Investigating the Quality of AI-Generated Distractors for a Multiple-Choice Vocabulary Test

Wojciech Malec[a]

Institute of Linguistics, John Paul II Catholic University of Lublin, Al. Racławickie, Lublin, Poland

Keywords: AI-Generated Items, ChatGPT, Vocabulary Assessment, Multiple-Choice Testing, Distractor Analysis.

Abstract: This paper reports the findings of a study into the effectiveness of distractors generated for a multiple-choice vocabulary test. The distractors were created by OpenAI's ChatGPT (the free version 3.5) and used for the construction of a vocabulary test administered to 142 students learning English as a foreign language at the advanced level. Quantitative analysis revealed that the test had relatively low reliability, and some of its items had very ineffective distractors. When examined qualitatively, certain items were likewise found to have an ill-matched set of options. Moreover, follow-up queries failed to correct the original errors and produce more appropriate distractors. The results of this study indicate that although the use of artificial intelligence has an unquestionably positive impact on test practicality, ChatGPT-generated multiple-choice items cannot yet be used in operational settings without human moderation.

1 INTRODUCTION

The recent advances in Artificial Intelligence (AI) have opened up new possibilities for education, resulting in the provision of powerful technologies to support both students and teachers (Holmes, Bialik, & Fadel, 2019). Specific examples of AI applications include personalization of education through intelligent tutoring systems (Arslan, Yildirim, Bisen, & Yildirim, 2021), development of practice quizzes for test preparation (Sullivan, Kelly, & McLaughlan, 2023), automated essay scoring (Gardner, O'Leary, & Yuan, 2021), chatting robots and intelligent virtual reality (Pokrivcakova, 2019), and many others. The latest technological innovations have also affected language learning and teaching, where AI-powered tools can offer assistance in designing learning materials, creating classroom activities, developing assessment tasks (including narrative writing prompts), levelling texts for reading practice, and providing personalized feedback (Bonner, Lege, & Frazier, 2023).

One of the most widely recognized recent implementations of artificial intelligence is OpenAI's Generative Pretrained Transformer (GPT), known as ChatGPT. As a large language model (LLM), ChatGPT is capable of performing natural language processing (NLP) tasks, such as text interpretation and generation. For example, it has been successfully used to generate stories for reading comprehension, and these were found to be very similar to human-authored passages in terms of coherence, appropriateness, and readability (Bezirhan & von Davier, 2023). The GPT-3 model has also been reported in Attali et al. (2022) as a text generator for reading comprehension assessments.

Specifically in testing and assessment, ChatGPT has been utilized for automatic item generation (AIG), thanks to the fact that it "can generate novel and diverse items by leveraging its ability to understand and generate natural language" (Franco & de Francisco Carvalho, 2023, p. 6). Generally speaking, AIG can rely either on structured inputs (template-based methods) or on unstructured inputs (non-template-based methods) (Circi, Hicks, & Sikali, 2023). In the latter case, AIG draws on NLP techniques, which means that the application of ChatGPT to item writing can be acknowledged as being a type of AIG (Kıyak, Coşkun, Budakoğlu, & Uluoğlu, 2024).

ChatGPT has been used to generate items for content-area assessments (e.g. Kumar, Nayak, Shenoy K, Goyal, & Chaitanya, 2023; Kıyak et al.,

[a] https://orcid.org/0000-0002-6944-8044

Malec, W.
Investigating the Quality of AI-Generated Distractors for a Multiple-Choice Vocabulary Test.
DOI: 10.5220/0012762400003693

In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024) - Volume 1*, pages 836-843
ISBN: 978-989-758-697-2; ISSN: 2184-5026

2024), reading comprehension (Attali et al., 2022; Sayin & Gierl, 2024), as well as assessments of specific components of language ability, such as vocabulary (Attali et al., 2022). The item format that has been predominantly, though not exclusively, used in these assessments is multiple choice.

However, challenges still exist in ensuring the appropriateness of AI-generated items. More precisely, while examination of the psychometric properties of test items is mandatory for operational purposes, AI-generated items reported in the literature have not always undergone the necessary evaluation, as pointed out by Circi et al. (2023, p. 4). This paper thus contributes to the debate on the quality of AI-generated test items, in the context of vocabulary assessment, by reporting the psychometric characteristics of multiple-choice items administered to a group of students learning English as a foreign language.

2 MULTIPE-CHOICE TESTING

Despite many of its shortcomings (such as lack of authenticity and susceptibility to guessing), multiple choice (MC) is still arguably the most popular item format in educational assessment (Parkes & Zimmaro, 2016), including language testing. MC items can be found in large standardized tests (e.g. the Test of English as a Foreign Language, TOEFL; the Test of English for International Communication, TOEIC; the French-language proficiency test, TFI), well-known vocabulary measures (e.g. the Vocabulary Size Test of Nation & Beglar, 2007), as well as informal practice quizzes for language learners, such as those created using Quizlet (quizlet.com) or Quizizz (quizizz.com).

The greatest advantage of MC items is that they are easy to score and administer. Moreover, given that there can be only one correct answer for an item, problems with inter- or intra-rater reliability are avoided (Hoshino, 2013), which is not the case with constructed-response items, where raters are likely to disagree on the acceptability of the responses. On the other hand, good quality MC items are difficult to develop (see below).

Although various MC formats have been used to test vocabulary, there is a general consensus that testing words in context contributes to the authenticity of assessment, and the required context may actually be limited to a single sentence (Read, 2000). This means that a multiple-choice vocabulary item can be constructed by presenting the target word in a context sentence (the stem), replacing this word with a gap, and then providing the word as one of the options (the key). In addition to that, some other words are given as incorrect options (called distractors).

It is worth adding that the optimal number of MC options is three (e.g. Bruno & Dirkzwager, 1995; Rodriguez, 2005) and constructing extra distractors is not an appropriate way of increasing the quality of MC items (Papenberg & Musch, 2017). The reason for this is that even in well-developed tests additional distractors do not usually function as expected, i.e. they may be (1) rarely selected, (2) non-discriminating, and/or (3) more attractive to high scorers than to low scorers (Haladyna & Rodriguez, 2013). It must be admitted, however, that in language testing four-option MC items are still very popular.

2.1 Distractor Development

Some of the MC item-writing guidelines available in the literature (e.g. Haladyna & Downing, 1989; Haladyna, Downing, & Rodriguez, 2002) refer to the distractors, which must be plausible but plainly incorrect. In fact, the writing of plausible (but indisputably wrong) distractors constitutes the greatest challenge associated with MC item development. One way of obtaining distractors for MC items is through analysing common wrong answers from gap-filling items (e.g. March, Perrett, & Hubbard, 2021). The problem with this approach is that, when assessing new material in classroom contexts, the teacher may not always be able to afford the time to test the same content domain twice.

The item-writing guidelines which are particularly relevant to the key and distractors in assessments of vocabulary can be summarized as follows (cf. Fulcher, 2010): (1) all the options should belong to the same word class, (2) the distractors and the key should have similar grammatical properties, and (3) the distractors and the key should be similar semantically.

The third guideline may not always be advisable. For example, when constructing ‘closest-in-meaning’ vocabulary questions, as used in TOEFL, it may be useful to avoid distractors which are semantically similar to the correct answer (Susanti, Tokunaga, Nishikawa, & Obari, 2018). On the other hand, some degree of semantic relatedness may well be an essential feature of distractors in vocabulary depth tests employing synonym tasks (Ludewig, Schwerter, & McElvany, 2023).

When words are tested in sentential context, semantic relatedness between the options may or may not be a desired selection criterion. By way of illustration, the options in (1) are all semantically related (the key is given in square brackets):

(1) It [*says*] in the paper that they haven't identified the robber yet.

Distractor 1: *speaks*
Distractor 2: *tells*

Despite the obvious semantic similarity, both distractors seem to be acceptable because none of them can be used to correctly complete the context sentence. By contrast, this is not the case with one of the distractors in (2) below:

(2) It's the kind of tune that sticks in your [*mind*].

Distractor 1: *brain*
Distractor 2: *head*

In this example, the second distractor, even though arguably not as appropriate as the key, is not completely ruled out since the collocation *stick in your head* can be found in English language corpora, such as the British National Corpus (BNC).

In the study reported below, semantic similarity was one of the criteria for distractor selection. However, this was not a necessary condition as distractors were supposed to be highly plausible, yet plainly wrong in the contexts given.

2.2 MC Item Analysis

MC item analysis essentially boils down to inspecting the performance of the options. Various methods of detecting badly-performing options can be found in the literature (Malec & Krzemińska-Adamek, 2020). The analysis typically begins by examining the frequencies (or proportions) of option choices. Specifically, if any of the options is selected by fewer than 5% of the test takers, this option is unlikely to be useful (Haladyna & Downing, 1993). Options which satisfy the frequency criterion can be further analyzed with a view to finding out whether they are more attractive to high scorers or to low scorers. Generally speaking, the correct answer should attract high scorers, whereas the distractors should attract low scorers (e.g. Bachman, 2004).

One way of determining whether the key and distractors perform appropriately is by inspecting the trace lines (Haladyna, 2016; Gierl, Bulut, Guo, & Zhang, 2017). A trace line indicates the number or percentage of test takers in several score groups who selected a given option. As a rule, correct answers should have ascending trace lines, whereas distractors should have descending trace lines.

In addition, the point-biserial correlation can be used as an indicator of item and distractor discrimination. Broadly speaking, the point-biserial calculated for the correct answer should be positive. When calculated for a distractor, by contrast, the correlation is supposed to be negative. A correlation that is close to zero means that the option in question does not discriminate.

3 STUDY

The purpose of this study was to assess the quality of MC vocabulary items created by artificial intelligence. In particular, the investigation was intended to ascertain whether AI tools can be trusted as sources of MC questions for classroom-based vocabulary assessment.

3.1 Method

The assessment instrument developed for this study included 15 MC vocabulary items. First, the target lexical items were selected from a coursebook for advanced learners of English (Clare & Wilson, 2012). Next, an AI-powered online platform, Twee (Twee, 2024), was employed to write context sentences for the target words. The reason why Twee was chosen instead of ChatGPT is that the system offers a user-friendly tool designed specifically for this task. In the next step, ChatGPT (the free version 3.5, OpenAI, 2024) was instructed to suggest distractors for the target words (as used in the context sentences). Table 1 shows the instructions included in the input for ChatGPT. The wording of the input was based on the assumption that ChatGPT is capable of understanding natural language and respond to it in a human-like manner. The output from ChatGPT included the (same) context sentences followed by the options (key and distractors).

Following this, an online test was constructed on the WebClass platform (webclass.co) using the AI-created items (see the entire test in the Appendix). Next, the test was administered to a group of advanced learners of English as a foreign language during their regular lessons at school. The test takers (n = 142) included 63 males and 79 females, aged between 16 and 18. Finally, the test and individual items were analyzed using the statistics and trace lines generated on the testing platform.

Table 1: Instructions for ChatGPT.

You are a teacher of English who is writing a vocabulary test at the advanced level.

The test is composed of multiple-choice items. Each item consists of a context sentence and a gap for which three options are provided, i.e. the key (correct answer) and two distractors (incorrect answers).

The sentences are given below:

1. The court was in [*session*] when the judge entered the courtroom.
2. (…)

Each sentence contains one word in square brackets. This is the key. For each sentence add TWO distractors (INCORRECT answers), using the following guidelines:

- the distractors must belong to the same word class as the key (noun, verb, adjective, adverb, preposition);
- the distractors and the key should have similar grammatical properties (such as tense for verbs and number for nouns);
- if possible, the distractors and the key should be similar semantically (they should belong to the same semantic field);
- the distractors should be plausible (likely to be selected by the students taking the test) but evidently wrong, which means that when the key is replaced with a distractor, the sentence becomes plainly incorrect.

3.2 Results and Discussion

The statistics calculated for the test as a whole are given in Table 2.

Table 2: Test statistics.

Number of test takers (*n*)	142
Number of items (*k*)	15
Cut score [test] (λ_t)	7.5
Cut score [item] (λ)	0.5
Mean ($\bar{x}$)	7.53
Mean of proportion scores ($\bar{x}_p$)	0.50
Standard deviation (*SD*)	2.23
Cronbach's alpha (α)	.488
Standard error of measurement (*SEM*)	1.60
Phi coefficient (Φ)	.398
Phi lambda (Φ_λ)	.258
Kappa squared (κ^2)	.488
Standard error for absolute decisions (SEM_{abs})	1.91

The results of test analysis show that the assessment instrument was only moderately reliable for norm-referenced interpretations (as indicated by Cronbach's alpha) and definitely below acceptable levels of reliability (or dependability) for criterion-referenced interpretations (as indicated by the phi coefficient). Moreover, if used for classification purposes (pass/fail), the decisions would tend to be incorrect (cf. the low values of phi lambda and kappa squared).

Table 3: Item analysis (the correct options are asterisked).

Question	Option	Percentage of selection	*PB*
1	A*	40.1	.23
	B	22.5	−.32
	C	37.3	−.16
2	A	7.7	−.31
	B*	73.2	.47
	C	19.0	−.46
3	A*	2.8	.15
	B	55.6	−.16
	C	41.5	−.28
4	A*	83.8	.56
	B	2.1	−.34
	C	14.1	−.53
5	A	11.3	−.29
	B	37.3	−.43
	C*	51.4	.41
6	A*	68.3	.55
	B	15.5	−.48
	C	16.2	−.50
7	A	16.2	−.56
	B	29.6	−.27
	C*	53.5	.39
8	A*	82.4	.48
	B	7.7	−.30
	C	9.2	−.49
9	A*	26.8	.07
	B	58.5	−.13
	C	14.8	.18
10	A	35.2	−.46
	B*	53.5	.46
	C	11.3	−.38
11	A	41.5	.19
	B	57.0	.12
	C*	1.4	−.11
12	A	23.2	−.30
	B*	20.4	.19
	C	56.3	−.19
13	A	7.0	−.28
	B	23.9	−.50
	C*	69.0	.49
14	A*	44.4	.32
	B	48.6	−.32
	C	7.0	−.22
15	A	12.7	−.35
	B	4.9	−.11
	C*	81.7	.33

The results of item analysis are presented in Table 3. The point-biserial correlation (*PB*) for the distractors was calculated using a modified formula (as proposed by Attali & Fraenkel, 2000), which makes a comparison between distractor and correct choices.

As can be seen in Table 3, the test was composed of a blend of satisfactory and mediocre items. In particular, Item 3 was extremely difficult – the key was selected by fewer than 5% of the test takers, which is unusual. This same problem was even more significant in the case of Item 11, where both distractors had positive values of the point-biserial (indicating that these options performed as should the correct answer). A rudimentary qualitative analysis of the options (cf. Appendix) points to the conclusion that ChatGPT failed to produce lexically appropriate distractors for the items in question. By contrast, some of the items (e.g. 2, 4, 5, 6, 8, 10, 13) were quite good as all of their options had satisfactory discrimination.

In addition, item analysis involved examining the trace lines. Figures 1–4 show selected trace lines to illustrate effective and ineffective options. Specifically, Option 10B (Figure 1) had a correctly ascending trace line, typical of a well-performing key, and Option 15A (Figure 2) had a descending trace line, as expected of an effective distractor.

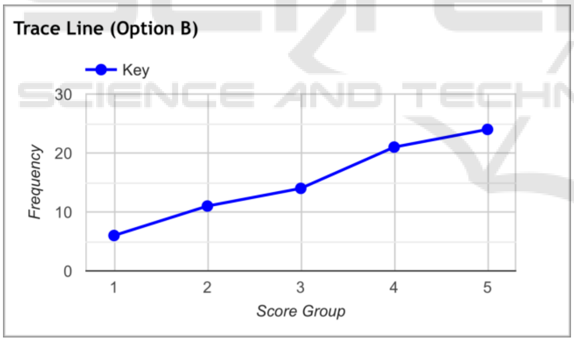

Figure 1: An effective key (10B).

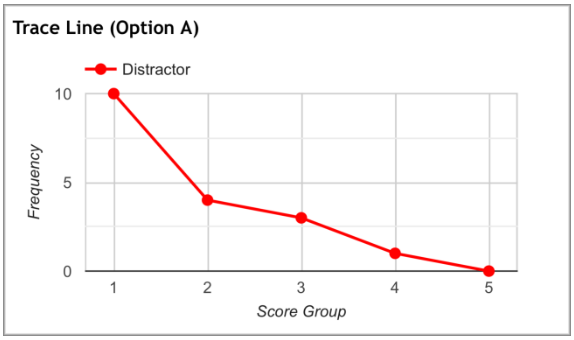

Figure 2: An effective distractor (15A).

In contrast, Option 9A (Figure 3) was rather inconsistent, and Option 3B (Figure 4) was entirely unacceptable as it was even more unstable in the way it attracted low and high scorers.

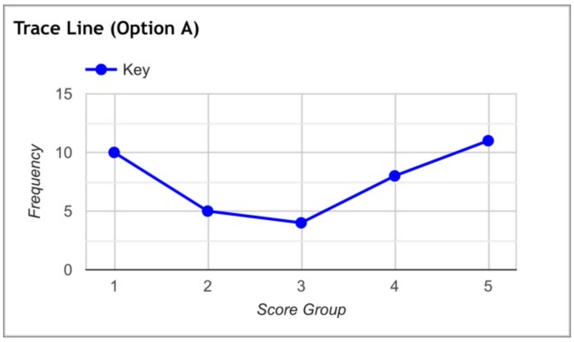

Figure 3: An inconsistent key (9A).

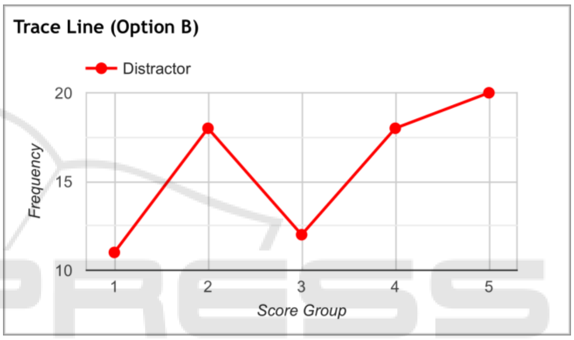

Figure 4: An inadequate distractor (3B).

3.2.1 Follow-Up Queries

When asked again to (re)generate distractors for a selection of the most unsatisfactory items, ChatGPT produced inconsistent results. For example, its output for Item 3 included the following distractors (key in brackets):

(3) There are [*sound*] reasons for implementing new safety measures at the factory.

Distractor 1: *solid*
Distractor 2: *melody*

The response shows that, in this particular case, ChatGPT failed to follow the instructions, according to which all the options should belong to the same word class. Even if we acknowledge that *sound* and *melody* are both nouns, the distractor (*melody*) is highly implausible because the average test taker expects an adjective in this particular context. Furthermore, the word *solid* did not seem to be a good distractor (i.e. an indisputably wrong option). In order to determine whether the problem with this distractor was an example of mere oversight, ChatGPT was asked to confirm that the word *solid* would indeed be an error in the given context (see Figure 5).

You
Which word is correct in the following sentence?
There are __ reasons for implementing new safety measures at the factory.
A) solid
B) melody
C) sound

ChatGPT
The correct word in the sentence is:
C) sound

You
Is "solid" incorrect?

ChatGPT
Yes, in the given sentence, "solid" would be considered incorrect. The most appropriate choice is "sound" in the context of reasons for implementing new safety measures at the factory. "Solid" doesn't convey the intended meaning in this context.

Figure 5: Part of a conversation with ChatGPT.

As can be seen in Figure 5, ChatGPT maintained that the word *solid* does not collocate correctly with *reasons*. However, according to the British National Corpus, BNC*web* (CQP-edition, Hoffmann & Evert, 2006), this is an acceptable collocation, although admittedly not very common (see Figure 6).

Hits 1 to 7 Page 1 / 1		
sist on denying him. Besides, there is no	solid reason	to assume that Mr Hussein's will is goir
the poll-tax pantomime. There are more	solid reasons	, though. The Labour Party is paying the
n has to paint a picture — it must give a	solid reason	why the report has to be considered, so
of going anywhere but it's nice to have a	solid reason	to stay put.' Talbot thought briefly. 'Sol
XPERIENCE Quine and Kuhn give very	solid reasons	for agreeing that scientific theories cont
n research field but for three rather more	solid reasons	. First, the episode shows how easy it is
don't tell you anything. It has to be some	solid reason	. Some absolute measurable thing. Okay

Figure 6: Concordance lines for *solid reason(s)* in BNC*web* (CQP-edition).

A similar problem remained unresolved with Item 13, for which ChatGPT's second attempt resulted in the following distractors:

(4) He [*disobeyed*] the rules and faced consequences for his actions.

Distractor 1: *followed*
Distractor 2: *ignored*

In fact, the outcome of the second attempt was even less satisfactory than of the first one (cf. Appendix) as neither of the distractors in (4) seems to be plainly wrong in the context given.

3.3 Limitations and Way Forward

Several limitations of this study are in order. First, it bears pointing out that GPT-3.5 is not the most powerful version of OpenAI's LLM, yet it is freely available without a fee, in contrast to the more recent (but paid) GPT-4. It is perfectly possible that more advanced models of AI should be capable of generating better distractors for MC items. Second, this study focussed exclusively on an AI-generated test, without comparing it to a human-made test. In fact, another study is currently in preparation, where teacher-created distractors are compared to those obtained from ChatGPT. Another possibility would be to compare ChatGPT to a different AI tool. Finally, this study does not offer any suggestions for prompt enhancements. The reason for this is that, by and large, the follow-up queries did not result in any improvement to the distractors originally suggested by ChatGPT. But such enhancements are definitely worth further investigation.

4 CONCLUSIONS

A fundamental consideration in the construction of MC tests is the quality of distractors, which have a direct influence on the psychometric properties of MC items. The choice of distractors plays a critical role in evaluating the test-takers' true comprehension of a word, making the problem particularly relevant in the context of vocabulary assessment.

A relatively novel way of generating distractors for MC items is by using artificial intelligence. This paper has demonstrated that AI, represented by ChatGPT (the free version 3.5), has the potential to contribute to assessment practicality, reducing the time and effort required for test development, but this method of creating distractors is still far from ideal. The results of the study reported here indicate that AI-generated MC items are quite likely to be in need of human moderation. Accordingly, the conclusion has to be that, as observed by Khademi (2023), AI tools such as ChatGPT "can only be trusted with some human supervision" (p. 79). In view of that, at least for the time being, AIG utilizing ChatGPT might better be termed *assisted* item generation (Segall, 2023).

ACKNOWLEDGEMENTS

I am grateful to the teachers of XXI Liceum Ogólnokształcące im. św. Stanisława Kostki in Lublin, Poland, who administered the vocabulary test and to the students who completed it.

REFERENCES

Arslan, E. A., Yildirim, K., Bisen, I., & Yildirim, Y. (2021). Reimagining education with artificial intelligence. *Eurasian Journal of Higher Education, 2*(4), 32–46.

Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement, 37*(1), 77–86.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, 903077. doi:10.3389/frai.2022.903077

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5*, 100161. doi:https://doi.org/10.1016/j.caeai.2023.100161

Bonner, E., Lege, R., & Frazier, E. (2023). Large Language Model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology, 23*(1), 23–41.

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*, 959–966.

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*, 858273. doi:10.3389/feduc.2023.858273

Clare, A., & Wilson, J. (2012). *Speakout Advanced: Students' Book*. Harlow: Pearson.

Franco, V. R., & de Francisco Carvalho, L. (2023). A tutorial on how to use ChatGPT to generate items following a binary tree structure. *PsyArXiv Preprints*. doi:https://doi.org/10.31234/osf.io/5hnkz

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning, 37*(5), 1207-1216. doi:https://doi.org/10.1111/jcal.12577

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116.

Haladyna, T. M. (2016). Item analysis for selected-response items. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (2nd ed., pp. 392–407). New York, NY: Routledge.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 37–50.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*, 999–1010.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309–334.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York, NY: Routledge.

Hoffmann, S., & Evert, S. (2006). BNC*web* (CQP-edition): The marriage of two corpus tools. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods* (pp. 177–195). Frankfurt am Main: Peter Lang.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign.

Hoshino, Y. (2013). Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. *Language Testing in Asia, 3*(1), 16. doi:10.1186/2229-0443-3-16

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching, 6*(1), 75–80.

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*. doi:10.1007/s00228-024-03649-x

Kumar, A. P., Nayak, A., Shenoy K, M., Goyal, S., & Chaitanya. (2023). A novel approach to generate distractors for Multiple Choice Questions. *Expert Systems with Applications, 225*, 120022. doi:https://doi.org/10.1016/j.eswa.2023.120022

Ludewig, U., Schwerter, J., & McElvany, N. (2023). The features of plausible but incorrect options: Distractor plausibility in synonym-based vocabulary tests. *Journal of Psychoeducational Assessment, 41*(7), 711–731. doi:10.1177/07342829231167892

Malec, W., & Krzemińska-Adamek, M. (2020). A practical comparison of selected methods of evaluating multiple-choice options through classical item analysis. *Practical Assessment, Research, and Evaluation, 25*(1, Art. 7), 1–14.

March, D. M., Perrett, D., & Hubbard, C. (2021). An evidence-based approach to distractor generation in multiple-choice language tests. *Journal of Higher Education Theory and Practice, 21*(10), 236–253. doi:10.33423/jhetp.v21i10.4638

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

OpenAI. (2024). ChatGPT (Feb 15 version) [Large language model]. https://chat.openai.com.

Papenberg, M., & Musch, J. (2017). Of small beauties and large beasts: The quality of distractors on multiple-choice tests is more important than their quantity. *Applied Measurement in Education, 30*(4), 273–286.

Parkes, J., & Zimmaro, D. (2016). *Learning and Assessing with Multiple-Choice Questions in College Classrooms*. New York, NY: Routledge.

Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education, 7*(3), 135–153.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice, 43*(1), 5–18. doi:https://doi.org/10.1111/emip.12590

Segall, D. (2023). Innovating test development: ChatGPT's promising role in assisted item generation. Retrieved from https://www.linkedin.com/pulse/innovating-test-development-chatgpts-promising-role-assisted-segall

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching, 6*(1), 31–40.

Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2018). Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning, 13*(1), 15. doi:10.1186/s41039-018-0082-z

Twee. (2024). AI Powered Tools For English Teachers (Feb 15 version). https://twee.com.

APPENDIX

Vocabulary Test

Multiple Choice

Select the best option.

(1) The court was in ___ when the judge entered the courtroom.
A. session B. break C. recess

(2) He's taking a lot of ___ for his chronic back pain.
A. prescription B. medication C. treatment

(3) There are ___ reasons for implementing new safety measures at the factory.
A. sound B. valid C. logical

(4) During questioning, he ___ everything the police accused him of.
A. denied B. accepted C. rejected

(5) The accused couldn't afford a private lawyer, so he was assigned a public ___ .
A. prosecutor B. advocate C. defender

(6) The film is ___ in a futuristic world where technology dominates society.
A. set B. located C. placed

(7) The ___ for the defense presented strong arguments in the courtroom.
A. advice B. guidance C. counsel

(8) The prosecutor will call the first ___ to testify against the accused.
A. witness B. observer C. participant

(9) The parade began with the traditional ___ music played by the band.
A. martial B. military C. warlike

(10) The company's ___ include properties, equipment, and cash reserves.
A. debts B. assets C. liabilities

(11) She was ___ on him to help her succeed in the business venture.
A. relying B. counting C. banking

(12) The attorney presented a compelling ___ for the prosecution during the trial.
A. presentation B. case C. argument

(13) He ___ the rules and faced consequences for his actions.
A. adhered B. followed C. disobeyed

(14) After the breakup, she ___ a page in her life and focused on self-improvement.
A. turned B. closed C. opened

(15) There is no ___ to suggest that he was involved in the robbery.
A. confirmation B. indication C. evidence