Just after a paragraph: tests comprehension of the paragraph with local integration of meaning
- Just after reading the text: tests comprehension of key points and integration across the text
- Delayed posttest: tests retention of content

DQGen 2014 differs from prior work on generating multiple choice cloze questions in terms of when and why to ask them. Prior work has generated questions from isolated sentences (e.g., to test vocabulary (Pino, Heilman and Eskenazi 2008)), questions to answer after reading a text (e.g., to test story comprehension (Huang *et al.* 2012)), or questions to answer without reading a particular text (e.g., to test domain knowledge (Mitkov *et al.* 2009; Aldabe and Maritxalar 2010)). In contrast, DQGen 2014 inserts questions into connected text to test readers' comprehension while reading connected text.

The most closely related work was by Mostow *et al.* (2004). Their Reading Tutor dynamically generated multiple choice cloze questions to test children's comprehension of randomly chosen sentences while reading a story. It randomly chose an approximate level of difficulty (sight, easy, hard, and defined) for which word to delete from the sentence, and which words to choose randomly from the same story as distractors. The deleted word could be anywhere in the sentence, which could disrupt reading the sentence. Likewise, the sentence could be anywhere in a paragraph, which could disrupt inter-sentential processing. To minimize both types of disruption, DQGen 2014 deletes only the last word of the last sentence in a paragraph.

DQGen 2014 generates multiple types of distractors designed to diagnose different types of comprehension failure. Goto *et al.* (2010) also generated multiple types of distractors. They used a training corpus of existing cloze questions to learn how to select sentences to turn into cloze questions, words to delete, and types of distractors distinguished by their relation to the answer word: inflectional (e.g., *ask* → *asked*); derivational (e.g., *work* → *worker*); orthographic (e.g., *circulation* → *circumcision*); and semantic (e.g., synonyms and antonyms).

DQGen 2014 generates plausible distractors to test inter-sentential processing by considering their relation not only to the stem but to the sentences that precede it. Likewise, Huang *et al.* (2012) generated reading comprehension questions to test a reader's understanding of inter-sentential relations, but by exploiting co-referential relations between sentences. They used the Stanford noun phrase coreference resolution system (Raghunathan *et al.* 2010) to determine whether two expressions refer to the same entity in real life.

5.2 Methods

In recent years, there has been wide interest in the natural language processing community about how to generate questions automatically. Question generation and shared task workshops (Piwek and Boyer 2012) were held in 2008 (hosted by NSF), 2009 (co-located with AIED), 2010 (QG-STEC; (Rus *et al.* 2010)), and 2011 (as an AAAI Symposium). The aim of the task is to generate a series of

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

questions based on the raw text of sentences or paragraphs. Question types include why, who, when, where, when, what, which, how many/long, and yes/no. Generally, the procedure of question generation can be characterized in three phases: content selection, identification of a question type, and question formulation. In DQGen, the analogues of these three components are as follows. First, content selection consists of deciding which sentence to use as stem, and which word to delete. Second, the question type is always multiple choice cloze, which is not one of the question types listed above. Finally, in place of question formulation, DQGen has the problem of distractor generation.

There has been considerable research on automatic generation of multiple choice cloze questions to test vocabulary, grammar, and comprehension. A multiple choice cloze question is constructed by picking a sentence to turn into a test item, deleting part of it (typically a single word such as a target vocabulary word), and selecting distractors for it. Some researchers have worked on both automatic generation of multiple choice cloze questions about vocabulary (Pino *et al.* 2008; Agarwal *et al.* 2011a) and question generation (Heilman and Smith 2009, 2010; Agarwal, Shah and Mannem 2011b), but the question generation focused only on generating questions (stems) and not on distractor selection.

Previous work on automated generation of cloze questions has focused on the individual sentences from which they are generated, and hence has not identified context-level constraints on distractors. However, prior work has identified various lexical and sentence-level constraints.

5.2.1 Lexical constraints

Prior work has selected distractors similar to the answer word in various ways. A common strategy is to choose distractors with the same POS and approximate frequency as the answer word (Coniam 1997; Brown, Frishkoff and Eskenazi 2005; Liu *et al.* 2005; Correia *et al.* 2010). Correia *et al.* (2010) and Huang *et al.* (2012) preferred distractors with the most similar spellings to the correct answer out of all words with the same POS, based on their Levenshtein distances to it.

Smith, Sommers and Kilgarriff (2008) looked for distractors semantically similar to the answer word, based on distributional similarity. Mitkov *et al.* (2009) considered seven types of similarity: phonetic similarity; other words from the same text; four WordNet-based measures of similarity; and distributional similarity. Their measure of distributional similarity used the same British National Corpus corpus as DQGen 2014, but instead of co-occurrence frequency within a simple context window of ± 3 words, they used co-occurrence within the same dependency relation computed by the FDG dependency parser (Tapanainen and Järvinen 1997). However, both Smith *et al.* (2008) and Mitkov *et al.* (2009) computed similarity of a candidate distractor only to the correct answer, not to the sentence or preceding context as DQGen 2014 does.

Some studies also took common student mistakes into consideration, a recommended guideline for writing multiple-choice questions (Haladyna *et al.* 2002). Liu *et al.* (2005) added a culture-dependent strategy for generating distractors: choose English

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

words with semantically similar translations in the learner's native language to the translation of the answer word. Aldabe *et al.* (2007) included students' common mistakes as candidate distractors. Besides real words, Correia *et al.* (2010) also considered misspellings of the answer using a table of common spelling mistakes.

In contrast to work on single-word distractors, Gates *et al.* (2011) generated phrase-type distractors. They generated questions from a dictionary definition of a target vocabulary word. Rather than delete the answer, they parsed the definition, deleted a phrase from it, and chose distractors with the same syntactic phrase type from definitions of other words, filtered to exclude synonyms of the answer. Mitkov *et al.* (2006; 2009) chose noun phrases as answers and distractors for multiple choice questions to test domain knowledge, but they were not cloze questions.

5.2.2 Sentence constraints

Some work considered how well distractors fit into the cloze sentence.

Lee and Seneff (2007) selected a preposition distractor based on collocations. If a candidate preposition co-occurs with the next word in the sentence but does not co-occur with the preceding word, it qualifies as a distractor.

Pino *et al.* (2008) selected distractors that made the completed sentence grammatical and tended to co-occur with the words in the sentence, i.e., were topically relevant, but semantically distant from the answer as measured by WordNet.

To verify the grammaticality and semantic correctness of an artificial sentence, Sumita *et al.* (2005) and Lin *et al.* (2007) used the Google search engine. A sentence found on the web was probably produced by a human being, is likely grammatical, and hopefully makes sense. However, this approach is insufficient because it is hard to find an exact matching sentence for any completed cloze sentence due to the sparseness of natural language.

Like DQGen 2014, Aldabe *et al.* (2009) also considered local context when choosing distractors. They used an N-gram language model to predict the probability of occurrence of a distractor after its preceding words, and chose more probable words, which are presumably likelier to seem plausible. Somewhat similarly, Zesch and Melamud (2014) learned topic-sensitive lexical inference rules from a corpus, e.g., the dependency triple '<N:subj> <V:*acquired*> <N:obj>' implies '<N:subj> <V:*purchased*> <N:obj>' if the relatedness of subj and obj exceeds some threshold, where the relatedness metric is based on their pointwise mutual information and a topic model of all the words that fill these semantic roles in the corpus. They used these rules to reject candidate distractors that could be inferred from the triple. Thus, they would reject '*purchased*' as a distractor for '*acquired*' in '*Microsoft ___ Skype for 8.5b dollars*,' but allow it as a distractor in '*Children ___ skills quickly*.'

5.3 Evaluation methodologies

Some prior work has evaluated automatically generated multiple choice cloze questions in the aggregate based on student performance. For example, Mostow *et al.* (2004) demonstrated the collective psychometric reliability and validity of such

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

questions in predicting students' scores on a published test of reading comprehension, even though most questions were seen by only one student. Mitkov *et al.* (2006, 2009) used the conventional, more labor-intensive approach of administering generated questions to 243 students on a paper test. They analyzed the individual item difficulty, discriminability, and utility of each generated question based on how many students answered it correctly, and whether their total scores fell above or below the median. They used the results to evaluate the various types of distractors they generated.

Other work has relied on human judges to evaluate individual questions. One evaluation method simply asks experts to decide whether automatic generated questions meet test purposes (Liu *et al.* 2005; Goto *et al.* 2010; Gates *et al.* 2011). These judgments are subjective, so inter-rater agreement is important to measure but sometimes difficult to achieve.

A more quantitative measure of quality is to compare human-authored questions with system-generated ones, without knowing which is which (Pino *et al.* 2008). The judgments are still subjective, but avoid bias for or against automatic generation.

To measure efficiency of generation, one study (Mitkov *et al.* 2006) had three experts estimate their time to write multiple choice cloze questions for different texts with versus without automated assistance. Our study was much more tightly controlled, comparing distractors generated by DQGen 2012, DQGen 2014, and humans for the same sixteen cloze stems. We measured a distractor's quality as the percentage of categorizations that agreed with its intended type. We measured efficiency based on the exact logged time to write or categorize a distractor.

6 Conclusion

We now summarize this work's contributions, limitations, future work, and applications.

6.1 Contributions

This paper reports work on developing, pilot-testing, refining, and evaluating a method for generating multiple choice cloze questions to test students' comprehension while reading.

Unlike previous cloze question generators, some of which also use multiple types of distractors, DQGen 2014 is explicitly designed to generate diagnostic distractors: it uses ungrammatical, nonsensical, and plausible distractors to detect failures of syntactic, semantic, and intersentential processing, respectively. Moreover, in contrast to previous systems that consider only individual sentences out of context, DQGen 2014 takes preceding sentences into account in generating plausible distractors. We specified DQGen 2014's lexical, sentence, and relevance constraints, explained how some of them emerged from pilot-testing and error analysis, and described how DQGen 2014 combines them algorithmically to generate the three types of distractors.

The paper also contributes to automated diagnostic assessment of children's reading comprehension by analyzing children's responses to questions generated

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

by DQGen 2014, and by comparing DQGen 2014 against human performance in generating each type of distractor. We measured performance by how often each distractor was categorized as its intended type. Previous evaluations of automatically generated cloze questions relied on expert critiques or crowdsourced human performance at answering them. Our study was much more tightly controlled by having human judges categorize and write distractors for the same sixteen cloze stems, and logging the time to read passages and categorize and write distractors. Moreover, to obtain results likely to generalize to similar stems and judges, we used mixed-effects models to analyze 1,486 categorizations by twenty-seven judges of sixteen correct answers, forty-eight distractors generated by DQGen 2014, and 504 distractors authored by twenty-one humans.

Surprisingly, DQGen 2014 did not differ significantly overall from human performance. DQGen 2014 was better at generating ungrammatical and nonsensical distractors, but its plausible distractors were too plausible, i.e., categorized as correct answers eighteen per cent of the time. We showed that vetting DQGen 2014's output and writing distractors only if needed, rather than writing them all, would take only half the time and yield better distractors.

Besides the methods and evaluations, contributions of this work include insights about the shallow methods used. First, the combination of the N-gram filter and local relevance usually suffices to produce sentences that are at least locally coherent. Second, and more intriguingly, candidate plausible distractors with high local and global relevance scores tend to be sensible completions of the stem – even though they do not exploit information about the correct answer, and are based on a very shallow representation of the meaning of the paragraph. This finding is surprising insofar as one would expect good performance on such items to require a deep representation such as the situation model constructed by human readers (Graesser and Bertus 1998). Third, choosing from candidates whose global relevance scores fall in the bottom half of candidates with scores worse than the correct answer is fairly effective at eliminating distractors that are too plausible, i.e., unfair because they are arguably correct even though not the original text word. In short, a parser, a large N-gram corpus, and a simple relevance measure typically suffice to produce intelligent choices for a cloze question without really understanding it.

6.2 Limitations

Error analysis elucidated performance differences: DQGen 2014 considers each possible POS and distinguishes local from contextual plausibility, sometimes better than humans. However, DQGen 2014 had at least two problems in generating plausible distractors. Judges categorized some of them as nonsensical because they did not make sense in the completed sentence, despite appearing topically relevant to it based on lexical similarity, and ending with an N-gram in the Google N-grams corpus. Judges categorized others as correct answers because they made sense in the larger context despite appearing topically irrelevant. DQGen 2014 needs stronger heuristics or linguistically deeper methods to reliably generate distractors both plausible in the local context and implausible in the larger context.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Having multiple judges categorize and write distractors enabled controlled comparison of both the quality of each type of distractor and the time to write or categorize them. However, it exposed the influence of the preceding text on judges' ability to distinguish nonsensical from plausible distractors. Future studies should eliminate this influence by having humans categorize or write distractors for a cloze stem before seeing the context that precedes it, and only then decide which plausible distractors do not fit the context.

As Section 5.1 discussed, DQGen generates cloze questions for a reader to answer while reading a text. Asking those questions after a text instead might enable spuriously high performance based merely on the reader's verbatim recall or recency of seeing the correct answer, rather than actual comprehension. One potential approach to this problem is to replace the correct answer with one of the 'too-plausible' distractors discussed in Section 6.1, on the assumption that it is a valid answer. The effectiveness of this approach would of course depend on the empirical accuracy of this assumption in practice.

6.3 Future challenges

Limitations of our evaluations and of DQGen 2014 itself leave ample room for future research.

We evaluated just the distractors, not the cloze stems themselves, nor the overall quality of the resulting questions in diagnostic comprehension assessment. To enable controlled comparison of distractors, we took the stems as givens, so we did not evaluate the percentage of sentences turned into cloze stems (yield). DQGen 2014's yield was considerably lower than DQGen 2012's because its stronger constraints disqualified so many distractors, and only paragraph-final words were even considered as candidate answers to delete in the first place. The problem of identifying appropriate words to delete has received recent attention (Agarwal *et al.* 2011a; Becker, Basu and Vanderwende 2012; Niraula *et al.* 2014). However, deleting words earlier in the paragraph poses a greater disruption to the flow of reading, and so may especially harm the comprehension of readers who have trouble maintaining that flow even when the text is intact.

Future work should improve DQGen along a number of dimensions. It should refine the methods for selecting stems and ungrammatical, nonsensical, and plausible distractors. For example, in selecting stems it should use stronger criteria for which sentences to turn into end-of-paragraph cloze questions than just the ability to generate a distractor of each type. Likewise, although it already surpasses human performance in selecting ungrammatical distractors, perhaps it might do even better by requiring not only that the distractor have the wrong POS, but that the resulting POS sequence never (or seldom) occurs in a POS n-gram corpus, as a reviewer suggested. Better methods for selecting nonsensical and plausible distractors are needed to rival or surpass human performance.

Future work should also identify higher level criteria for the generated cloze items to satisfy, and develop constraints to enforce them. One criterion is informativeness: what do wrong answers reveal about comprehension? DQGen 2014 currently

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

generates only ungrammatical, nonsensical and plausible distractors. Other types of distractors, based on deeper models of comprehension processes such as inter-sentential inference, may enable more reliable and informative diagnostic assessments. Another criterion is psychometric reliability: how well does performance on a question correlate with performance on other questions about the same text? Item-level reliability is normally determined empirically – and expensively – by administering a test to a sufficiently large sample of participants, correlating their performance on each item against their overall performance, and eliminating items with weak or negative correlations. Heuristics for generating reliable items in the first place would improve the utility of the cloze items generated by DQGen.

6.4 Potential applications

One possible use of DQGen 2014 is machine-assisted generation of comprehension questions, or more precisely, human-assisted machine generation, for example with the human vetting or selecting among candidate questions generated automatically, thereby reducing the amount of human effort currently required to compose comprehension questions, and producing them more systematically.

Success in getting DQGen 2014 to produce cloze questions on a large scale would have useful applications. Periodic comprehension checks should deter children from reading as fast as they can and ignoring what the text means. Diagnostic feedback based on incorrect answers should shed light on the nature of their comprehension failures and may be valuable as feedback to teachers or as guidance to the reading tutor.

Another use for large numbers of automatically generated cloze questions is to develop methods to monitor reading comprehension unobtrusively. Student responses to cloze questions could provide automated labels for data collected while they read the preceding text. Such data could include oral reading (Zhang, Mostow and Beck 2007) or even electroencephalography (Chang *et al.* 2013). Models trained and tested on the labeled data could estimate reading comprehension based on unlabeled data – that is, without interrupting to ask questions.

References

Agarwal, M., and Mannem, P. 2011a. Automatic gap-fill question generation from text books. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics. 209 N. Eighth Street, Stroudsburg, PA 18360, USA, pp. 56–64.

Agarwal, M., Shah, R., and Mannem, P. 2011b. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics. 209 N. Eighth Street, Stroudsburg, PA 18360, USA, pp. 1–9.

Aldabe, I., and Maritxalar, M. 2010. Automatic distractor generation for domain specific texts advances in natural language processing. In H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir (eds.), *The 7th International Conference on NLP*, Reykjavk, Iceland, pp. 27–38, Berlin/Heidelberg: Springer.

Aldabe, I., Maritxalar, M., and Martinez, E. 2007. Evaluating and improving distractor-generating heuristics. In N. Ezeiza, M. Maritxalar, and S. M. (eds.), *The Workshop on*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

NLP for Educational Resources. In conjunction with RANLP07, Amsterdam, Netherlands, pp. 7–13. Borovets, Bulgaria.

Aldabe, I., Maritxalar, M., and Mitkov, R. 2009, July 6–10. A study on the automatic selection of candidate sentences and distractors. In V. Dimitrova, R. Mizoguchi, B. D. Boulay, and A. Graesser (eds.), In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED2009)*, pp. 656–8. Brighton, UK: IOS Press.

Becker, L., Basu, S., and Vanderwende, L. 2012. Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 742–51. Montreal, Canada: Association for Computational Linguistics.

Biemiller, A. 2009. *Words Worth Teaching: Closing the Vocabulary Gap*. Columbus, OH: SRA/McGraw-Hill.

Brown, J. C., Frishkoff, G. A., and Eskenazi, M. 2005, October 6–8. Automatic question generation for vocabulary assessment. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 819–26. Vancouver, BC, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., and Wood, B. 1991. *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Salt Lake City, UT: Brigham Young University Testing Services and The Department of Instructional Science.

Cassels, J. R. T., and Johnstone, A. H. 1984. The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education* 61(7): 613.

Chang, K.-M., Nelson, J., Pant, U., and Mostow, J. 2013. Toward exploiting eeg input in a reading tutor. *International Journal of Artificial Intelligence in Education* 22(1, "Best of AIED2011 Part 1"): 29–41.

Chen, W., Mostow, J., and Aist, G. S. 2013. Recognizing young readers' spoken questions. *International Journal of Artificial Intelligence in Education* 21(4): 255–69.

Coniam, D. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *CALICO Journal* 14(2–4): 15–33.

Correia, R., Baptista, J., Mamede, N., Trancoso, I., and Eskenazi, M. 2010, September 22–24. Automatic generation of cloze question distractors. In *Proceedings of the Interspeech 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Waseda University, Tokyo, Japan.

Fellbaum, C. 2012. Wordnet. *The Encyclopedia of Applied Linguistics*: Blackwell Publishing Ltd. Hoboken, New Jersey, USA.

Gates, D., Aist, G., Mostow, J., Mckeown, M., and Bey, J. 2011. How to generate cloze questions from definitions: a syntactic approach. In *Proceedings of the AAAI Symposium on Question Generation*, pp. 19–22. Arlington, VA, AAAI Press.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., and Yamada, T. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal (KM& EL)* 2(3): 210–24.

Graesser, A. C., and Bertus, E. L. 1998. The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading* 2(3): 247–69.

Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement In Education* 15(3): 309–34.

Heilman, M., and Smith, N. A. 2009. *Question Generation Via Overgenerating Transformations and Ranking* (Technical Report CMU-LTI-09-013). Pittsburgh, PA: Carnegie Mellon University.

Heilman, M., and Smith, N. A. 2010, June. Good question! Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 609–17. Los Angeles, CA, Association for Computational Linguistics.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Hensler, B. S., and Beck, J. E. 2006, June 26–30. Better student assessing by finding difficulty factors in a fully automated comprehension measure [best paper nominee]. In K. Ashley and M. Ikeda (eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, pp. 21–30. Jhongli, Taiwan, Springer-Verlag.

Huang, Y.-T., Chen, M. C., and Sun, Y. S. 2012, November 26–30. Personalized automatic quiz generation based on proficiency level estimation. In *Proceedings of the 20th International Conference on Computers in Education (ICCE 2012)*, pp. 553–60. Singapore.

Huang, Y.-T., and Mostow, J. 2015, June 22–26. Evaluating human and automated generation of distractors for diagnostic multiple-choice cloze questions to assess children's reading comprehension. In C. Conati, N. Heffernan, A. Mitrovic, and M. F. Verdejo (eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education*, pp. 155–64. Madrid, Spain, Lecture Notes in Computer Science, vol. 9112. Switzerland: Springer International Publishing.

Kendall, M. G., and Babington Smith, B. 1939. The problem of m rankings. *The Annals of Mathematical Statistics* 10(3): 275–87.

Kintsch, W. 2005. An overview of top-down and bottom-up effects in comprehension: the ci perspective. *Discourse Processes* 39(2–3): 125–8.

Klein, D., and Manning, C. D. 2003, July 7–12. Accurate unlexicalized parsing. In E. W. Hinrichs and D. Roth (eds.), *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–30. Sapporo, Japan, Association for Computational Linguistics.

Kolb, P. 2008. Disco: a multilingual database of distributionally similar words. In *Proceedings of KONVENS-2008 (Konferenz zur Verarbeitung natürlicher Sprache)*, pp. 5–12. Berlin.

Kolb, P. 2009. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA'09*, Odense, Denmark.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–74.

Lee, J., and Seneff, S. 2007, August 27–31. Automatic generation of cloze items for prepositions. In *Proceedings of INTERSPEECH*, pp. 2173–6. Antwerp, Belgium,

Li, L., Roth, B., and Sporleder, C. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1138–47. Uppsala, Sweden, Association for Computational Linguistics.

Li, L., and Sporleder, C. 2009. Classifier combination for contextual idiom detection without labelled data, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 315–23. Singapore, Association for Computational Linguistics.

Lin, Y.-C., Sung, L.-C., and Chen, M. C. 2007. An automatic multiple-choice question generation scheme for english adjective understanding. In *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, Amsterdam, Netherlands, pp. 137–42.

Liu, C.-L., Wang, C.-H., Gao, Z.-M., and Huang, S.-M. 2005, June 29. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, Ann Arbor, Michigan, pp. 1–8. Stroudsburg, PA: Association for Computational Linguistics.

Ming, L., Calvo, R. A., Aditomo, A., and Pizzato, L. A. 2012. Using wikipedia and conceptual graph structures to generate questions for academic writing support. *IEEE Transactions on Learning Technologies* 5(3): 251–63.

Mitkov, R., Ha, L. A., and Karamanis, N. 2006. A computer-aided environment for generating multiple choice test items. *Natural Language Engineering* 12(2): 177–94.

Mitkov, R., Ha, L. A., Varga, A., and Rello, L. 2009, March 31. Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In R. Basili and M. Pennacchiotti

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

(eds.), EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics, pp. 49–56. Athens, Greece, Association for Computational Linguistics.

Mostow, J. 2013, July. Lessons from project listen: what have we learned from a reading tutor that listens? (keynote). In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik (eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, pp. 557–8. Memphis, TN, LNAI, vol. 7926. Springer.

Mostow, J., Beck, J. E., Bey, J., Cuneo, A., Sison, J., Tobin, B., and Valeri, J. 2004. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning* 2(1–2): 97–134.

Mostow, J., and Chen, W. 2009, July 6–10. Generating instruction automatically for the reading strategy of self-questioning. In V. Dimitrova, R. Mizoguchi, B. D. Boulay, and A. Graesser (eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pp. 465–72. Brighton, UK: IOS Press.

Mostow, J., and Jang, H. 2012, June 7. Generating diagnostic multiple choice comprehension cloze questions. In *NAACL-HLT 2012 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 136–46. Montréal, Association for Computational Linguistics.

Niraula, N. B., Rus, V., Stefanescu, D., and Graesser, A. C. 2014. Mining gap-fill questions from tutorial dialogues. In *Proceedings of the 7th International Conference on Educational Data Mining*, pp. 265–8. London, UK.

Pearson, P. D., and Hamm, D. N. 2005. The history of reading comprehension assessment. In S. G. Paris and S. A. Stahl (eds.), *Children's Reading Comprehension and Assessment*, pp. 13–69. London, United Kingdom, CIERA.

Pino, J., Heilman, M., and Eskenazi, M. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems*, pp. 22–34. Montreal, Canada.

Piwek, P., and Boyer, K. E. 2012. Varieties of question generation: introduction to this special issue. *Dialogue and Discourse* 3(2): 1–9.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 492–501. MIT, Cambridge, MA, Association for Computational Linguistics.

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., and Moldovan, C. 2010. The first question generation shared task evaluation challenge. In *Proceedings of the 6th International Natural Language Generation Conference*, pp. 251–7. Dublin, Ireland, Association for Computational Linguistics.

Shrout, P. E., and Fleiss, J. L. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86(2): 420–8.

Sleator, D. D. K., and Temperley, D. 1993, August 10–13. Parsing english with a link grammar. *Third International Workshop on Parsing Technologies*, Tilburg, NL, and Durbuy, Belgium.

Smith, S., Sommers, S., and Kilgarriff, A. 2008. Learning words right with the sketch engine and webbootcat: automatic cloze generation from corpora and the web. In *Proceedings of the 25th International Conference of English Teaching and Learning & 2008 International Conference on English Instruction and Assessment*, pp. 1–8. Lisbon, Portugal.

Sumita, E., Sugaya, F., and Yamamoto, S. 2005. Measuring non-native speakers' proficiency of english by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pp. 61–8. Ann Arbor, Michigan, Association for Computational Linguistics.

Tapanainen, P., and Järvinen, T. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 64–71. Washington, DC, Association for Computational Linguistics.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, pp. 252–9.

Unspecified. 2006. Tiny invaders, *National Geographic Explorer (Pioneer Edition)* http://ngexplorer.cengage.com/pioneer/.

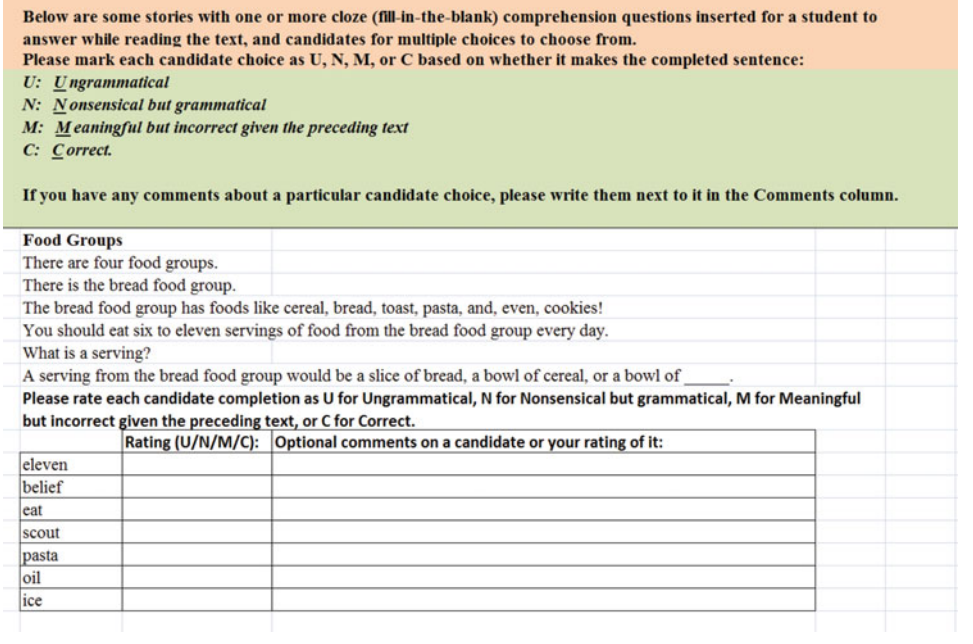
van den Broek, P., Everson, M., Virtue, S., Sung, Y., and Tzeng, Y. 2002. Comprehension and memory of science texts: inferential processes and the construction of a mental representation. In J. Otero, J. Leon, and A. C. Graesser (eds.), *The Psychology of Science Text Comprehension*, pp. 131–154. Mahwah, NJ: Erlbaum.

Zesch, T., and Melamud, O. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 143–8. Baltimore, MD.

Zhang, X., Mostow, J., and Beck, J. E. 2007, July 9–13. Can a computer listen for fluctuations in reading comprehension?. In R. Luckin, K. R. Koedinger, and J. Greer (eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, pp. 495–502. Marina del Rey, CA: IOS Press.

Appendix A

Figure 8 illustrates the spreadsheet used in Section 4.2's comparison of DQGen 2014 to DQGen 2012. The website that ran Section 4.3's comparison of DQGen 2014 to humans used essentially the same instructions for categorizing distractors but presented them one at a time in an online format.



Below are some stories with one or more cloze (fill-in-the-blank) comprehension questions inserted for a student to answer while reading the text, and candidates for multiple choices to choose from.
Please mark each candidate choice as U, N, M, or C based on whether it makes the completed sentence:

U: ***Ungrammatical***
N: ***Nonsensical but grammatical***
M: ***Meaningful but incorrect given the preceding text***
C: ***Correct.***

If you have any comments about a particular candidate choice, please write them next to it in the Comments column.

Food Groups
There are four food groups.
There is the bread food group.
The bread food group has foods like cereal, bread, toast, pasta, and, even, cookies!
You should eat six to eleven servings of food from the bread food group every day.
What is a serving?
A serving from the bread food group would be a slice of bread, a bowl of cereal, or a bowl of _____.
Please rate each candidate completion as U for Ungrammatical, N for Nonsensical but grammatical, M for Meaningful but incorrect given the preceding text, or C for Correct.

	Rating (U/N/M/C):	Optional comments on a candidate or your rating of it:
eleven		
belief		
eat		
scout		
pasta		
oil		
ice		

Fig. 8. (Colour online) Instructions for categorizing distractors in the evaluations of DQGen 2014.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

A.1 Instructions for categorizing distractors

A.2 Cloze questions from each text

To control the evaluation, we gave humans, DQGen 2012, and DQGen 2014 the same sixteen cloze stems constructed from seven texts, drawn from Project LISTEN's Reading Tutor and chosen to ensure that DQGen 2014 could generate each type of distractor for each stem. Table 12 lists each stem with the correct answer in brackets; space limitations preclude including the full texts. Below each stem is a row for each distractor type, showing the distractors of that type generated by DQGen 2012, DQGen 2014, and humans, and the percentages of distractors categorized as intended. The first parenthesized pair of percentages compares DQGen 2012 versus DQGen 2014 based on categorizations by five judges as described in Section 4.2. The second parenthesized pair of percentages compares DQGen 2014 versus humans based on twenty-seven judges as described in Section 4.3. The questions were generated in essentially the same way as the questions used in the evaluation on children's data in Section 4.2, but with very little overlap because the comparison experiments used only sixteen stems. The key point here is that the controlled experiments evaluated many fewer DQGen questions, but more than a dozen judges rated each question; in contrast, the 118 distinct questions analyzed in Section 4.2 were typically seen by at most one or two children.

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 12. *Distractors generated by DQGen 2012, DQGen 2012, and humans, with percentages categorized as intended type*

From *Silly Pilly Goes to School, by Jack Mostow:*					
1. '*The next morning, Silly Pilly was ready to go to ___.' [school]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*Bring*	*(80% versus 100%)*	*along*	*(54% versus 69%)*	*Bananas, blue, every, happy, Laugh, of, running, school, sleepy, walked, week*
Nonsensical	*daffodil*	*(60% versus 60%)*	*Slang*	*(69% versus 54%)*	*apples, bed, chocolate, church, circus, college, Disneyland, hospital, Jupiter, Mars, sign, Vapours*
Plausible	*calls*	*(0% versus 80%)*	*Breakfast*	*(92% versus 54%)*	*bed, church, college, heaven, hospital, lunch, sleep, work*
2. '*It took six great big strong guys to load it all into the ___.' [truck]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*ready*	*(100% versus 80%)*	*they*	*(100% versus 85%)*	*and, Beautiful, Christmas, funny, green, heavy, only, Pepper, running, silly, smart, warm*
Nonsensical	*birth*	*(60% versus 100%)*	*Plum*	*(100% versus 92%)*	*air, backpack, beach, bowl, Cabbage, Cheese, factory, jail, jar, mushroom, pasta, purse, soup*
Plausible	*association*	*(0% versus 40%)*	*computer*	*(31% versus 77%)*	*airplane, basement, boat, bus, car, house, suitcase, Train, trunk, Van, wagon, wheelbarrow*
From *Bill Clinton's Inaugural Speech, Part 1:*					
3. '*Now, the sights and sounds of this ceremony are broadcast instantaneously to billions around the ___.' [world]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*increasing*	*(100% versus 100%)*	*Slowly*	*(100% versus 54%)*	*brave, breezy, dirty, down, flows, light, politics, right, run, seeing, Sputter, that, Worldliness*
Nonsensical	*slick*	*(20% versus 80%)*	*chandelier*	*(92% versus 69%)*	*beanbag, cake, circus, house, Kleenex, movies, Nailbrush, Navy, puddle, table, tree, truck*
Plausible	*foot*	*(0% versus 20%)*	*Globe*	*(15% versus 69%)*	*church, city, continent, country, Galaxy, globe, hotel, microphone, Neighborhood, radio, states, world*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 12. *Continued*

4. '*We earn our livelihood in peaceful competition with people all across the ___.' [earth]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*almost*	*(100%* versus *100%)*	*and*	*(100%* versus *62%)*	*convey, feeling, jump, of, political, proper, quiet, search, Sit, stay, technology, Universal, wet*
Nonsensical	*triple*	*(0%* versus *100%)*	*Skunk*	*(92%* versus *85%)*	*bathtub, bridge, Cupboard, dog, freeway, Hospital, lake, moon, pool, room, rope, slice, table*
Plausible	*posts*	*(0%* versus *20%)*	*globe*	*(23%* versus *92%)*	*city, College, Continent, country, playground, state, store, street, town*
5. '*our people have always mustered the determination to construct from these crises the pillars of our ___.' [history]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*came*	*(100%* versus *100%)*	*bring*	*(92%* versus *54%)*	*and, awesome, bravely, coffee, Eat, enjoy, Helpful, hoping, laugh, sign, sit, slept, sneeze*
Nonsensical	*repeat*	*(20%* versus *100%)*	*skunk*	*(100%* versus *92%)*	*Baggies, Bananas, blanket, bridge, clothing, fathers, humor, ice, kitchen, mansion, salt, tent*
Plausible	*products*	*(40%* versus *0%)*	*democracy*	*(15%* versus *62%)*	*buildings, community, congress, courage, Destiny, existence, faith, fate, foundation, politics, temple*
From *Corky, by Katherine Ayres:*					
6. '*Corky's had a long ride and she needs a ___.' [run]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*sandy*	*(80%* versus *80%)*	*follow*	*(77%* versus *92%)*	*and, blue, busy, cold, Does, driving, Eating, heavy, jumpy, pretty, satisfy, short, without*
Nonsensical	*wine*	*(40%* versus *100%)*	*pocket*	*(100%* versus *62%)*	*bird, Church, coin, dog, donkey, dress, fish, mountain, paw, pony, suit, umbrella, Virtue*
Plausible	*library*	*(40%* versus *60%)*	*ride*	*(23%* versus *62%)*	*bath, break, collar, drink, friend, hug, party, snack, Swim*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 12. *Continued*

7. '*Bite tails and paws? You'd better stop or you could get in ___.*' *[trouble]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*this*	(0% versus 100%)	*barked*	(100% versus 85%)	*aggravating, Beauty, blue, confused, eating, face, good, hot, Hurt, poor, space, the, yell*
Nonsensical	*smell*	(40% versus 60%)	*brandy*	(77% versus 62%)	*bed, cars, Church, clouds, college, four, sadness, school, sleep, Trinkets, water*
Plausible	*quick*	(0% versus 40%)	*touch*	(38% versus 69%)	*cages, car, class, Discomfort, easily, favor, fights, hospital, mischief, nothing, place, shape, trouble*
8. '*We nip if they stray too far from ___.*' *[home]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*do*	(100% versus 100%)	*asked*	(100% versus 62%)	*bowl, cookies, day, dead, happy, hard, Herding, into, lonely, Plan, run, silly, up*
Nonsensical	*space*	(80% versus 0%)	*share*	(8% versus 62%)	*Bananas, beaches, beans, England, Hamburger, heaven, muscle, pizza, restaurants, rivers, sea, sheep*
Plausible	*family*	(60% versus 60%)	*reality*	(69% versus 46%)	*Chicago, Closeness, farms, home, horses, houses, land, school, us*
From *Tiny Invaders*:					
9. '*But they get your body to make ___.*' *[antibodies]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*do*	(80% versus 100%)	*with*	(92% versus 85%)	*antibodies, are, badly, float, healthy, if, in, is, of, running, sharp, Standing, therefore, warmer*
Nonsensical	*freckles*	(20% versus 100%)	*hurricanes*	(100% versus 85%)	*apples, bananas, blue, church, clouds, Cupcakes, fighting, flowers, germs, happy, jello, pillows, up*
Plausible	*energy*	(80% versus 100%)	*cholesterol*	(85% versus 62%)	*blood, breakfast, energy, healthy, improvements, invaders, Muscles, sick, sweat, trouble*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 12. *Continued*

From *Bill Clinton's Inaugural Speech, Part 3:*					
10. '*Clearly America must continue to lead the world we did so much to ___.' [make]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*old*	(*100%* versus *100%*)	*world*	(*85%* versus *92%*)	*angry, banana, blue, brain, building, chair, cheese, Die, globe, larger, old, over, stability, world*
Nonsensical	*volcano*	(*0%* versus *60%*)	*sit*	(*62%* versus *54%*)	*borrow, Clarify, clean, collapse, destroy, escape, invent, laminate, paint, skip, smother, upset, verify, write*
Plausible	*finish*	(*60%* versus *60%*)	*win*	(*77%* versus *77%*)	*bomb, build, free, guide, harm, heal, help, monitor, organize, protect, ruin, terrify, undermine*
11. '*And now, each in our way, and with God's help, we must answer the ___.' [call]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*trumpets*	(*0%* versus *100%*)	*heard*	(*85%* versus *85%*)	*called, carry, cooking, engulf, globe, go, moving, pretend, servicing, stake, stopped, Underneath, very, walked*
Nonsensical	*top*	(*80%* versus *40%*)	*pressure*	(*54%* versus *77%*)	*apples, bell, cheesecake, Door, hair, mountaintop, phone, pomegranate, post, reason, swamp, weather, wizard*
Plausible	*messages*	(*100%* versus *100%*)	*question*	(*77%* versus *46%*)	*call, guard, joy, others, phone, question, telephone, torch*
From *Jack and the Beanstalk, Part 3:*					
12. '*Their miserable shack was transformed into a luxurious ___.' [home]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*carpenters*	(*100%* versus *100%*)	*beautiful*	(*100%* versus *85%*)	*ant, are, Before, easy, flying, full, hilly, hunting, of, paintings, rich, running, slowly, touring*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 12. *Continued*

Nonsensical	*scramble*	*(40%* versus *80%)*	*volunteer*	*(92%* versus *85%)*	*apple, automobile, car, cheese, cow, goat, grape, now, octopus, Peanut, slum, tablecloth, trailer*
Plausible	*hand*	*(0%* versus *80%)*	*hotel*	*(92%* versus *31%)*	*abode, airplane, car, cottage, Cushion, hotel, house, hovel, mansion, roof, tent, tower, train*
13. '*When Jack got down to earth he called to his ___.' [mother]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*slipped*	*(100%* versus *100%)*	*not*	*(92%* versus *100%)*	*afraid, desperate, golden, her, lovely, motherly, musical, Outside, over, penning, quickly, skipped, suddenly, yellow*
Nonsensical	*hare*	*(100%* versus *60%)*	*bushel*	*(92%* versus *77%)*	*bag, book, chair, ear, grocer, Hands, lizard, music, ogre, shoes, sun, toes, zinc*
Plausible	*editor*	*(60%* versus *80%)*	*friend*	*(77%* versus *92%)*	*brother, cat, dog, friend, master, pet, teacher*
From *Food Groups:*					
14. '*A serving from the bread food group would be a slice of bread, a bowl of cereal, or a bowl of ___.' [pasta]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*eleven*	*(60%* versus *100%)*	*eat*	*(100%* versus *100%)*	*delicious, eleven, going, healthy, lively, many, run, running, tasty, under, why, yellow*
Nonsensical	*scout*	*(80%* versus *60%)*	*belief*	*(83%* versus *92%)*	*beef, cats, cups, dirt, dogs, fun, jelly, roses, serving, shoes, soup, spoons, Trees*
Plausible	*oil*	*(80%* versus *60%)*	*ice*	*(31%* versus *54%)*	*apples, carrots, cherries, chowder, coins, custard, food, granola, nuts, potatoes, soup, strawberries, sugar*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024

Table 12. *Continued*

15. '*A serving in the milk food group means a glass of milk or a cup of ___.' [yogurt]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*eat*	*(100%* versus *100%)*	*eat*	*(100%* versus *92%)*	*after, also, angry, are, because, Down, first, happy, healthy, quickly, red, the, they, well*
Nonsensical	*palm*	*(60%* versus *60%)*	*football*	*(92%* versus *69%)*	*air, balloons, barley, cats, kids, moon, nails, serving, soda, sugar, tapioca, tree, trees*
Plausible	*faith*	*(0%* versus *100%)*	*water*	*(62%* versus *54%)*	*cheese, coffee, coke, juice, milk, nothing, something, soup, tea, water, yogurt*
16. '*If you eat foods from each food group every day, you will be very ___.' [healthy]*					
Intended type:	2012:		2014:		Humans:
Ungrammatical	*day*	*(100%* versus *100%)*	*eggs*	*(100%* versus *92%)*	*church, Come, happiness, helping, is, jump, proteins, rapidly, seven, sleep, statewide, walk, water*
Nonsensical	*dizzy*	*(40%* versus *40%)*	*red*	*(85%* versus *69%)*	*backwards, baggy, cold, funny, light, purple, runny, sad, Seaworthy, serving, shy, small, smart, tall*
Plausible	*friendly*	*(80%* versus *80%)*	*competitive*	*(54%* versus *62%)*	*big, excited, favorite, fit, Full, happy, hungry, ill, short, sick, silly, strong*

Downloaded from https://www.cambridge.org/core. Texas A&M University Evans Libraries, on 10 Aug 2017 at 03:55:18, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1351324916000024