Interactive Learning Environments

ISSN: 1049-4820 (Print) 1744-5191 (Online) Journal homepage: https://www.tandfonline.com/loi/nile20

Evaluation of auto-generated distractors in multiple choice questions from a semantic network

Lishan Zhang & Kurt VanLehn

To cite this article: Lishan Zhang & Kurt VanLehn (2019): Evaluation of auto-generated distractors in multiple choice questions from a semantic network, Interactive Learning Environments, DOI: 10.1080/10494820.2019.1619586

To link to this article: https://doi.org/10.1080/10494820.2019.1619586

Published online: 21 May 2019.

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2019.1619586

Routledge
Taylor & Francis Group

Evaluation of auto-generated distractors in multiple choice questions from a semantic network

Lishan Zhang[a] and Kurt VanLehn[b]

[a]National Engineering Research Center for E-learning, Central China Normal University, Wuhan, People's Republic of China; [b]Computing, Informatics and Decision Science Engineering, Arizona State University, Tempe, AZ, USA

ABSTRACT

Despite their drawback, multiple-choice questions are an enduring feature in instruction because they can be answered more rapidly than open response questions and they are easily scored. However, it can be difficult to generate good incorrect choices (called "distractors"). We designed an algorithm to generate distractors from a semantic network for four types of multiple choice questions in biology. By recruiting 200 participants from Amazon Mechanical Turk, the machine-generated distractors were compared to human-generated distractors in terms of question difficulty, question discrimination and distractor usefulness. The machine-generated and human-generated distractors performed very closely on all the three measures, suggesting that generating distractors from a semantic network for simple multiple choice questions is a viable method.

ARTICLE HISTORY
Received 17 August 2017
Accepted 13 May 2019

KEYWORDS
Question generation; multiple choice question; semantic network; crowd sourcing; distractor evaluation; item analysis

1. Introduction

Multiple choice questions are common in both of assessment and learning contexts. However, fabricating good distractors for multiple choice questions is non-trivial and time consuming. In this article, we briefly review both the issues arising when multiple choice questions are manually generated and the existing work of machine-generated distractors. During this literature review, we also introduce some of the basic design features of our approach. After the literature review, we describe in detail our algorithm for generating distractors automatically from a semantic network, and our evaluation of the auto-generated distractors using participants from Amazon Mechanical Turk. In the final section, we discuss the significance and the limitation of our study.

1.1 Multiple choice questions

A multiple choice question has three components: stem, answer, and distractors. The answer and the distractors are also called alternatives or foils. For example,

The light reactions of photosynthesis occur in the _____

a. Cytochromes
b. Thylakoid membranes
c. Reaction centers
d. Stroma
e. Antenna complexes

CONTACT Lishan Zhang lishan.zhang3@gmail.com

"The light reaction of photosynthesis occur in the _____" is the stem of the question, and the five phrases below are the alternatives, where foil b is the correct answer and all the rest are distractors.

Multiple choice questions are widely used in many different teaching domains. Not only are they easy to score, students can answer them more quickly than typing answers to open response questions. However, a good multiple choice question requires plausible distractors so that students have to put some thought into determining the correct answer. Previous work has shown that making good multiple choice questions is a skill that is difficult to acquire and requires formal training (Abdulghani et al., 2015; Naeem, van der Vleuten, & Alfaris, 2012). Bad multiple choice questions may lower the quality of an exam (Jozefowicz et al., 2002). Indeed, although guidelines for writing good quality of multiple choice questions exist, inexperienced question writers still write poor items while trying to follow the guidelines (Abdulghani et al., 2015; Tarrant, Ware, & Mohammed, 2009). Based on Downing's analysis of four examination for medical students, 46% of the multiple choice questions contained item writing flaws (Downing, 2005). Item writing flaws are usually generated due to faculty's limited time and lack of training (Vyas & Supe, 2008).

This motivated us to develop an algorithm to auto-generate distractors for multiple choice questions. Besides saving instructor's time, certain writing flaws can also be potentially avoided when distractors are generated from an algorithm. As Tarrant, Knierim, Hayes, and Ware (2006) pointed out, question authors often make simple mistakes such as:

- Long correct answer: The correct answer is significantly longer than the distractors.
- Absolute word: Absolute terms like never and always are used in the stem or alternatives.
- Word repetition: Word repeats in the stem and the correct answer.

As long as distractor fabrication guidelines are properly encoded in the distractor generation algorithm, these flaws can be avoided.

Guidelines for generating good multiple choice questions describe features of good stems and good alternatives (Al-Rukban, 2006; McMillan & Lawson, 2001; McMillan, Hellsten, & Klinger, 2007). In this paper, we focus on generating good distractors with the assumption that the stem has been generated already, so we summarize below five features of good distractors that appeared frequently in these guidelines:

(1) All alternatives should be plausible.
(2) Alternatives should be homogeneous in content.
(3) Alternatives should be mutually exclusive.
(4) Alternatives should be free from clues about which response is correct.
(5) Alternatives should be placed in some logical order.

These guidelines were translated into constraints on the search for qualified distractors in a semantic network, which is described later.

1.2 Choice of task domain and question depth

We chose to work with questions that involved simple scientific facts, such as the one about photosynthesis above. A typical science course covers a large number of such facts.

Such factual questions could be called shallow. This limits their utility and effectiveness. In terms of Chi's ICAP framework, answering shallow questions probably should be classified as an Active engagement behavior rather than Constructive (Chi & Wylie, 2014). Nonetheless, answering shallow questions is probably more effective than simply rereading the textbook, as that would be classified as Passive, which is an even less effective activity according to the ICAP framework.

In addition, if we can manage to auto-generate good shallow multiple choice questions, human item writers should have more time for writing deep questions.

1.3 Related work on generation

Many methods have been developed for finding distractors for a question based upon the answer to the question. The most popular way is to calculate semantic similarity between a candidate distractor and the answer according to WordNet (Miller, 1995) or other lexical sources. If the similarity exceeds some threshold, the candidate would become a distractor. Pho, Ligozat, and Grau (2015) used this method to generate distractors for multiple choice questions in English language learning. For the sake of Chinese vocabulary assessment, Liu, Rus, and Liu (2018) used mixed similarity strategy to capture the characteristic (such as character glyph, pronunciations and semantic meaning) of the Chinese characters to generate good quality of multiple choice questions.

Other work that generated multiple choice questions from ontologies used the questions for teaching English (Alsubait, Parsia, & Sattler, 2016; Brown, Frishkoff, & Eskenazi, 2005; Lee & Seneff, 2007; Lin, Sung, & Chen, 2007). The generation algorithms were mainly focused on finding semantically similar words and used knowledge bases like WordNet to generate distractors. Cubric and Tosic (2011) implemented a graphical user-interface to make this type of generation process interactive. For vocabulary learning, generating and selecting distractors based on semantic similarity makes considerable sense. However, when the goal is to learn a set of facts, the distractors should be based on the facts rather than the words.

For learning factual knowledge, a more suitable method is to write rules to extract distractors from a domain-specific ontology (Al-Yahya, 2011; Papasalouros, Kanaris, & Kotis, 2008). For example, Alsubait, Parsia, and Sattler (2014) generated multiple choice questions from OWL ontologies for representing everyday knowledge and simple relationships, such as *marriedTo*(*Mark, Sara*). This can be called as pattern-based generation method. Venugopal and & Kumar (2017) put a further step. They also generated multiple choice questions using aggregation-based method. It first grouped the results of pattern-based generation, selected boarder values, and generated questions with suitable adjectives like highest or lowest. However, their generation schema focus on generating ''what is'' and ''which is'' questions with loose constraints, which would lead to too many tedious questions in the domain of photosynthesis.

1.4 Related work on evaluation

Several methods have been used for evaluating distractor or question quality. One common method is to have human instructors score the quality of each question or distractor (Karamanis, Ha, & Mitkov, 2006; Lee & Seneff, 2007; Papasalouros et al., 2008). However, even though instructors can easily recognize correct answers, they often do not know which distractors correspond to misconceptions (Chi, Siler, & Jeong, 2004; Putnam, 1987; Siler & VanLehn, 2015) nor would they know which distractors are popular. When distractors are generated from an algorithm, because the algorithm essentially describes the logic of how the distractors were generated, it becomes much clearer which misconception a distractor is connected to.

A subtle problem with using human judgement for evaluation is that it is not clear what application the judges have in mind for the questions. Questions can be used for instruction, assessment or both. If they are used for instruction, the amount of learning they tend to cause is an important consideration. For instance, a distractor that corresponds to an important, common misconception would be more valuable that a distractor that is plausible but not a symptom of a misconception. On the other hand, when questions are used for assessments, their reliability and validity are paramount concerns. It is often not clear what criteria human judges are using as the basis for their judgements.

In order to have a more objective basis for evaluating questions than human judgement, we assume that the questions will be used for assessment and thus use traditional psychometric measures of question quality, as done by (Mitkov, Ha, Varga, & Rello, 2009). Some previous studies adopted measures like item difficulty, item discrimination and distractor usefulness in their

evaluation (Liu et al., 2018). In addition to the three measures, we focus on two more measures in this study: reliability and construct validity. Measuring these requires administering a test to a sample of students. These traditional measures of test quality will be explained later.

However, the quality of a question depends on the stem as well as the distractors. Also, we need some standard against which to measure the question's quality. Both purposes can be achieved by comparing a question with machine-generated distractors to the same stem with human-generated distractors. This was done once before, by Huang and Mostow (2015), but their quality measure was based on the ability to generate certain types of distractors whereas our quality measures are more general. In particular, our evaluation involved two tests that had the same question stems, but one test had machine-generated distractors and the other had human-generated distractors. In order to measure the reliability and discrimination of the questions, we administered both tests to a large sample of students. We compared questions with machine-generated distractors to matching questions with human-generated distractors. Our hypothesis is that the machine-generated distractors will be just as good the human-generated distractors.

1.5 Contribution of this study

Compare to the existing methods of multiple choice question generation and evaluation, our main contributions are the following:

- Generated distractors by defining relatively deep heuristic rules that extracted the appropriate parts from the given semantic network, instead of only looking for similar phases of the correct answers.
- Compared the machine-generated distractors to human-generated distractors by using five classical psychometrical measures: reliability, validity, difficulty, discrimination and distractor usefulness.
- Recruited participants from Amazon Mechanical Turk to conduct the evaluation and designed the workflow to ensure the quality of the participants' responses.

2. Multiple choice question generation

Our method for generating distractors relies upon a semantic network to represent the factual knowledge that students need to know. A semantic network represents domain knowledge by describing the relations among a set of concepts. For this project, we obtained a semantic network about photosynthesis from Baral and Liang (2012). The fact that "photosynthesis produces sugar and oxygen" is represented in this semantic network by the following two relations:
(photosynthesis, result, sugar)
(photosynthesis, result, oxygen)
where *result* is a predefined predicate in the semantic network, *photosynthesis* is the subject of the predicate, and *sugar* and *oxygen* are the objects of the predicate. Photosynthesis, sugar and oxygen are concepts.

Besides relations among concepts, the semantic network also defines a hierarchical ontology of these concepts. For example, *photosynthesis* is a sub-class of *process* and *sugar* is a sub-class of *entity*. This ontology is also described with triples.

In prior work using this semantic network, we developed a method to generate photosynthesis stems and answers from it (Zhang & VanLehn, 2016). More specifically, our method generated 4 different types of questions: (1) questions about the inputs and outputs of a biological process; (2) questions about where a process was located (3) questions what a process does, and (4) and questions about how two processes are related.

Multiple choice question generation guidelines suggest that a good distractor should be a plausible answer to the question for a low-knowledge person in the domain but also be a clearly wrong answer for a domain expert. This motivated the methods we developed for generating distractors. The next section describes how distractors are generated for each type of question. There are five specific distractor generation rules in total, two for "input/output" questions, one for "where" questions, one for "what" questions, and one for "connection" questions. The rules are described in detail below and illustrated from Figures 1–5 in turn.

2.1 "Input/output" questions

Input/output questions ask about the raw materials (inputs) or the products (outputs) of a process. There were two heuristics for generating distractors for this type of questions:

(1) The raw materials of a process can be used as the distractors for a question asking for the products of the process, and vice versa.
(2) The intermediate raw materials or products of a process can be used as distractors for a question asking for the initial raw materials or the final products of the process.

Figure 1 illustrates application of the first heuristic, where Entity1 could be treated as the distractors of Entity2, and vice versa. For example, if the correct answer is that photosynthesis has energy as a product, then sunlight, which is one of its inputs, could be used as a distractor.

Figure 2 illustrates the second heuristic: Given a process, Entity2 is one of the products of the sub process of the given process, so Entity2 could be treated as the distractors of Entities1, which were the final products of the given process. For example, given the question stem "what are the two main products of photosynthesis?", the correct answer should be oxygen and sugar, and the corresponding distractors are sunlight, water and carbon dioxide.

By relaxing the constraints on the link labels (predicates), we could generate even more distractors. However, we found that this dramatically decreased the subjective quality of the distractors it generated. So we generate distractors for input/output questions using only the predicates shown in Figures 1 and 2.

2.2 "Where" questions

This type of question asks for the location of a process. Because the location is a biological structure, a good distractor should also be a biological structure and somehow related to the correct answer.

We defined a good distractor for "where" questions as an entity that is part of the same structure as the correct answer. The relation between the correct answer and a distractor is illustrated in Figure 3, where Entity1 is the correct answer, Entity2 is a distractor, and both are part of Entity3. For example, given the question stem: "where does the Calvin cycle in photosynthesis take place?", the correct answer should be stroma, and the generated distractors are thylakoid, chloroplast membrane, ribosome and DNA, which are all are part of the cell along with stroma.

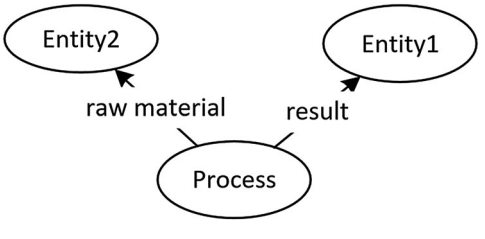

Figure 1. If Entity 1 is the correct answer, then Entity2 is a candidate distractor.

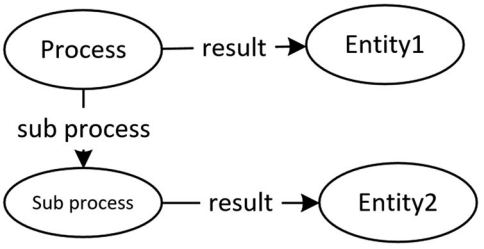

Figure 2. Entity2 can be a distractor, if Entity1 is the correct answer.

We also tried to relax the constraints on predicate names in order to generate more distractors. Unlike a similar attempt with input-output questions, relaxing predicate names tended to generate plausible distractors. Thus, our method was extended so that any predicate could replace the "part_of" predicates in the pattern of Figure 3.

2.3 "What" questions

"What" questions ask what sub-processes a parent process contains. A good distractor should be a sub-process contained by a related process to the parent process. Now the issue is to find the related processes. Two processes were considered to be related to each other if they satisfied one of the three conditions below:

(1) The two processes shared the same parent process.
(2) The two processes shared at least one product
(3) The two processes shared at least one raw material

The relations of two related processes are illustrated in Figure 4, where Process1 and Process2 represent the two related processes where one is the correct answer and the other is a distractor. Figure 4(a) illustrates condition (1), Figure 4(b) illustrates condition (2), and Figure 4(c) illustrates condition (3). Given the question stem "What are the 3 stages of Calvin cycle in photosynthesis?", some distractor examples are:

Condition (1): light reaction
Condition (2): krebs cycle
Condition (3): noncyclic photophosphorylation

2.4 "Connection" questions

Connection questions ask what products of one process are consumed by another process. Two types of entities were used as distractors. One was the set of products of the first process that were not used as the inputs to the second process, and the other was the set of raw materials of the latter process

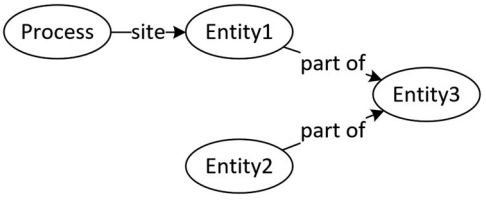

Figure 3. If Entity1 is the correct answer, then Entity2 is a candidate distractor.

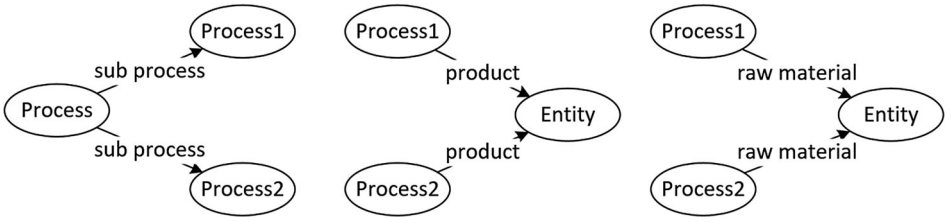

Figure 4. Process1 and Process2 are related processes. If one is correct, the other is a candidate distractor.

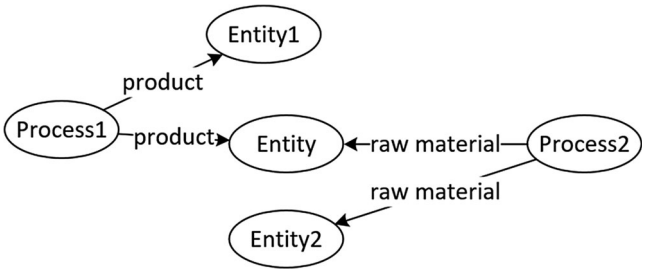

Figure 5. If Entity is the correct answer, then Entity1 and Entity2 are candidate distractors.

that were not able to be produced by the first process. The relations are illustrated in Figure 5, where Entity1 represented the first type of distractors and Entity2 represented the second type of distractors. Give the question stem "Which of the following are products of the light reactions of photosynthesis that are utilized in the Calvin cycle?", the correct answer is NADPH and ATP. An example of an Entity1 distractor is oxygen. An example of an Entity2 distractors is carbon dioxide.

2.5 Post-possessing

For various reasons explained below, not all the distractors generated from the 4 schemas above could be used in the evaluation. Thus, we use "raw distractors" for those generated by the schemas, and the "final distractors" for those selected to be included in the evaluation. This section describes how raw distractors were further processed to be final distractors.

In terms of stems, the original machine-generated open response questions could potentially be processed to be the stems in the corresponding multiple choice questions. However, we wanted to evaluate how machine-generated distractors compared to human-generated distractors. Thus, we looked for human-generated multiple choice questions whose stems were essentially asking the same thing as the machine-generated open response questions. We then replaced the original distractors of the human-generated questions with machine-generated raw distractors.

Human-generated distractors often have more than one phrase. Here is an example:
What are the two main products of photosynthesis?

(a) Carbon dioxide and oxygen
(b) Oxygen and glucose
(c) Glucose and carbon dioxide
(d) Nitrogen and glucose

Each foil of this question contains two phrases, where each phrase corresponds to one concept in the semantic network. The corresponding machine-generated distractors are single concepts: water, sunlight and carbon dioxide. In order to replicate the foil structure of the human-generated

questions, machine-generated distractors replaced concepts in the human-generated distractors. If a phase in a human-generated distractor was a correct response to the question, that phase would be kept without replacement. For the other phrases, a replacement was randomly selected out of all the machine-generated raw distractors. For the question above, (b) is the correct answer, so the human-generated concepts nitrogen and carbon dioxide where randomly replaced with one of water, sunlight and carbon dioxide. For example, one random replacement yields:

What are the two main products of photosynthesis?

(a) Water and oxygen
(b) Oxygen and glucose
(c) Glucose and water
(d) Sunlight and glucose

Sometimes the human-generated distractors are too similar to the machine-generated distractors. For example, here is a human-generated question:
The light-independent reactions of photosynthesis take place in the:

(a) Stroma
(b) Thylakoids
(c) Nucleus
(d) Mitochondria
(e) Grana

The correct answer is (a).

The machine-generated raw distractors are: thylakoids, chloroplast membrane, ribosome, thylakoid space and DNA. The two groups of distractors shared one common distractor in this case, thylakoids. Having distractors that were in both the human-generated set and the machine-generated set would make evaluating questions more complicated.

Thus, in order to make the set of machine-generated distractors disjoint from the human-generated set, common distractors were excluded before random selection. In this example, once thylakoids is excluded, there are still four machine-generated distractors left, which is just enough. However, all machine-generated questions needed the same number of distractors as their matched human-generated questions. Thus, shared distractors were allowed if necessary to equate the number of distractors. This occurred in only 2.94% of the distractors, which is so low that we ignored it in subsequent analyses.

3. Evaluation

The human-generated questions were collected by using keyword search on the Internet. The machine-generated question stems were used as the keywords. A human went through the top 20 returned web pages and harvested multiple choice questions. The collected questions were mainly from McGraw-Hill's online learning center.

Our algorithm generated 37 open response questions, and 9 of them were found to correspond to human-generated multiple choice questions. Thus the evaluation was based on the nine pairs of questions. These nine pairs should be viewed as a sample from the whole population of such pairs. Thus, this is a low-powered evaluation, and its results should be considered suggestive rather than definitive.

Before comparing the qualities of the two types of distractors, we first wanted to explore the agreement between machine-generated distractors and human-generated distractors at the level of individual concepts (raw distractors). Thus, human-generated distractors with multiple concepts,

such as "oxygen and glucose," where divided into individual concepts. Among the 9 pairs of questions, there were 44 machine-generated concepts (i.e. an average of about 5 per question) and 41 human-generated concepts. The intersection of these two sets contained 11 concepts. Therefore, 25% of the human-generated raw concepts were covered by machines, and 26.8% of the machine-generated raw concepts were covered by humans.

3.1 Experiment design rationale

In order to objectively measure the characteristics of the test questions, participants needed to take tests composed of the questions. Clearly, these would only be a sample of all possible participants. Thus, if we used only a small sample, there would be some doubt about the generality of the resulting measurements. Because we are already dealing with a small sample of questions, we wanted to test a large number of participants so that we could justifiably ignore possible participant-sampling artifacts.

Since the generated multiple choice questions were assumed to be used for assessment without particular target population, it motivated us to recruit participants via the Internet, which can reduce the biases found in school samples (Gosling, Vazire, Srivastava, & John, 2004). In order to collect enough participants, Amazon Mechanical Turk (MTurk) was used to recruit and run participants. The advantage of using this platform was that we were able to rapidly collect data from many participants in a short time with low cost. The disadvantage was that we had to detect and eliminate participants who were only pretending to take the test. Fortunately, many researchers have begun to use crowd sourcing platforms like MTurk to collect data, so methods for using it have become a hot topic.

One methodological question is whether respondents to MTurk (called Turkers) represent a different sample of the population than more traditional methods of recruitment. Both Buhrmester, Kwang, and Gosling (2011) and Paolacci, Chandler, and Ipeirotis (2010) concluded that Turkers were demographically diverse and slightly older than the respondents in traditional subject pool, but the data obtained from MTurk were at least as reliable as those obtained by traditional methods. Moreover, methods for eliminating non-compliant Turkers, like instructional manipulation checks (Oppenheimer, Meyvis, & Davidenko, 2009) and Kapcha (Kapelner & Chandler, 2010), could be used with traditional respondents as well and could improve the sample quality of both. Hauser and Schwarz (2016) contend that data from Turkers were often more representative of the population than data from participants in traditional subject pool.

By referring to previous works (Kapelner & Chandler, 2010;Oppenheimer et al., 2009), we designed three mechanisms to improve data quality. According to Mason and Watts (2010), data quality seems to be not affected by payments. We took an average payments of similar surveys available in MTurk. Each of participants got $0.50 upon finishing the experiment.

3.2 Experiment procedure

All the participants were recruited from Amazon Mechanical Turk. The experiment procedure of one participant is illustrated in Figure 6. The rest of the section described the mechanisms in detail.

First of all, qualified participants should have basic knowledge about photosynthesis. To rule out participants who had poor domain knowledge, two simple fill-in-the-blank questions were asked in the very beginning of the survey:

1. What are the raw materials of photosynthesis?
2. What does photosynthesis produce?

Participants were not expected to answer these two questions completely correctly. But they should at least point out one correct product and one correct raw material. Participants who failed to do so

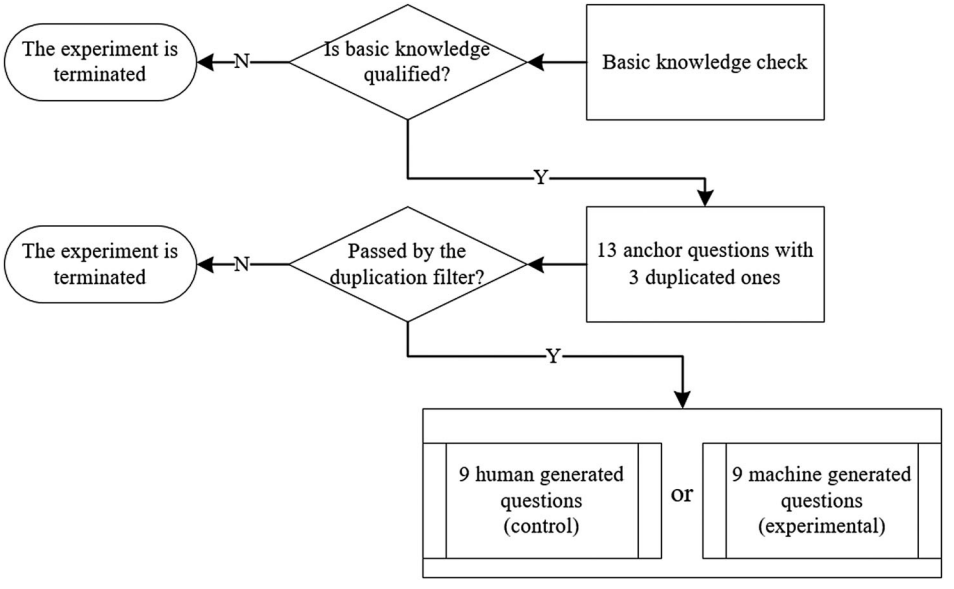

Figure 6. The experiment procedure in Amazon Mechanical Turk.

were sent to a page politely telling them about their non-qualifications. Qualified participants were sent to a page with the consent form. The test started only if they consented.

Secondly, we prevented participants from going through the questions too rapidly. A study conducted by Kapelner and Chandler (2010) showed that adding a delay for printing each word in a question was more effective than simply adding a delay for the submit button of the question. They conducted four groups of experiments to study the effect of different methods of conducting a survey of sentiment: the first group was the control group without any treatment. The second group had a reminder for each question to ask participants to respond seriously. The third group disabled the submit button for a while, so that participants were forced to slow down. The fourth group had the text fade in word by word. Trick questions were used to check the quality of the responses. Those questions could not be answered correctly without careful reading the instructions. According to this measure, the fourth group had the highest compliance. Our tests were similar to their surveys, so we faded in the questions to draw participants' attentions. Figure 7 showed how a question faded in.

Our pilot study disclosed that some participants chose the foil in the same ordinal position (e.g. the first foil) on all the multiple choice questions, which is a clear case of non-compliance. On the other hand, a compliant participant who is given the same question twice with the foils in a different order should be able to find and select the same answer to the second question as the first. So a filter utilizing duplicated questions that were close to each other was added.

In the duplication filter, two versions of the same question would appear consecutively or with one intervening question. The filter questions' foils were in different orders, but their stems were identical. Participants were explicitly asked to select the same alternative as they did before. If a participant failed to do so, the participant was asked to leave the study. There were three pairs of duplicated questions. All of them were located in the early phase of the survey. More specifically, question 4, question 6 and question 8 are duplicates of question 2, question 5 and question 7 respectively.

Figure 7. The image on the left appears for 0.5 s, then the image on the right appears.

There were two conditions: control (human-generated) and experimental (machine-generated). Each condition contained 22 questions in total. The first 13 questions, 3 of them were duplicated, were the same across conditions and they were not the same as the 9 target questions. For evaluating participants' knowledge levels and item analysis, we needed some questions that were answered by both the control and the experimental group. These questions were called "anchor questions." The nine questions that followed the anchor questions were where the manipulation occurred. The control group had nine questions with human-generated distractors and the experimental group had nine questions with machine-generated distractors.

3.3 Item analysis

Item analysis is often used to measure the quality of a test. We are interested in comparing two tests – one composed of questions with machine-generated distractors and the other test composed of the same questions but with human-generated distractors. Item analysis defines several measures of the characteristics of questions. This section reviews the measures we selected for our evaluation.

3.3.1 Reliability

A test is reliable if all the items measure the same construct. That is, correctly answering one question correlates well with correctly answering the other questions on the test. Cronbach's alpha is the standard measure of test reliability, and 0.70 or larger is considered an acceptably high value.

3.3.2 Construct validity

Although the question stems are the same, it is possible that the two tests measure different kinds of biology knowledge. That is, the constructs measured by the tests might be different. The standard method of comparing the constructs measured by two tests is to have the same students take both tests. That would not work in our case because the question stems are the same. Thus, in order to compare the construct validity of the two tests, we generated a third test which both the control and experimental groups would take. This is the anchor questions mentioned earlier. If the two tests measured the same construct, and the anchor test also measured the same construct, then student scores on the two tests should both be highly correlated with their scores on the anchor questions.

3.3.3 Question difficulty

A question's difficulty reflects the proportion of students who is able to answer the question correct. However, the higher the value is, the more difficult the question is. A question's difficulty is calculated as Equation (1)

$$Dif(Q) = 1 - \frac{N_c}{N} \tag{1}$$

where N_c is the number of students who answered the question correctly and N is the total number of students who answered the question.

A good test should have questions whose difficulties are distributed uniformly across a range. Thus, we want our distraction generator to produce questions that range over the same difficulty as the human questions and are just as uniformly distributed. Given that we have only 9 questions on each test, we do not have a large enough sample for standard measures of range and distribution, so we will convey a qualitative impression by displaying the both tests' distribution of question difficulties.

Because the questions in the two tests shared the same stems, distractors were the only source that made difficulties different. So we paired the questions based on their stems, and used Chi-square to test the significance of their differences. More specifically, for each question, students were divided into two sets in terms of their correctness on the question. The number of students

who correctly and incorrectly answered the question with human-generated distractors formed the expected values in Chi-square test. In contrast, the corresponding numbers in experimental group formed the observed values. So a 2 by 2 contingency table was used for each calculation, see Table 1 for example.

3.3.4 Question discrimination

Good multiple choice questions should allow students who know the fact to answer correctly and those who do not know the fact to answer incorrectly. Question discrimination describes how useful a question is in distinguishing the students. A common way to measure question discrimination is to use the point bi-serial (Glass & Hopkins, 1970). It can be calculated as the Pearson correlation between a student's correctness of a particular question and his/her score on all the rest questions.

Test items in a good test are expected to have their discriminations stay at relatively high levels. Therefore, we want our distraction generator to produce questions whose discriminations are as high as the human-made questions. To make the comparison, we calculated the average question discriminations of the two tests, and conducted a *t*-test.

3.3.5 Usefulness of distractors

In addition to the difficulty and discrimination of a whole multiple choice question, we are also interested in the usefulness of each distractor, as defined by (Mitkov et al., 2009). The usefulness of a distractor is measured in terms of whether the distractor had been selected by any students. Because our control and experimental groups had almost the same number of participants, a distractor was considered as useful if at least one student selected it.

3.4 Results

The evaluation took about two weeks in total. 200 Turkers passed the initial screening and started the test. 1 participant did not finish the test and was excluded in the analysis. 100 out of the 199 participants were in the control group; the remaining 99 were in the experimental group. Participants took 21.4 min on average to finish all the 24 questions. In terms of the performance on the 10 different anchor questions, participants in the control group scored 5.94 (SD = 1.848) and participants in the experimental group scored 5.87 (SD = 2.111). There was no significant difference observed ($t = 0.261$, $p = 0.79$)

The five sections below reported the difference between the machine-generated distractors and the human-generated distractors in terms of five metrics: reliability, construct validity, difficulty, discrimination and alternative usefulness.

3.4.1 Reliability

In terms of reliability, both 9-question subtests were acceptable based on Cronbach's alpha. The alpha of the experimental group was 0.701, and the alpha of the control group was 0.706. Their reliabilities were also very close to each other.

3.4.2 Construct validity

The scores of 9-question subsets from the students in the experimental group and the students in the control group correlated 0.40 ($p < 0.001$) and 0.54 ($p < 0.001$) with the scores of the corresponding anchor questions. Because both correlations were high enough, the anchor questions, the human-generated questions, and the machine-generated questions were measuring the same construct.

Table 1. A sample contingency table for Chi-square test.

	Number of students whose answers are correct	Number of students whose answers are incorrect
Control group	46	54
Experimental group	48	51

3.4.3 Question difficulty

The range of difficulty of the two tests is depicted in Figure 8. The questions with human-generated distractors have a wider range, but the distribution is not strictly uniform. Our distraction generator (upper cluster) produced questions whose difficulties are more uniformly distributed but with a smaller range. Chi-square test was run on each single question. The results were shown in Table 2. It suggested that out of the nine pairs of questions, one pair had significantly different difficulties. The average difficulties of the two tests were very close to each other when item difficulties were aggregated ($\text{Difficulty}_{\text{control}} = 0.43$, $\text{Difficulty}_{\text{exp}} = 0.44$).

3.4.4 Question discrimination

By calculating the point bi-serial, we measured the discrimination of the 9 questions in the two tests. It turned out that the questions in control group (Mean = 0.47, SD = 0.102) were slightly better than those in the experimental group (Mean = 0.41, SD = 0.117), but the difference was small and not significant ($p = 0.283$). The ranges of discrimination in the two tests were also very close. The highest and lowest discrimination in control group were 0.59 and 0.26, respectively. The highest and lowest discrimination in experimental group were 0.59 and 0.27, respectively. All the questions were at least in the acceptable level.

3.4.5 Usefulness of distractors

Usefulness of the distractors reflected similar information as discrimination of the questions, but in distractor level. There were 34 distractors in both control and experimental groups. Out of the 34 distractors, there was 1 useless distractor in the control group. In a contrast there were 2 useless distractors in the experimental group. Chi-square test was run to test the proportion of useful vs. useless distractors in the two groups, and the difference was not significant ($\chi^2 = 0.531$, $p = 0.467$).

To sum up, none of the three measures (question difficulty, question discrimination, and usefulness of distractors) reported significant difference. This suggests that there was no important difference between the two distractor generations methods based on our sample questions.

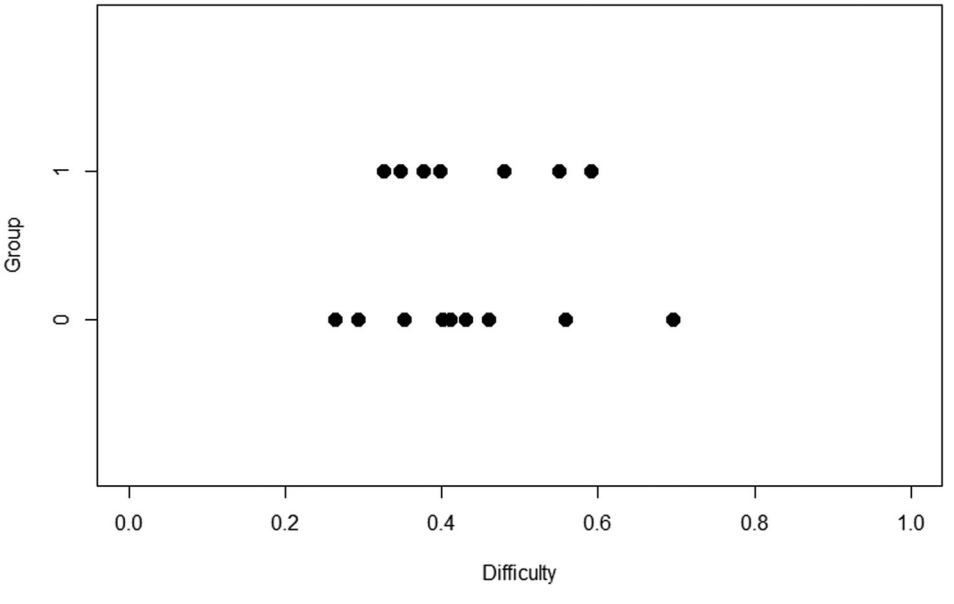

Figure 8. The range of question difficulty over the two tests. Group 0 represents control group, and group 1 represents experimental group.

Table 2. Comparison of difficulty between experimental and control group.

Question	Control	Experimental	Chi-square *p*-value
Q1	0.46	0.48	0.79
Q2	0.26	0.40	**0.045**
Q3	0.43	0.40	0.63
Q4	0.40	0.38	0.72
Q5	0.56	0.55	0.91
Q6	0.41	0.33	0.21
Q7	0.29	0.35	0.42
Q8	0.35	0.48	0.070
Q9	0.70	0.59	0.12

4. Discussion

Most of the previous works generated distractors for multiple choice questions by calculating the similarity between the correct answer and the prospective distractors (Aldabe & Maritxalar, 2010; Shah, Shah, & Kurup, 2017) or based on a hierarchical ontology whose encoded semantic relations were relatively simple (Alsubait et al., 2014; Vinu & Kumar, 2015). The main contribution of this study is the demonstration of how distractors can be generated from a semantic network that encode more complicated semantic relations and the effectiveness of this method. This generation technique can also be potentially used for constructing the authoring tools for an intelligent tutoring system. Previous works, such as CTAT (Aleven, McLaren, Sewall, & Koedinger, 2006; Terzis, Moridis, & Economides, 2013), have shown how production rules can facilitate authorization in the subjects like mathematics, but it is not clear yet how biology questions can be authorized in a similar manner efficiently.

Our entire user study was conducted though Amazon Mechanical Turk, which has been widely used in many areas for data collection, but rarely been used in evaluating educational systems. The biggest advantage of using Amazon Mechanical Turk is its speed of data collection. We ran 200 participants within two weeks. The downside was the need of dealing with unqualified and non-compliant participants. Using Turkers required using a screening test and gaming detectors because not every Turker would read the instructions carefully and treat the test seriously, even though only Turkers with good track records (historical acceptance rate >95%) were allowed to participate. Our item analysis showed that the participants produced reasonable results while screening test and gaming detectors were adopted. Other researchers who want to evaluate their educational system with a large number of participants during a short period of time may wish to consider our methodology.

In terms of the results, all of the measures suggested that the questions with machine-generated distractors were equivalent to the questions with human-generated distractors. This is good news. However, there are some limitations to our result. First, our multiple choice questions were relatively shallow. Machine-generated multiple choice questions may not completely free instructors from generating questions, but could save instructors time in fabricating shallow questions. Second, readers should be also aware that the machine-generated questions evaluated were a small sample of a large population of possible questions. Third, the generation methods were only applied to one domain, i.e. photosynthesis. Although the methods are potentially to be used in other domains with the same schema, further studies are needed for testing the generalization issues. So this work is merely the first step in understanding the characteristics of machine-generated distractors from a semantic network.

It is well known that test questions can be adaptively selected based upon a student's competence (Conejo et al., 2004; Corbalan, Kester, & Van Merriënboer, 2006; Zhang & VanLehn, 2017). If distractors were connected with students' competence represented at a fine grain size (i.e. hundreds of knowledge components), they too could be adaptively selected based on a student's competence. When distractors are generated from an algorithm as we did, they could be automatically annotated with the corresponding knowledge components. It is also possible to automatically make error-specific feedback for each distractor based on the distractor generation rules. This would be an advantage for machine-generated distractors over human-generated distractors, and would be a good direction for future work.

Notes on contributors

Lishan Zhang received a Ph.D. in Computer Science from Arizona State University (ASU) in 2015. He is an associate professor at Central China Normal University. He has published over 20 peer reviewed journal articles and conference papers. His main research area is technology-enhanced learning and assessment.

Kurt VanLehn received a Ph.D. in Computer Science from M.I.T. in 1983. He has been a professor at CMU, the University of Pittsburgh, and now at Arizona State University (ASU), where he is the Diane and Gary Tooker Chair for Effective Education in Science, Technology, Engineering and Math. He is on the editorial boards of the International Journal of A.I. in Education and Cognition and Instruction. He has published over 165 refereed journal articles and conference papers, including 11 that have won best paper awards. His main research area is intelligent interactive instructional systems. He is a fellow in the Cognitive Science Society.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Science Foundation [Grant Number DRL-0910221, IIS1123823]; Bill and Melinda Gates Foundation OPP1061281; National Natural Science Foundation of China [Grant Number 61807004].

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., … Haque, S. (2015). Faculty development programs improve the quality of multiple choice questions items' writing. *Scientific Reports*, *5*(1). Article no. 9556. doi:10.1038/srep09556

Aldabe, I., & Maritxalar, M. (2010). *Automatic distractor generation for domain specific texts*. Paper presented at the Advances in Natural Language Processing, International Conference on NLP.

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2006). *The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains*. Paper presented at the Intelligent Tutoring Systems.

Al-Rukban, M. O. (2006). Guidelines for the construction of multiple choice questions tests. *Journal of Family & Community Medicine*, *13*(3), 125.

Alsubait, T., Parsia, B., & Sattler, U. (2014). *Generating multiple choice questions from ontologies: lessons learnt*. Paper presented at the The 11th OWL: Experiences and Directions Workshop (OWLED2014).

Alsubait, T., Parsia, B., & Sattler, U. (2016). *A similarity based approach to omission finding in ontologies*. International Experiences & Directions Workshop on Owl.

Al-Yahya, M. (2011). *OntoQue: A question generation engine for educational assessment based on domain ontologies*. Paper presented at the Advanced Learning Technologies (ICALT).

Baral, C., & Liang, S. (2012). *From Knowledge Represented in Frame-Based Languages to Declarative Representation and Reasoning via ASP*. Paper presented at the KR.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). *Automatic question generation for vocabulary assessment*. Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363–387.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking Cognitive engagement to Active learning Outcomes. *Educational Psychologist*, *49*(4), 219–243.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., & Ríos, A. (2004). SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, *14*(1), 29–61.

Corbalan, G., Kester, L., & Van Merriënboer, J. J. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science*, *34*(5), 399–422.

Cubric, M., & Tosic, M. (2011). Towards automatic generation of e-assessment using semantic web technologies. *International Journal of e-Assessment*, *1*(1), 1–9.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143.

Glass, G. V., & Hopkins, K. D. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104.
Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.
Huang, Y.-T., & Mostow, J. (2015). *Evaluating human and automated generation of distractors for diagnostic multiple-choice cloze questions to assess children's reading comprehension*. Paper presented at the Artificial Intelligence in Education.
Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house médical school examinations. *Academic Medicine*, *77*(2), 156–161.
Kapelner, A., & Chandler, D. (2010). *Preventing satisficing in online surveys*. Paper presented at the Proceedings of.
Karamanis, N., Ha, L. A., & Mitkov, R. (2006). *Generating multiple-choice test items from medical text: A pilot study*. Paper presented at the Proceedings of the Fourth International Natural Language Generation Conference.
Lee, J., & Seneff, S. (2007). *Automatic generation of cloze items for prepositions*. Paper presented at the Interspeech.
Lin, Y.-C., Sung, L.-C., & Chen, M. C. (2007). *An automatic multiple-choice question generation scheme for English adjective understanding*. Paper presented at the Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007).
Liu, M., Rus, V., & Liu, L. (2018). Automatic Chinese multiple choice question generation using mixed similarity strategy. *IEEE Transactions on Learning Technologies*, *11*(2), 193–202.
Mason, W., & Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, *11*(2), 100–108.
McMillan, J. H., Hellsten, L., & Klinger, D. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction*. Boston, MA: Pearson/Allyn & Bacon.
McMillan, J. H., & Lawson, S. R. (2001). Secondary science teachers' classroom assessment and grading practices.
Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.
Mitkov, R., Ha, L. A., Varga, A., & Rello, L. (2009). *Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation*. Paper presented at the Proceedings of the Workshop on Geometrical Models of Natural Language Semantics.
Naeem, N., van der Vleuten, C., & Alfaris, E. A. (2012). Faculty development on item writing substantially improves item quality. *Advances in Health Sciences Education*, *17*(3), 369–376.
Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.
Papasalouros, A., Kanaris, K., & Kotis, K. (2008). *Automatic generation of multiple choice questions from domain ontologies*. Paper presented at the e-Learning.
Pho, V.-M., Ligozat, A.-L., & Grau, B. (2015). *Distractor quality evaluation in multiple choice questions*. Paper presented at the Artificial Intelligence in Education.
Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal*, *24*(1), 13–48.
Shah, R., Shah, D., & Kurup, L. (2017, January). *Automatic question generation for intelligent tutoring systems*. Paper presented at the International Conference on Communication Systems.
Siler, S., & VanLehn, K. (2015). Investigating micro-adaptation in one-to-one tutoring. *The Journal of Experimental Education*, *83*(3), 344–367. doi:10.1080/00220973.2014.907224
Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, *6*(6), 354–363.
Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, *9*(1), 40.
Terzis, V., Moridis, C. N., & Economides, A. A. (2013). Continuance acceptance of computer based assessment through the integration of user's expectations and perceptions. *Computers & Education*, *62*, 50–61.
Venugopal, V. E., & Kumar, P. S. (2017). Automated generation of assessment tests from domain ontologies. *Sprachwissenschaft*, *8*(6), 1023–1047.
Vinu, E. V., & Kumar, S. (2015). A novel approach to generate MCQs from domain ontology: Considering DL semantics and open-world assumption. *Web Semantics: Science, Services and Agents on the World Wide Web*, 34, 40-54
Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review on the optimal number of options. *National Medical Journal of India*, *21*(3), 130–133.
Zhang, L., & Vanlehn, K. (2016). How do machine-generated questions compare to human-generated questions. *Research & Practice in Technology Enhanced Learning*, *11*(1), 7.
Zhang, L., & VanLehn, K. (2017). Adaptively selecting biology questions generated from a semantic network. *Interactive Learning Environments*, *25*(7), 828–846.

Appendix

The nine pairs of questions and their raw distractors.

	Question stem	Correct answer	Machine raw distractors	Human raw distractors	Machine alternatives	Human alternatives
1	Which of the following stages is part of the Calvin cycle?	-Carbon dioxide fixation -Carbon dioxide reduction -RuBP regeneration	-Glycolysis -Krebs cycle -Cyclic photophosphorylation -Light reaction -Noncyclic photophosphorylation	Carbon dioxide release	-Carbon dioxide fixation -Krebs cycle -Carbon dioxide reduction -RuBP regeneration	-Carbon dioxide fixation -Carbon dioxide release -Carbon dioxide reduction -RuBP regeneration
2	The light independent reactions of photosynthesis take place in the _____	-stroma	-Thylakoid -Chloroplast membrane -Ribosome -Thylakoid space -DNA	-Stroma of the mitochondrion -Thylakoids of the mitochondrion -Cristae of the mitochondrion -Thylakoid membrane -Cytoplasm surrounding the chloroplast -Chlorophyll molecule -Chloroplast membrane -Grana -Thylakoids -Nucleus -Mitochondria	-Stroma -Thylakoid space -Chloroplast membrane -Ribosome -DNA	-Stroma -Thylakoids -Nucleus -Mitochondria -Grana
3	Photosynthesis takes place in which organelle?	-chloroplast	-Plasma membrane -Mitochondria -Tonoplast -Endoplasmic reticulum -Golgi apparatus -Cell wall -Peroxisome -Chromosome -Centrosome	-Ribosome -Endoplasmic reticulum -Vacuole -Mitochondria -Nucleus -Lysosomes -The golgi apparatus -Carbon dioxide -Nitrogen	-Plasma membrane -Tonoplast -Endoplasmic reticulum -Chloroplast -Peroxisome	-Mitochondria -Ribosome -The golgi apparatus -Chloroplast -Nucleus
4	Plants produce what two products in photosynthesis?	oxygen and glucose	-NADP+ -Sunlight -Water	-Carbon dioxide -Nitrogen	-Water and oxygen -Oxygen and glucose	-Carbon dioxide and oxygen -Oxygen and glucose

(*Continued*)

Continued.

	Question stem	Correct answer	Machine raw distractors	Human raw distractors	Machine alternatives	Human alternatives
			-ADP -ATP -Carbon dioxide -Glyceraldehyde 3 phosphate -NADPH		-Glucose and sunlight -NADPH and glucose	-Glucose and carbon dioxide -Nitrogen and glucose
5	The final product of the Calvin cycle is _____	G3P	-NADPH -RuPB -3 phosphoglycerate -ATP -Carbon dioxide	-RuBP -PGA -ATP	-NADPH -Carbon dioxide -3 phosphoglycerate -G3P	-RuBP -PGA -ATP -G3P
6	What two energy carrying molecules are produced by the light of reaction of photosynthesis?	NADPH and ATP	-Sunlight -water	-NADP+ -ADP	-Sunlight and water -Water and ATP -NADPH and ATP -Sunlight and ATP -Water and NADPH	-NADP+ and ADP -NADP+ and ATP -NADPH and ATP -ADP and ATP -NADP+ and NADPH
7	Photosynthesis takes what 3 thing to create energy?	Carbon dioxide, water and sunlight	-ATP -NADPH -RuBP -Sugar -oxygen	-Carbon monoxide -Cytoplasm	-Oxygen, water and sunlight -Carbon dioxide, water and sunlight -Carbon dioxide, sugar and sunlight -Oxygen, sugar and sunlight	-Carbon monoxide, water and sunlight -Carbon dioxide, water and sunlight -Carbon dioxide, cytoplasm and sunlight -Carbon monoxide, cytoplasm and sunlight
8	Which of the following molecules is produced during the noncyclic electron pathway?	-ATP -NADPH -Oxygen	-Sunlight -water	-glucose	-ATP -NADPH -Sunlight -Oxygen	-ATP -NADPH -Glucose -Oxygen
9	Which of the following are products of the light reactions of photosynthesis that are utilized in the Calvin cycle?	ATP and NADPH	-Carbon dioxide -RuBP -Oxygen	-Oxygen -Carbon dioxide -NADH -Light energy -Water -Sugar -ADP -NADP+ -Electrons -RuBP	-Carbon dioxide and ATP -RuBP and NADPH -RuBP and oxygen -oxygen and ATP -ATP and NADPH	- Carbon dioxide and ATP -Water and NADPH -Sugar and oxygen -Light energy -ATP and NADPH