Using ChatGPT to Write Examination Questions: Opportunities and Pitfalls

Henk G. Schmidt
Professor of Psychology and Medical Education, Department of Psychology, Erasmus University Rotterdam, the Netherlands, henkschmidt@icloud.com

Mohi M. A. Magzoub
Professor of Medical Education, Department of Medical Education, College of Medical and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates

Follow this and additional works at: https://hpe.researchcommons.org/journal

Recommended Citation
Schmidt, Henk G. and Magzoub, Mohi M. A. () "Using ChatGPT to Write Examination Questions: Opportunities and Pitfalls," Health Professions Education: Vol. 10: Iss. 1, Article 1.
Available at: https://hpe.researchcommons.org/journal/vol10/iss1/1

This Editorials is brought to you for free and open access by Health Professions Education. It has been accepted for inclusion in Health Professions Education by an authorized editor of Health Professions Education.
Using ChatGPT to Write Examination Questions: Opportunities and Pitfalls

Henk G. Schmidt a,*, Mohi M.A. Magzoub b

a Department of Psychology, Erasmus University Rotterdam, the Netherlands
b Department of Medical Education, College of Medical and Health Sciences, United Arab Emirates University, Al Ain, United Arab Emirates

Keyword: ChatGPT, Multiple-choice items, Short-answer questions

1. Introduction

Writing multiple-choice items or short-answer questions for assessment purposes is an important but time-consuming task for a teacher. Not only should the questions fairly represent the subject matter tested, they also should be simple in format, challenge students but allow them to succeed often, and target cognitive processes that correspond to particular learning objectives. Finally, because assessment drives learning, they should enable meaningful feedback to students about their performance [1]. Apart from the usual formatting guidelines (no more than three alternative answers, no complex answer structures), it is difficult to a priori assess the extent to which a question fulfills these requirements. Only post hoc, often statistical, analyses allow for answers in this respect. Further, a test should have sufficient reliability, which is directly dependent on the number of questions included. However, teachers have only limited time and creativity to produce large numbers of such questions. The latter is problematic because students tend to copy questions for use by the next generations of students. This is surely a sign of solidarity, but forces teachers to eternally produce new questions.

Could artificial intelligence help solving this problem? A few early studies suggest that it can, but for now with mixed results [2,3]. Questions produced in these early studies were derived from large text bases such as Harrison’s Principles of Internal Medicine. In addition, these studies limited themselves to the writing of multiple-choice items. Would ChatGPT 3.5 (https://chat.openai.com) one of the large-language models using artificial intelligence techniques, be able to write items based on a more limited text such as a book chapter? And would it also be able to produce open-ended short-essay questions? Would it be able to provide the correct answers to these questions? And finally, are these questions any good?

2. Methods and results

To test the capabilities of ChatGPT in this respect, we copied a text on bacterial endocarditis, written by Vilcant and Hai, from the website of the National Library of Medicine. This 1112-word text, intended to provide continuing education to physicians, describes etiology, epidemiology, pathophysiology, and history and physical examination related findings. In addition, it supports the learner with inclusion-exclusion clinical criteria for infective endocarditis. The full text can be found at: https://www.ncbi.nlm.nih.gov/books/NBK470547/#:~:text=It%20usually%20involves%20heart%20valves,with%20in%20weeks%20if%20not%20treated.

We first submitted to ChatGPT the following instruction: “We need to write 30 multiple-choice items for an examination for medical students. The items should be the true-false type statements and should be based on the following text:” (here we copied the endocarditis text to the Chat box). Some examples that ChatGPT returned are presented in Table 1. Responses of ChatGPT are all in Italic print.
Finally, we asked ourselves (and ChatGPT) if it could also produce short-answer questions with their answers. The answer is again affirmative. See for an example Table 3.

### 3. Discussion

The results of these attempts to have ChatGPT help in the production of examination questions give rise to the following remarks. First, comparing the full set of items with the bacterial endocarditis text reveals that it faithfully covered the important elements. Second, most items tend to be simple reformulations of textual propositions. For instance, the textual proposition: “Most cases are caused by (...) Streptococcus galloyticus (…)” is transformed into “S. galloyticus is a common cause of bacterial endocarditis.” Second, items can sometimes already be answered correctly without relevant background knowledge, e.g., “Acute endocarditis has a slower disease process as compared with subacute endocarditis.” Third, ChatGPT can make mistakes. Inspection of the first item of Table 2 reveals that all alternatives are in fact correct. Fourth, answers provided to open-ended questions tend to be literal quotations from the text rather than constructive summaries. A fifth pitfall is that large-language models using artificial intelligence techniques are known to be “hallucinating” when they are not able to comply with requests [4].

In conclusion, ChatGPT may be useful to increase the pool of examination questions considerably and in a short time without requiring much time and energy. However, each of the questions produced subsequently requires careful supervision regarding relevance and truth to avoid the pitfalls described here.

### Conflicts of interest

There is no conflict of interest with regard to this contribution.

### References


