JOURNAL OF ATHLETIC TRAINING EDUCATION AND PRACTICE

www.natajournals.org
ISSN: 3067-1833
DOI: 10.4085/1947-380X-24-066

ORIGINAL RESEARCH

Comparing Human- and ChatGPT-Generated Multiple-Choice Questions in Athletic Training Education

Christina Davlin-Pater, PhD, ATC; Lisa S. Jutte, PhD, ATC
Sport Science & Management, Xavier University, Cincinnati, OH

Context: Creating well-written multiple-choice questions (MCQs) requires time and attention to detail. Artificial intelligence tools such as ChatGPT have the potential to assist faculty members in creating exam or practice questions.

Objective: To compare human-generated athletic training–related MCQs with those generated by ChatGPT for quality, clarity, relevance, and difficulty of the questions.

Design: Cross-sectional study.

Patients or Other Participants: Ninety-three athletic training faculty teaching in Commission on Accreditation of Athletic Training Education–accredited entry-level athletic training programs completed the survey. Eleven second-year graduate-level athletic training students completed the 20-question quiz.

Main Outcome Measure(s): Faculty participants completed a 2-part survey in which they evaluated 10 pairs of MCQs for grammar, clarity, difficulty, terminology, and suitability using a 5-point Likert scale, and indicated which question they preferred. Each pair included a human-generated question and a ChatGPT-generated question on a similar topic. A student quiz was developed to evaluate question quality/difficulty. Second-year master's students nearing graduation were asked to complete the 20-question quiz using the same questions found in the faculty survey.

Results: ChatGPT-generated Board of Certification–style questions used in this study have similar values for grammar, stem quality, answer quality, question difficulty, proper use of medical terminology, and suitability for content to human-generated questions for all 5 athletic training domains. Most ChatGPT-generated questions were easy to understand, used appropriate terminology, and had answer options that were similar in style and length.

Conclusions: ChatGPT is another tool that athletic training faculty may consider using to improve the quality and efficacy of exam question preparation. The data from this study suggest that faculty can effectively use ChatGPT for exam question preparation; however, faculty should understand that ChatGPT, like all tools, has its limitations.

Key Words: Question writing, artificial intelligence, exam development, Board of Certification question style

Dr Davlin-Pater is currently a professor in Sport Science & Management at Xavier University. Address correspondence to Christina Davlin-Pater, PhD, ATC, Xavier University, 3800 Victory Pkwy, Cincinnati, OH 45207-6311. davlin@xavier.edu.

Full Citation:
Davlin-Pater C, Jutte LS. Comparing human- and ChatGPT-generated multiple-choice questions in athletic training education. *J Athl Train Educ Pract*. 2025;21(2):51–60.

Comparing Human- and ChatGPT-Generated Multiple-Choice Questions in Athletic Training Education

Christina Davlin-Pater, PhD, ATC; Lisa S. Jutte, PhD, ATC

KEY POINTS

- ChatGPT-generated questions are similar to human-generated questions in terms of grammar, stem quality, answer quality, question difficulty, proper use of medical terminology, and suitability for content.
- Most ChatGPT-generated questions were easy to understand, used appropriate terminology, and included answer options that were similar in style and length.
- ChatGPT can be used by athletic training faculty to generate multiple-choice questions, but questions and answers should be carefully reviewed and refined.

INTRODUCTION

Faculty in athletic training need to regularly and accurately assess students to facilitate learning, demonstrate student mastery of athletic training concepts, and provide evidence of compliance with accreditation standards. Instructors commonly use multiple-choice questions (MCQs) in both low-stakes quizzes and high-stakes exams. Well-written questions can produce meaningful test scores and valid measurements of student learning.[1,2] However, writing quality MCQs can be difficult and time-consuming.[3–5] To assess crucial content, questions must be well structured, easy to understand, and free of construction errors.[6] Terminology should be accurate and precise to reduce the chance of confusion or misinterpretation.[7] To decrease the likelihood of guessing the correct answer, each incorrect answer option (distractor) must be similar to the correct answer in terms of style and length while also being plausible to students who have not yet mastered the material and also clearly incorrect to students who have learned the content.[2,8] Quality questions are essential for exams to be fair and for scores to be interpreted correctly.[9,10]

There are tools available to educators to evaluate MCQs after an exam to help assess question quality and identify problematic items. For example, the corrected item-total correlation coefficient examines how each MCQ is related to overall test performance.[11] Values range from −1.0 to 1.0. A positive value indicates that students who score higher on the exam are more likely to answer the item correctly. This suggests that the question is relevant and aligns with the goals of the exam. Test questions with a value of 0.25 or above indicate that the question has good distractors and provides good discrimination.[12] Negative items indicate that a question is miskeyed or ambiguous and confusing for students. Exam questions with a negative corrected item-total correlation should be revised or eliminated.[12] Reviewing item difficulty data can help faculty refine MCQs to align with exam goals. Item difficulty shows the percentage of students who answered a particular question correctly. This allows faculty to identify questions that may be easier or more difficult than what is appropriate or intended for an exam.[11] Effective use of this postexam data can help faculty identify quality questions and determine where to focus their revision efforts. The process of creating, evaluating, and revising questions is important, but can require considerable time and attention to detail.

Artificial intelligence (AI) tools provide opportunities for faculty to save time and enhance the way they work.[13] ChatGPT (Chat Generative Pre-trained Transformer, a predictive language generation software program developed by OpenAI) is an example of an AI tool that has received attention for helping faculty create classroom activities, simulation scenarios, discussion forums, knowledge assessments, and more.[14–16] Recently, faculty in science- and health care–related fields have evaluated the effectiveness of ChatGPT in creating exam questions.[17–19] For example, Cox et al compared ChatGPT-generated National Council Licensure Examination–type questions with human-generated National Council Licensure Examination–type questions.[18] The authors determined that both methods produced relevant, clear, and grammatically correct questions with understandable options. ChatGPT has also successfully produced valid and relevant biology exam questions and computer science questions.[19,20] However, not all questions produced by ChatGPT are perfect.[3,20] For example, Ngo et al found that 25% of the multiple-choice medical exam questions created by ChatGPT were wrong or misleading, thus highlighting the possible limitations of ChatGPT and the need for faculty to review and refine questions to ensure accuracy.[3]

To create an MCQ in ChatGPT, the user should provide clear and detailed instructions in their prompt. Complicated, multipart prompts may lead to errors, as ChatGPT might misunderstand or ignore some instructions.[21] When creating MCQs for medical school exams, Zuckerman et al found that they needed to rephrase the prompts used in the initial attempts to correct MCQs that focused on the wrong topic or omitted expected information.[22] When faculty in their study felt that adequate quality was achieved, they edited the questions to remove distractors that were not taught, changed item wording to match what students had learned, and added clinically relevant details to the question stem. Despite the work to refine the questions they created with ChatGPT, the authors noted that they still spent less time creating a question than they would have without the use of ChatGPT.[22] Cheung et al also found that creating medical exam MCQs in ChatGPT took significantly less time compared with the time needed to generate human-created questions of similar quality.[17] Given the workload of faculty today, this is an encouraging finding.

Currently, researchers do not know if ChatGPT can produce quality Board of Certification (BOC)–style questions for use by athletic training faculty. This study aims to compare human-generated athletic training–related MCQs with those generated by ChatGPT. The goal is to examine the quality, clarity, relevance, and difficulty of the questions produced by both methods to learn more about the potential for using AI tools such as ChatGPT to help faculty create fair and valid exam or practice questions for their courses.

Table 1. Example of Question Pairs (Domain 2: Assessment, Evaluation, and Diagnosis)

	Human-Generated MCQ	ChatGPT-Generated MCQ
Pair 3	A patient presents with the following signs and symptoms: • Swelling behind the medial malleolus. • Point tenderness under the medial malleolus. • Burning, numbness, and tingling over the bottom of the foot and toes. • Pain worsens as the day progresses (especially with running or standing for long periods. • Joint and muscle function: all within normal limits • Positive Tinel sign Based on this information, what condition should the athletic trainer suspect? A. Tarsal Tunnel Syndrome B. Plantar Fasciitis C. Medial Tibial Nerve Syndrome D. Sural Nerve Injury E. Morton's Neuroma	During a routine evaluation, an athlete presents with complaints of burning pain, numbness, and tingling sensations in the sole of the foot. Physical examination reveals Tinel's sign and reproduction of symptoms upon tapping the posterior tibial nerve. Based on these findings, what additional diagnostic test would be most appropriate for confirming a diagnosis of tarsal tunnel syndrome? A. Electromyography (EMG) and nerve conduction studies. B. MRI (Magnetic Resonance Imaging) of the foot and ankle. C. Ultrasound imaging of the tarsal tunnel area. D. X-ray examination of the foot to assess bone abnormalities. E. Blood tests to rule out autoimmune or systemic disorders.
Pair 4	Upon visual inspection, an athletic trainer observes an athlete's second digit in extension of the MCP and DIP joints and flexion of the PIP joint. Based on this information, what condition should the athletic trainer suspect? A. Pseudo-boutonniere deformity B. Mallet finger C. Swan deck deformity D. Jersey Finger E. Trigger Finger	A gymnast sustains a severe finger injury during practice, leading to a noticeable deformity. As an athletic trainer, you are tasked with assessing the injury to determine the appropriate intervention. When assessing a severe finger injury resulting in deformity, which of the following signs would most likely indicate a volar plate injury? A. Tenderness over the distal phalanx B. Hyperextension of the proximal interphalangeal joint C. Lateral deviation of the finger D. Pain with axial loading of the fingertip E. Swelling and tenderness over the proximal interphalangeal joint

Abbreviations: DIP, distal Interphalangeal joint; MCP, metacarpophalangeal joint; MCQ, multiple-choice question; PIP, proximal Interphalangeal joint.

METHODS

We used a cross-sectional design that included a web-based survey to examine faculty views and a digital exam to assess student performance on the human-generated and ChatGPT-generated MCQs. The institutional review board at Xavier University reviewed and approved the methods, protocols, and instruments for each part of this study. The Checklist for Reporting of Survey Studies was used as a guideline to prepare the present manuscript.[23]

Instrumentation

We constructed a survey for athletic training faculty with 3 sections. The first section included consent and 2 questions to determine inclusion. Participants were asked if they had ever taught in a didactic setting and if they were familiar with the format of BOC-style exam questions. Participants were included only if they answered yes to both questions. The next section asked participants to evaluate 10 pairs of MCQs. Each pair included a human-generated question and a ChatGPT-generated question on a similar topic. Please see Table 1 for an example of 2-question pairs.

Ten human-generated MCQs were selected by the research team from previously used program exams. The research team used Microsoft Word's grammar and spell check functions when the questions were originally created. Two questions were chosen from each of the 5 athletic training domains. Each question met the following criteria: (1) multiple-choice format with 5 answer options, (2) Bloom taxonomy application level or higher, (3) corrected item-total correlation coefficient of 0.3 or above on a recent exam, and (4) aligned with BOC exam question creation guidelines.

ChatGPT version 3.5 (current free version in February 2024) was used to create 10 MCQs with similar content to human-generated questions. To align ChatGPT-generated questions with human-generated questions, the prompts inputted into ChatGPT used a similar template to that of Cox et al, which included a specific topic and Bloom taxonomy level.[18] For example, "Create an athletic training BOC-style, multiple-choice question about [topic] at the application level of Bloom's taxonomy with a short or medium prompt. Create 5 answer options." For example, one ChatGPT prompt stated: "Create an athletic training BOC-style, multiple-choice question about evaluation of tarsal tunnel syndrome at the application level of Bloom's taxonomy with a short or medium

prompt. Create 5 answer options." See Table 1 for the resulting MCQ.

Participants were not aware of the origin of the question (human or ChatGPT). Faculty participants evaluated each question on grammar, clarity, difficulty, terminology, and suitability using a 5-point Likert scale (1 = *very poor* to 5 = *very good*). Participants responded to 2 separate questions about clarity, 1 for the stem, and 1 for the answer options. They also evaluated whether each question would be at an appropriate level of difficulty for an entry-level athletic trainer. Participants rated the medical terms and abbreviations for accuracy/appropriateness and the suitability of each question to effectively address entry-level athletic training content. Participants were then asked which question from the pair they would be more likely to use in an examination (question 1, question 2, neither). The final section included demographic questions. The survey was pilot tested by 2 athletic training faculty members not affiliated with this study. The survey was revised to improve clarity and decrease completion time.

A student quiz was also developed to further evaluate question quality/difficulty. Each MCQ used in the faculty survey was uploaded to Canvas, a web-based learning management system. Current second-year master's students nearing graduation were asked to complete the 20-question quiz. Scores from this 20-question quiz were used to determine the corrected item-total correlation coefficient and item difficulty scores for the ChatGPT-generated and human-generated MCQs used in the survey sent to athletic training faculty.

Participants

Masters-level professional athletic training programs were identified in each state using the Commission on Accreditation of Athletic Training Education website. Faculty contact information was collected from publicly available directories on the selected institutions' website. Second-year students enrolled in host institutions' master's-level professional athletic training program were recruited. All student participants are known to the lead investigator.

Procedures

The research team emailed 653 athletic training faculty members teaching at master's-level professional athletic training programs to request their participation in this study. We sent a reminder email 3 weeks later. The survey was hosted on the Qualtrics platform (Qualtrics). After providing informed consent, faculty members were asked to complete the survey. Completion of all survey items took approximately 20 minutes. Data were collected anonymously.

We sent an email to 13 current second-year master's-level athletic training students requesting they complete a 20-question, multiple choice quiz. This quiz did not affect student participants' grade in any course. Confidentiality was assured, and informed consent was obtained from each participant.

Data Analysis

Faculty survey results were exported from Qualtrics into SPSS version 26 (IBM Corp). Questions related to quality, clarity, relevance, and difficulty were rated on a 5-point Likert scale (1 = *very poor* to 5 = *very good*) and compared using the Wilcoxon signed rank test. Questions about MCQ preference were analyzed descriptively as frequency and percentage.

For the 20-question student quiz, we used the Canvas quiz and item analysis report that provides the corrected item-total correlation coefficient and item difficulty score for each question. The corrected item-total correlation coefficient for each question was examined to determine if a question had good distractions and provided good discrimination (score of 0.25 or above) or if a question may have been miskeyed or confusing for students (negative score). Item difficulty scores were used to identify questions that may have been too easy or too difficult.

RESULTS

The survey garnered 93 responses from athletic training faculty (73 complete responses + 20 partial responses), with a total response rate of 7%. Responses were not required for all questions. Table 2 summarizes faculty participant demographics.

Grammar

Faculty participants rated grammar as acceptable to good (range, 3.4–4.3) for all questions. Human-generated questions ranged from 3.4 to 4.0, whereas ChatGPT-generated questions were rated from 3.6 to 4.3. Table 3 shows significant differences in grammar quality for 4 question pairs. In these 4 question pairs, faculty participants rated the ChatGPT-generated questions higher, indicating better grammar.

Stem

Faculty rated the quality of the question stem as acceptable to good, ranging from 3.2 to 4.2. Human-generated questions ranged from 3.2 to 4.1, whereas ChatGPT-generated questions ranged from 3.2 to 4.2. Five question pairs exhibited statistically significant differences in the quality of the question stem (see Table 3). In 3 cases, faculty participants rated the ChatGPT-generated questions higher, indicating a better stem.

Answers

Faculty rated the quality of answers for both human- and ChatGPT-generated questions as acceptable to good, ranging from 3.2 to 4.3. Seven question pairs showed statistically significant differences in the quality of answers (see Table 3). In 4 of these pairs, human-generated questions were scored higher.

Difficulty

Athletic training faculty rated the difficulty of the questions as acceptable for entry-level athletic training students, ranging from 3.0 to 3.9. Human-generated questions ranged from 3.0 to 3.9, whereas ChatGPT-generated questions ranged from 3.1 to 3.9. Statistically significant differences in perceived difficulty ratings were observed in 5 question pairs; in 3 of these pairs, ChatGPT-generated questions were perceived as having a better or more appropriate level of difficulty for entry-level athletic training students than their human-generated counterparts (see Table 3).

Table 2. Faculty Participant Demographics

	No. (%)
Age	
Under 25	0
25–29	4 (6)
30–34	17 (23)
35–39	12 (16)
40+	39 (53)
Prefer not to respond	1 (1)
Gender	
Man	30 (41)
Woman	43 (59)
Transgender man	0
Transgender woman	0
Nonbinary/nonconforming	0
Identity not listed, write in:	0
Prefer not to respond	0
Ethnicity	
Asian	0
Black or African American	4 (6)
Hispanic American or Latino/a	1 (1)
Middle Eastern or North African	0
Native American or Alaskan Native	0
Native Hawaiian or Pacific Islander	0
White or Caucasian	64 (88)
Biracial or multiracial	2 (3)
Race/ethnicity not listed, write in:	0
Prefer not to respond	2 (3)
Faculty rank	
Full-time instructor/nontenure track	24 (33)
Tenure-track faculty (not tenured)	21 (29)
Tenured faculty	27 (37)
Part-time instructor or adjunct faculty member	1 (1)
Highest degree	
Bachelor's	0
Master's	13 (18)
Doctoral	60 (82)
Role in program	
Program director	18 (25)
Clinical education coordinator	27 (37)
Faculty, nonadministrative	28 (38)
Years teaching at the college level	
0–3	11 (15)
4–7	11 (15)
8–12	20 (27)
12+	31 (43)
Type of institution where you are primarily employed	
Public (state) university	42 (57)
Community college	0
Private nonprofit university	30 (41)
Other	1 (1)

Terms

Athletic training faculty rated the use of terms as poor to good, ranging from 2.9 to 4.1. Human-generated questions ranged from 2.9 to 3.9, whereas ChatGPT-generated questions ranged from 3.4 to 4.1. Four question pairs showed statistical differences. In 3 of the 4 question pairs, ChatGPT-generated questions were rated higher, indicating more appropriate terminology use (see Table 3).

Suitability

Athletic training faculty rated the suitability of questions as poor to good, ranging from 2.8 to 4.0. Human-generated questions ranged from 2.8 to 3.8, whereas ChatGPT-generated questions ranged from 2.9 to 4.0. In 7 question pairs, significant differences were observed, with human-generated questions deemed more suitable in 4 of the question pairs (see Table 3).

Question Preference

There was no clear preference for either type of MCQ. When comparing the question pairs, participants preferred 5 questions generated by ChatGPT and 4 questions generated by humans. In 1 instance, preferences were evenly split between the 2 types of questions, as shown in Table 4. Each of the 5 athletic training domains included 2 question pairs. In 3 of these domains, participants favored 1 ChatGPT-generated question and 1 human-generated question. Overall, question preference was balanced across the different athletic training domains.

Student Results

Eleven student participants completed the 20-question student quiz, with a response rate of 85%. The majority of student participants were under the age of 25 (82%). Table 5 summarizes student participant demographics.

Three human-generated questions and 1 ChatGPT question had an item difficulty score above 0.85, indicating that they may have been too easy (see Table 6). Only 1 human-generated question and 1 ChatGPT question had an item difficulty score below 0.30, indicating it may have been difficult. The remaining questions fell within an acceptable range. Six human-generated questions and 5 ChatGPT-generated questions achieved a corrected item-total correlation coefficient value of 0.25 or above, indicating that these questions had good distractors.[12] There were 2 human-generated and 5 ChatGPT-generated questions with a value below 0.25 and 3 ChatGPT-generated questions that received negative values, which may indicate that a question is miskeyed or confusing. All participants answered 2 human-generated questions correctly. In these cases, a corrected item-total correlation coefficient score could not be calculated and "NA" (not applicable) appears in the table.

Overall, human-generated and ChatGPT-generated questions showed a range of difficulty (item difficulty values between 0.27 and 1.0). Higher scores (eg, 0.80) mean that more students answered the question correctly and the question was easier. Lower scores (eg, 0.20) mean that fewer students answered the question correctly and the question was more difficult. There were 4 human-generated and 3 ChatGPT-generated MCQs with item difficulty scores over 0.80. There were no questions with an item difficulty score below 0.27.

DISCUSSION

The results of this study demonstrate that faculty found ChatGPT-generated, BOC-style questions had similar values for grammar, stem quality, answer quality, question difficulty, proper use of medical terminology, and suitability for content to human-generated questions. This was true for questions

Table 3. Question Item Ratings[a]

	Grammar		Stem		Answers		Difficulty		Terms		Suitability	
Question Set	Mean ± SD	*P*	Mean ± SD	*P*	Mean ± SD	*P*	Mean ± SD	*P*	Mean ± SD	*P*	Mean ± SD	*P*
1 (n = 93)												
Human	3.5 ± 0.92		3.7 ± 1.04		3.2 ± 0.98		3.4 ± 0.89		3.6 ± 0.89		3.2 ± 0.96	
ChatGPT	4.3 ± 0.73	.000[b]	4.1 ± 0.88	.005[b]	4.3 ± 0.68	.000[b]	3.8 ± 0.84	.000[b]	4.0 ± 0.88	.000[b]	4.0 ± 0.88	.000[b]
2 (n = 85)												
Human	4.0 ± 0.81		3.9 ± 0.88		4.1 ± 0.94		3.7 ± 0.92		3.9 ± 0.87		3.7 ± 1.02	
ChatGPT	4.1 ± 0.77	.355	3.9 ± 0.97	.835	3.3 ± 1.31	.000[b]	3.1 ± 1.05	.000[b]	3.8 ± 0.86	.407	3.2 ± 1.26	.004[b]
3 (n = 82)												
Human	3.8 ± 0.85		3.8 ± 0.94		4.3 ± 0.74		3.8 ± 0.88		3.8 ± 0.90		3.8 ± 0.87	
ChatGPT	3.9 ± 0.95	.513	3.7 ± 1.04	.549	3.7 ± 1.08	.000[b]	3.5 ± 0.97	.053	3.6 ± 1.01	.103	3.5 ± 0.98	.008[b]
4 (n = 82)												
Human	3.8 ± 1.05		3.6 ± 1.19		4.2 ± 0.84		3.7 ± 0.96		3.9 ± 0.95		3.8 ± 0.94	
ChatGPT	3.8 ± 0.93	.941	3.6 ± 1.12	.884	3.7 ± 1.06	.000[b]	3.8 ± 0.90	.140	3.8 ± 0.92	.526	3.8 ± 0.88	.764
5 (n = 81)												
Human	3.8 ± 0.90		3.6 ± 1.16		3.6 ± 1.09		3.4 ± 0.88		3.6 ± 0.95		3.3 ± 1.14	
ChatGPT	4.1 ± 0.75	.007[b]	3.9 ± 0.86	.012	4.0 ± 0.96	.015[b]	3.9 ± 0.72	.000[b]	4.0 ± 0.82	.013[b]	3.9 ± 0.77	.000[b]
6 (n = 78)												
Human	4.0 ± 0.86		4.1 ± 0.81		4.3 ± 0.77		3.9 ± 0.86		3.9 ± 0.91		3.8 ± 0.89	
ChatGPT	4.2 ± 0.72	.033[b]	4.2 ± 0.85	.512	4.1 ± 0.95	.066	3.8 ± 0.93	.444	4.0 ± 0.90	.599	3.8 ± 1.06	.680
7 (n = 76)												
Human	3.9 ± 1.04		3.8 ± 1.10		3.7 ± 1.15		3.6 ± 1.05		3.6 ± 1.06		3.6 ± 1.11	
ChatGPT	3.8 ± 0.98	.795	3.4 ± 1.22	.014[b]	3.5 ± 1.24	.123	3.5 ± 1.10	.274	3.6 ± 1.09	.377	3.3 ± 1.23	.037
8 (n = 74)												
Human	3.4 ± 1.02		3.2 ± 1.12		3.5 ± 1.10		3.0 ± 1.04		2.9 ± 1.14		2.8 ± 1.06	
ChatGPT	4.1 ± 0.75	.000[b]	4.0 ± 0.87	.000[b]	4.1 ± 0.87	.000[b]	3.8 ± 0.91	.000[b]	4.1 ± 0.79	.000[b]	3.8 ± 0.87	.000[b]
9 (n = 73)												
Human	3.7 ± 1.07		3.6 ± 1.09		3.7 ± 1.08		3.5 ± 1.06		3.7 ± 1.02		3.4 ± 1.15	
ChatGPT	3.6 ± 0.99	.560	3.2 ± 1.15	.006[b]	3.2 ± 1.11	.000[b]	3.1 ± 1.08	.028[b]	3.4 ± 1.05	.037[b]	2.9 ± 1.16	.015[b]
10 (n = 73)												
Human	4.0 ± 0.85		3.8 ± 0.97		3.9 ± 1.00		3.6 ± 0.98		3.9 ± 0.85		3.7 ± 0.99	
ChatGPT	4.0 ± 0.75	.407	3.9 ± 0.92	.285	3.9 ± 0.91	.939	3.7 ± 0.87	.120	3.9 ± 0.95	.830	3.6 ± 1.08	.965

[a] Ratings: 1 = *very poor*; 2 = *poor*; 3 = *acceptable*; 4 = *good*; 5 = *very good*.

[b] $P < .05$.

Table 4. Question Item Preference

Athletic Training Domain	Question Set (Total Responses)	No. (%) Human Generated	ChatGPT Generated	Neither
Risk reduction, wellness, and health literacy	1 (93)	16 (17.2)	73 (78.5)	4 (4.3)
	2 (85)	51 (60.0)	28 (32.9)	6 (7.1)
Assessment, evaluation, and diagnosis	3 (82)	52 (63.4)	26 (31.7)	4 (4.9)
	4 (82)	40 (48.8)	40 (48.8)	2 (2.4)
Critical incident management	5 (81)	16 (19.8)	57 (70.4)	8 (9.9)
	6 (76)	31 (40.8)	39 (51.3)	6 (7.9)
Therapeutic intervention	7 (76)	39 (51.3)	23 (30.3)	14 (18.4)
	8 (74)	9 (12.2)	57 (77.0)	8 (10.8)
Health care administration and professional responsibility	9 (73)	32 (43.8)	28 (38.4)	13 (17.8)
	10 (73)	24 (32.9)	41 (56.2)	8 (10.9)

related to all 5 athletic training domains. Most of the questions created by ChatGPT were easy to understand, used appropriate terminology, and had answer options that were similar in style and length.

Grammar and Terminology

Athletic training faculty rated grammar as acceptable to good for both human-generated and ChatGPT-generated questions. The ChatGPT-generated questions had slightly higher grammar scores in 4 cases, indicating that faculty might want to consider using ChatGPT to improve grammar in their human-generated questions. Additionally, for most of the MCQs, the use of medical terms and abbreviations was considered accurate and appropriate. These results align with previous studies showing that ChatGPT can create MCQs that follow general rules of grammar and syntax.[18,19] For example, Cox et al found that nursing faculty rated ChatGPT-generated questions and nursing faculty–generated questions similarly in their clarity and grammar.[18] Nasution found that 73% of biology students thought the AI-generated questions in their study were without grammatical or conceptual errors.[19] The author noted that although they did encounter a language or sentence issue in a question created by AI, an expert could fix the mistake during the review process.[19] Therefore, ChatGPT appears to be an effective tool to generate MCQs that closely reflect natural language and use acceptable medical terminology, but careful review is still needed.

Table 5. Student Participant Demographics

	No. (%)
Age	
Under 25	9 (82)
25–29	1 (9)
30+	1 (9)
Prefer not to respond	0
Gender	
Man	5 (45)
Women	6 (54)
Transgender man	0
Transgender woman	0
Nonbinary/nonconforming	0
Identity not listed, write in:	0
Prefer not to respond	0
Ethnicity	
Asian	0
Black or African American	0
Hispanic American or Latino/a	0
Middle Eastern or North African	0
Native American or Alaskan Native	0
Native Hawaiian or Pacific Islander	0
White or Caucasian	11 (100)
Biracial or multiracial	0
Race/ethnicity not listed, write in:	0
Prefer not to respond	0

Question Stems and Answer Options

Athletic training faculty also found the question stems and answer options acceptable for both human-generated and ChatGPT-generated questions. Although faculty noted a wider range in the quality of ChatGPT-generated question stems, the ChatGPT questions scored higher in 3 of the 5 cases of statistical difference. Previous studies examining ChatGPT-generated MCQ quality have found mixed results.[3,18,19,24] For example, Cox et al found that ChatGPT generated clear stems with correct answers and appropriate distractors for nursing content, although the authors noted that questions could be improved by faculty with content knowledge.[18] Cheung et al found similar results for medical education MCQs.[17] However, a range of content accuracy rates was found in a review of 23 studies that used ChatGPT to create medical education MCQs.[25] Additionally, Ngo et al found incorrect answers and explanations in most immunology MCQs generated by ChatGPT and remarked that 43% of questions would need significant changes before they could be used.[3] Our results are consistent with Ngo et al in that our human-generated questions scored better for correct answers more often (57%; 4 of 7) than the AI-generated questions when question pairs were statistically different from each other.[3] Overall, ChatGPT provides a promising tool for automatically generating MCQs, but faculty should expect to review questions and provide necessary changes to ensure accuracy and clarity for both questions' stems and answers.

Question Difficulty

Athletic training faculty rated the difficulty of each pair of MCQs as acceptable. Most question pairs were perceived as having an equally appropriate level of difficulty, whereas 3

Table 6. Student Results

Athletic Training Domain	Set	Human Generated		ChatGPT Generated	
		Item Difficulty	Corrected Item-Total Correlation Coefficient	Item Difficulty	Corrected Item-Total Correlation Coefficient
Risk reduction, wellness, and health literacy	1	0.82	0.49	0.36	−0.44
	2	0.55	0.41	0.36	−0.7
Assessment, evaluation, and diagnosis	3	1	NA	0.82	0.67
	4	0.36	0.18	0.55	0.41
Critical incident management	5	0.45	0.28	0.36	−0.02
	6	0.91	0.84	0.82	0.49
Therapeutic intervention	7	0.27	0.02	0.36	0.18
	8	0.64	0.50	0.73	0.39
Health care administration and professional responsibility	9	0.64	0.43	0.27	0.23
	10	1	NA	0.91	0.84

Abbreviation: NA, corrected item-total correlation coefficient could not be calculated due to all participants answering the question correctly.

ChatGPT-generated questions and 2 human-generated questions were rated as having a more appropriate level of difficulty than their counterpart question. Overall, both types of questions could effectively address entry-level athletic training content.

Results from the student quiz data showed that most human-generated and ChatGPT-generated MCQs were within an acceptable level of difficulty. There were 4 questions that could have been considered too easy and only 2 questions that might have been too challenging. The faculty difficulty Likert score ratings did not forecast student performance for either the human-generated or ChatGPT-generated questions (Tables 3 and 6). This might be accounted for by the difference in measurement scales. Faculty were asked if the question had an appropriate amount of difficulty rather than if it was a more difficult question.

Item Discrimination

Results from the student data showed that most human-generated MCQs had good item discrimination, with only 2 questions showing a need for revision. However, half of the ChatGPT-generated questions had values indicating poor discrimination. Additionally, 3 of these questions had negative values, indicating that these questions may have been ambiguous or confusing. This underscores the importance of expert review prior to using new questions on exams and the need for analyzing the data generated when these questions are used in course assessments.

Suitability and Preference

Suitability ratings were mixed, ranging from poor to good for both human- and AI-generated questions. There was no clear preference overall, with participants favoring an almost equal number of human- and ChatGPT-generated questions when there was a statistical difference in suitability. The athletic training domain did not seem to impact human versus ChatGPT question preference. It was common for at least 1 human-generated question to be preferred per domain. In 7 of the 10 question pairs, there was a statistically significant difference in suitability. In each of these 7 pairs, the question considered "more suitable" was also the preferred question. This trend was also seen with difficulty ratings and question preference, where statistically significant differences were found in difficulty ratings in 5 question pairs, and the preferred question was also a question that was considered to have a more appropriate level of difficulty.

LIMITATIONS AND FUTURE DIRECTIONS

Although we tried to limit the time commitment for survey completion, some respondents did not rate every question pair, with fewer responses toward the end of the survey. This may have influenced the results of the question pairs that were later in the survey. Despite efforts to include all faculty currently teaching in accredited athletic training programs across the country, we relied on publicly available directories from each institution's website. Some directories may not have been up-to-date. Finally, student participants were a convenience sample from the investigators' institution and represented only a single university.

The quality of questions produced by ChatGPT is affected by the provided prompts. ChatGPT can return more useful output, in this case exam questions, when better prompts are used, ie, with prompt engineering.[26] In addition, ChatGPT's output is improved when the human has a dialogue with the program that allows for refinement of the output that addresses the user's concerns with the initial output. The results of this study might be limited by the simple prompt style that was used rather than a dialogue approach. Once a question was produced by ChatGPT, no further refinement was requested by the investigators. With careful review of ChatGPT output and appropriate prompt engineering, the ChatGPT questions could receive higher ratings. Faculty are encouraged to learn about prompt engineering to improve the effectiveness of using ChatGPT to write exam questions. Future researchers should explore and develop systematic approaches for prompt engineering, including guidelines to optimize the quality of ChatGPT-generated questions. Pilot studies using iterative testing and refinement of prompts could help identify the most effective strategies. Collaborating with ChatGPT experts to create and evaluate prompts may also improve the relevance and quality of the questions. Future authors should also consider examining other types of questions such as multiple selection, case based, or questions that are part

of a focused testlet. It would also be interesting to investigate how much time is saved in item creation. This research could provide further information regarding the utility of using ChatGPT to produce exam questions.

Recommendations for Athletic Training Faculty

Given the results of this study and the increased demands on faculty members, we recommend that athletic training educators consider using ChatGPT to generate quality quiz and exam questions. For faculty who have not used ChatGPT, they could consider using our simple template to generate an initial draft of a question, then using an approach similar to that proposed by Zuckerman et al, in which a human instructor reviews and edits the question to remove distractors that were not taught, changes item wording to match what students had learned, and adds clinically relevant details to the question stem.[22] With this approach, educators can leverage ChatGPT to reduce the burden of crafting effective questions, while still ensuring that questions are tailored to the content and level of the students.

Once a reasonable question has been produced, we can use multiple methods to ensure that a question is fair and valid. First, we can identify and correct common writing flaws such as excess verbiage in the stem, use of implausible distractors and absolute terms (eg, always, never), and making the correct answer more detailed or longer.[27,28] After a test has been administered, we can use item analyses to identify questions that need to be edited or removed.[12] If an exam is housed in a university's learning management system such as Canvas, item analysis results are provided for instructors. In this study, obtaining results for item difficulty and corrected item-total correlation coefficients identified multiple questions that could be improved with modifications.

CONCLUSIONS

ChatGPT is another tool that athletic training faculty may consider using to improve the quality and efficacy of exam question preparation. The data from this study suggest that faculty can effectively use ChatGPT for exam question preparation; however, faculty should understand that ChatGPT, like all tools, has its limitations.

REFERENCES

1. Collins J. Education techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*. 2006;26(2):543–551. doi:10.1148/rg.262055145
2. Towns MH. Guide to developing high-quality, reliable, and valid multiple-choice assessments. *J Chem Educ*. 2014;91(9):1426–1431. doi:10.1021/ed500076x
3. Ngo A, Gupta S, Perrine O, Reddy R, Ershadi S, Remick D. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Acad Pathol*. 2024;11(1):100099. doi:10.1016/j.acpath.2023.100099
4. Smith LS. How to write better multiple-choice questions. *Nursing*. 2023;48(11):14–17. doi:10.1097/01.NURSE.0000546471.79886.85
5. Torres C, Lopes A, Babo L, Azevedo J. Improving Multiple-Choice Questions. *US-China Education Review B1*. 2011;1(1):1–11.
6. Paniagua MA, Swygert KA, eds. *Constructing Written Test Questions for the Basic and Clinical Sciences*. 4th ed. Revised. National Board of Medical Examiners; 2016.
7. Holsgrove G, Elzubeir M. Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. *Med Educ*. 1998;32(4):343–350. doi:10.1046/j.1365-2923.1998.00203.x
8. Haladyna T. Creating multiple-choice items for testing student learning. *Int J Assess Tools Educ*. 2022;9(special issue):6–18. doi:10.21449/ijate.1196701
9. Downing SM. 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. *Adv Health Sci Educ Theory Pract*. 2005;10(2):133–143. doi:10.1007/s10459-004-4019-5
10. Downing SM, Haladyna TM. Test item development: validity evidence from quality assurance procedures. *Appl Meas Educ*. 1997;10(1):61–82. doi:10.1207/s15324818ame1001_4
11. Considine J, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*. 2005;12(1). doi:10.1016/s1322-7696(08)60478-3
12. Varma S. *Preliminary Item Statistics Using Point-Biserial Correlation and P-Values*. Educational Data Systems Inc; 2006.
13. Owan VJ, Abang KB, Idika DO, Etta EO, Bassey BA. Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia J Math Sci Technol Educ*. 2023;19(8):em2307. doi:10.29333/ejmste/13428
14. Abujaber AA, Abd-Alrazaq A, Al-Qudimat AR, et al. A strengths, weaknesses, opportunities, and threats (SWOT) analysis of ChatGPT integration in nursing education: a narrative review. *Cureus*. 2023;15(11):e48643. doi:10.7759/cureus.48643
15. Jeyaraman M, K SP, Jeyaraman N, et al. ChatGPT in medical education and research: a boon or a bane? *Cureus*. 2023;15(8):e44316. doi:10.7759/cureus.44316
16. Schneider K, Tomchuk D, Snyder B, Bisch T, Koch G. Incorporating artificial intelligence into athletic training education: developing case-based scenarios using ChatGPT. *Athl Train Educ J*. 2024;19(1):42–50. doi:10.4085/1062-6050-028.23
17. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). *PLoS One*. 2023;18(8):e0290691. doi:10.1371/journal.pone.0290691
18. Cox RL, Hunt KL, Hill RR. Comparative analysis of NCLEX-RN questions: a duel between ChatGPT and human expertise. *J Nurs Educ*. 2023;62(12):679–687. doi:10.3928/01484834-20231006-07
19. Nasution NEA. Using artificial intelligence to create biology multiple choice questions for higher education. *Agric Environ Educ*. 2023;2(1):em002. doi:10.29333/agrenvedu/13071
20. Tran A, Angelikas K, Rama E, Okechukwu C, Smith DH, MacNeil S. Generating multiple choice questions for computing courses using large language models. In: 2023 IEEE Frontiers in Education Conference (FIE). IEEE: 2023:1–8. doi:10.1109/FIE58773.2023.10342898
21. Gonsalves C. On ChatGPT: what promise remains for multiple choice assessment? *J Learn Dev High Educ*. 2023;(27). Preprint posted online April 27, 2023. doi:10.47408/jldhe.vi27.1009
22. Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, Lockspeiser T. ChatGPT for assessment writing. *Med Teach*. 2023;45(11):1224–1227. doi:10.1080/0142159X.2023.2249239

23. Sharma A, Minh Duc NT, Luu Lam Thang T, et al. A consensus-based checklist for reporting of survey studies (CROSS). *J Gen Intern Med*. 2021;36(10):3179–3187. doi:10.1007/s11606-021-06737-1
24. Kıyak YS. A ChatGPT prompt for writing case-based multiple-choice questions. *Rev Esp Educ Méd*. 2023;4(3):98–103. doi:10.6018/edumed.587451
25. Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgrad Med J*. 2024;100(1189):858–865. doi:10.1093/postmj/qgae065
26. Vidhani DV, Mariappan M. Optimizing HUMAN-AI collaboration in chemistry: a case study on enhancing generative AI responses through prompt engineering. *Chemistry*. 2024;6(4):723–737. doi:10.3390/chemistry6040043
27. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple choice item writing guidelines for classroom assessment. *Appl Meas Educ*. 2002;15(3):309–334. doi:10.1207/S15324818AME1503_5
28. Tarrant M, Ware J. Impact of item-writing flaws in multiple choice questions on student achievement in high-stakes nursing assessments. *Med Educ*. 2008;42(2):198–206. doi:10.1111/j.1365-2923.2007.02957.x