

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Smart and safe use of AI tools for setting assessments

Mark Schier, University of New England

Generative AI tools such as ChatGPT, Microsoft Copilot and Gemini have often been promoted in negative ways as they may be used by students in an attempt to breach academic integrity. One issue with assessment question pools is that they are often set and not updated frequently. Hence, they can be overused and are vulnerable to malicious use such as: being sold for money or ported to the dark web and therefore their practical value for assessment is significantly diminished. There is another side of gen AI that can assist academics in authoring assessment questions that is substantially different to the standard question pools that academics often rely on. This paper will illustrate potential uses, benefits, and precautions that need to be taken when using gen AI for assessment design.

Keywords: AI tools, Assessment, MCQ, ChatGPT, academic integrity

Background

The quality and integrity of an assessment is often affected by available time of the academic or team producing the assessment task. Many academics are time poor (Kinman, 2001; Newson & Menzies 2007) and have significant inherent pressures such as managing multiple roles. These include teaching, researching, scholarship, leadership, community engagement, mentoring, pastoral care, external editorial, professional societies (Boyd et al., 2011; McCune, 2018) and having life outside the workplace!

Many universities and institutions have incredibly short turnaround times for final assessment pieces in subjects (also known as: units, courses, or modules). Typical turnaround times at universities is 14 days, but for the final assessment piece it could be as short as 7-10 days due to resulting deadlines. This indicates that front-end effort is an effective way of moving the time-consuming element of assessment practice to the writing of the question and not the grading, which may be done automatically online, or with minimal intervention. This is often why multiple-choice questions (MCQ) are used in several final assessments in the exam period. Additionally, short and long answer questions can also be used in conjunction with MCQs.

This paper describes using a novel approach to question generation in the discipline of Anatomy and Physiology at an Australian University.

Multiple Choice Questions (MCQ)

This paper will not discuss the educational value of utilising MCQ for assessment, which is a separate topic entirely (Timmis et al, 2016) and is a debate that has been simmering away for many years (e.g., Glaser & Silver, 1994). This paper assumes that there are sufficient reasons for making use of this type of assessment and describes the use of gen AI to generate effective questions.

The danger of MCQs as secured assessment is that the pool may be used on many occasions without being updated, rewritten or only having a small percentage of questions added. This creates a risk of the pool becoming static, stagnant or compromised. In some instances, students may be aware that the pool is not updated regularly and therefore effectively closed. This has the unwanted and counterintuitive effect of making the question pool open. There are websites that provide clandestine (and often extant) question banks for subjects on both the free and dark web or equivalent (Ruggiero, 2016). Once these assessments are accessed and part of the ecosystem then students have the actual questions with answers, and their knowledge and understanding of the content is not securely assessable. If the questions are minimally modified, then there is a risk that students will only learn to pass the assessment, with a very superficial

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

understanding of the content. This is evident in areas where there are external exams, such as: International English Language Testing System (IELTS) and Graduate Medical School Admissions Test (GAMSAT). Commercial coaching schools exist for these external examinations and have been criticised as poor vehicles for actual learning and understanding the content, as distinct from learning how to do the test (Templar, 2004).

There are benefits and drawbacks for using MCQs in assessments, and these need to be considered for their practicality for the task at hand. Some of these are summarised in Table 1.

Table 1

Benefits and Drawbacks of MCQs

Benefits	Drawbacks
Easy to grade	Take time to write and check
Consistent across students	Poor understanding of quality
Sampled question pools can be used for variety	Re-used too frequently.
Known quantity for students	Can over-assess recall.
	Pools may be found by students online

One obvious solution to minimise learning of questions or online discovery of pools is to generate new question pools each instance to provide a fresh set without the inherent risk of assessing learned answers rather than knowledge or understanding of the subject content. This is not usually considered due to the time taken to generate a set of questions.

So, we return to the issue highlighted initially – that of scarcity of time for educators to produce either unique questions or series of questions that have a similar level of difficulty for consistency across the final assessment tasks. This paper addresses this issue in some detail.

Short and Long Answer Questions

An alternative or adjunct to MCQs are questions that require written responses. These are better at assessing a deep understanding of concepts as they may be of a form that asks students to do more than just provide a correct alternative. For example, to compare or contrast, explain briefly, or put in context a series of terms.

These are generally questions which are much easier to write than MCQs but take longer to grade as students answer in their own words reflecting their understanding of the question and thought processes. Assessing these answers does not fit well with the marking turnaround times that are imposed by administrative systems. These take less time to craft than MCQs, but take more time to assess, compounding the time factor. Model answers are usually produced that provide keywords, phrases or concepts that are assigned a portion of the overall answer grade, to reduce marking or grading times and provide uniformity of grading.

Methods: Creation of MCQs and other questions

The process by which academics create a series of questions for assessment tasks relates to the constructive alignment of learning outcomes, content, and assessment tools (Biggs, 2003). While there are many updates to the work by Biggs, the fundamentals hold true that there needs to be consistency (or alignment) between content, objectives and assessment. Leaving aside any vigorous debate on the theoretical aspects of Biggs' work, educators essentially want to assess the right things, in the right way.

The following is the practice of many academics in the Anatomy and Physiology discipline area for generating questions.

- Divide the content into subsections and determine roughly the number of points and questions in each section

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

- Determine the make-up of MCQ, short and long answer questions required,
- Determine if a pool of questions is to be used for each MCQ subsection and how many questions in the overall pool (e.g., 5 randomly selected from the pool of 8),
- Write the questions,
- Review the question for accuracy,
- Moderate the questions by independent colleague for quality assurance,
- Deploy the assessment,

The author compared writing questions from scratch with using a reasonably simple prompt to ChatGPT, an open gen AI tool (<https://openai.com/blog/chatgpt>). With ChatGPT and many open gen AI tools, the interrogator/user enters a prompt in plain English such as a question or statement into a dialog box and hits the enter key.

Results:

The questions and answers presented in this paper relate to the urinary or renal system and are compared back to the author's experience of creating similar questions.

Multiple Choice Questions (MCQ)

A prompt was entered into ChatGPT (as seen in the screen shot forming figure 1). This also shows the first part of the output from ChatGPT. All the multiple-choice questions and answers are provided in Table 2.

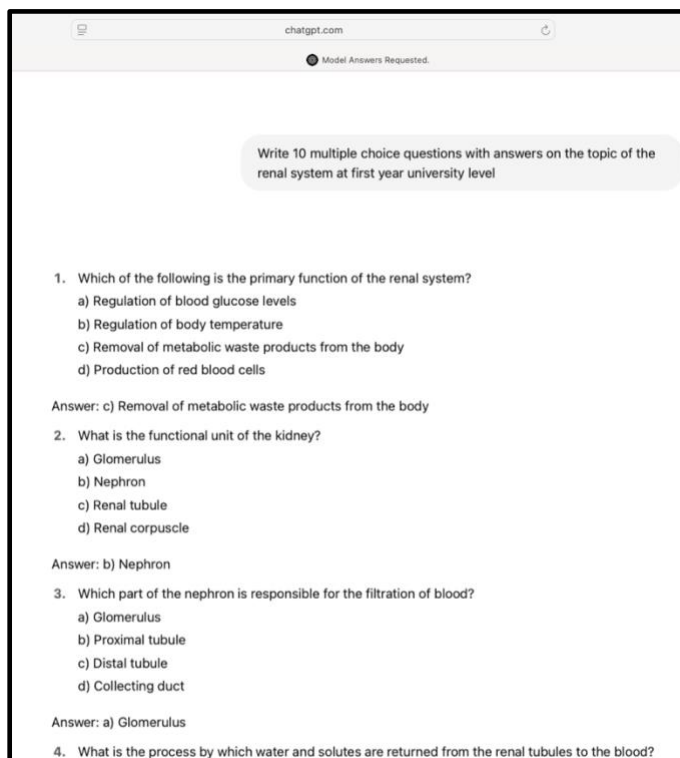


Figure 1. ChatGPT prompt for generating MCQ questions with tagged answers. Note that questions 4–10 are not shown in this screenshot.

The complete set of example questions with answers are detailed in Table 2. When checked, these questions are quite reasonable for this level of understanding and knowledge of the renal system, and the provided answers are correct for the context and the alternatives given.

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Table 2.

Output of MCQs for renal system generated by ChatGPT (note the italicised option is the correct answer).

1. Which of the following is the primary function of the renal system? a) Regulation of blood glucose levels b) Regulation of body temperature c) <i>Removal of metabolic waste products from the body</i> d) Production of red blood cells	2. Which of the following is a condition in which the kidneys fail to remove waste products from the blood? a) Hypertension b) Glomerulonephritis c) <i>Chronic kidney disease</i> d) Kidney stones
3. What is the functional unit of the kidney? a) Glomerulus b) <i>Nephron</i> c) Renal tubule d) Renal corpuscle	4. What is the name of the condition in which protein is present in the urine? a) <i>Proteinuria</i> b) Hematuria c) Pyuria d) Glycosuria
5. Which part of the nephron is responsible for the filtration of blood? a) <i>Glomerulus</i> b) Proximal tubule c) Distal tubule d) Collecting duct	6. Which of the following is a common symptom of kidney stones? a) Hematuria b) Proteinuria c) Hypertension d) <i>Abdominal pain</i>
7. What is the process by which water and solutes are returned from the renal tubules to the blood? a) Filtration b) <i>Reabsorption</i> c) Secretion d) Excretion	8. Which of the following is a condition in which the kidneys produce too much urine? a) Oliguria b) Anuria c) <i>Polyuria</i> d) Nocturia
9. Which hormone stimulates the kidneys to reabsorb more water and produce a more concentrated urine? a) Aldosterone b) <i>Antidiuretic hormone (ADH)</i> c) Atrial natriuretic peptide (ANP) d) Renin	10. Which of the following is a test that can be used to measure kidney function? a) Electroencephalogram (EEG) b) Electrocardiogram (ECG) c) <i>Blood urea nitrogen (BUN) test</i> d) Magnetic resonance imaging (MRI)

If the output is different to what is desired, then there is an option to regenerate response. For more serious reworking of the questions, an alternative is to modify the prompt and resubmit, in a similar way that you would make a general database search more specific to target your research question. If the generated questions are ambiguous or unclear, then minor editing can fix these issues.

The time saved here is summarized in Table 3 and is substantial for the simple generation of 10 questions, of the order of 35 minutes. It takes a similar time for ChatGPT to generate 100 questions for a larger pool and hence the savings here would be much greater, where the checking time becomes the limiting factor. Roughly 105 minutes (less than 2 hours) to generate and check the questions using ChatGPT, compared to 500 minutes (just over 8 hours) for manual generation. Note that this does not include mental fatigue in either case!

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Table 3.

Comparison of MCQ Manual and ChatGPT question creation.

Manual	ChatGPT
5 minutes per question to write and check.	1 minute to generate 10 questions 5 minutes to generate 100 questions 1 minute per question to check.
10 questions = 50 minutes.	10 questions = 15 minutes.
100 questions = 500 minutes	100 questions = 105 minutes

Short and Long Answer Questions

When creating short or long answer style questions, a slightly different approach is taken and the questions and answers are generated in a two-pass method, with a more specific prompt to ensure that the content being assessed is adequately covered.

Staying with the example of the renal system, the prompt this time to ChatGPT was:

“Write 5 short questions on the human renal system: first on glomerular filtration rate, second on fluid balance, third on renal anatomy, fourth on general kidney function, and fifth on urinary composition.”

The questions output from ChatGPT were as follows:

1. What is the glomerular filtration rate, and how is it measured?
2. How does the human renal system maintain fluid balance in the body?
3. What are the major structures of the human kidney and what are their functions?
4. What is the primary function of the human kidney, and how does it contribute to overall health?
5. What is the composition of urine, and how is it affected by various factors such as diet, hydration, and disease?

The second prompt to ChatGPT was:

“Provide 50–100-word answers to these generated questions.” These were generated to assist in marking or grading student answers.”

The generated answer to question 3 has been selected for review:

The pathway through the kidney from the renal artery to the ureter involves several segments, including the renal cortex, renal medulla, renal pelvis, and ureter. The renal artery branches into smaller arteries, eventually leading to the afferent arterioles that supply the glomeruli in the renal cortex. Blood is filtered in the glomeruli and then passes through the proximal and distal convoluted tubules, loop of Henle, and collecting ducts, where further processing and reabsorption occurs. The resulting urine is then funneled into the renal pelvis, which connects to the ureter and eventually the bladder.

This answer looks reasonable at an initial view but has an error in the presented order of the anatomical structures – the distal convoluted tubule is after the Loop of Henle. As an aside, naming structures and organs after early anatomists is known as anatomical eponymism (Burdan et al., 2016), and can be confusing as there are then multiple names for the same structure. For example, it is preferable to use nephron loop, instead of loop of Henle. This, however, runs into the danger of disappearing down a rabbit hole (Carroll, 1865) and

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

appearing in a completely different landscape. While interesting, eponyms are beyond the scope of this paper. Issues like this could be easily remedied during the checking phase, along with US versus Australian English variations (e.g., hematuria vs haematuria from MCQ 6).

The timing for the written response questions is probably less critical as these sorts of questions are not as front-loaded as the MCQs and will still require grading time. However, gen AI is capable of easily and quickly generating a variety of questions. This means we can create multiple questions in a question pool, so that students taking an online test or exam will be presented with a random selection from the pool. This should reduce the incidence of question sharing or collaboration for an individual assessment task, particularly for online, timed assessments.

Warnings for use of AI tools.

With information freely available on the web, including gen AI, we must be aware of the potential issues that could arise with the repeated use of secured assessment. These include question security, sharing of important information and intellectual property, integrity of information and potential confabulation by the AI tool (Lucchi, 2023).

Data security is a constant threat with anything that is asked ask of a search engine, and gen AI tools are no different in that respect. Prompts or searches carried out with many proprietary applications, including browsers, are stored for faster access by others searching for the same information as it is often timely, and may in fact list the common and trending searches available by these applications (e.g., <https://trends.google.com>). The author of this paper has only used the gen AI tools by prompting for questions or answers that the tool can provide, without submitting any new material into the system. One could posit that by using their own prompts to generate questions, students could gain similar questions and answers. At this level in Anatomy and Physiology, arguably this is a useful endeavour for students to test their knowledge and understanding of the content and probably should be encouraged!

As any prompts are ultimately harvested and stored, we need to be cautious of entering our own or our institution's intellectual property as this becomes present and searchable information, and could be found by colleagues, competitors, and students. It may also be against institution IT policy, so it is prudent to check your university or institutional policy carefully before embarking on this process.

A further concern is that of fake information, where the gen AI tool provides a response, that is at best, a guess, or at worst, a complete fabrication (to give the impression that the question has been answered). In either case, this is similar to some student exam answers! For example, at the extreme, there have been reports of authors receiving requests for papers that they had not written but were in their discipline area, or fictitious but reasonably titled papers (Májovský et al., 2023). Some other patently ridiculous outputs or responses have been documented by White, (2022).

Another issue is the size of the online knowledge base that is available for gen AI tools to generate the questions and answers. In a more specialised area, the knowledge base will be small with less available material to produce questions. The examples used in this paper (basic first year Anatomy and Physiology) are quite general and at a level where there is a vast amount of information available online so a large variety of questions should be possible. Therefore, a good estimate of the size of the knowledge base in the discipline area and a quick online search can give an order of magnitude sanity check.

Conclusion and where to next

This paper demonstrates that there is a place for generating questions for topics in the discipline of Anatomy and Physiology using gen AI. The author has communicated with academics in other disciplines, and the application there also appears to be possible, particularly where in foundation knowledge-based subjects that are mostly offered as first and second year university subjects. The main application compared for time-saving benefits was the generation of MCQs where this could lead to hours saved and a greater variety of similar level

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

questions generated in a seemingly random fashion. This has been put into practice by the author, but at this stage mainly for the generation of practice quizzes for tutorial and small lecture groups. This paper warrants the investigation for the use of gen AI in creating variable secured assessment especially in foundation-based subjects where there are not many methods of assessing recall knowledge and some higher order thinking ability.

Reference List

- Biggs, J. (2003). Aligning teaching and assessing to course objectives. *Teaching and learning in higher education: New trends and innovations*, 2(4), 13-17.
- Burdan, F., Dworżański, W., Cendrowska-Pinkosz, M., Burdan, M. and Dworżańska, A., (2016). Anatomical eponyms—unloved names in medical terminology. *Folia morphologica*, 75(4), pp.413-438.
- Boyd, C. M., Bakker, A. B., Pignata, S., Winefield, A. H., Gillespie, N., & Stough, C. (2011). A longitudinal test of the job demands-resources model among Australian university academics. *Applied psychology*, 60(1), 112-140. <https://doi.org/10.1111/j.1464-0597.2010.00429.x>
- Carroll, L., (1865). *Alice's Adventures in Wonderland*, London: William Heinemann. <https://www.gutenberg.org/ebooks/28885> (Accessed June 2025).
- Glaser, R., & Silver, E. (1994). Chapter 9: Assessment, testing, and instruction: Retrospect and prospect. *Review of research in education*, 20(1), 393-419. <https://doi.org/10.3102/0091732X020001393>
- Kinman, G. (2001). Pressure points: A review of research on stressors and strains in UK academics. *Educational psychology*, 21(4), 473-492. <https://doi.org/10.1080/01443410120090849>
- Lucchi, N. (2024). ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 15(3), 602-624.
- Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. *Journal of Medical Internet Research*, 25, e46924. <https://doi.org/10.2196/46924>
- McCune, V. (2018). Experienced academics' pedagogical development in higher education: Time, technologies, and conversations. *Oxford Review of Education*, 44(3), 307-321. <https://doi.org/10.1080/03054985.2017.1389712>
- Menzies, H., & Newson, J. (2007). No time to think: Academics' life in the globally wired university. *Time & Society*, 16(1), 83-98. <https://doi.org/10.1177/0961463X07074103>
- Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic medicine*, 65(9), S63-7. <https://doi.org/10.1097/00001888-199009000-00045>
- Ruggiero, D. (2016, April 13). The black market in academic papers – and why it's spooking publishers. *The Conversation*. <https://theconversation.com/the-black-market-in-academic-papers-and-why-its-spooking-publishers-57296>.
- Templer, B., (2004). High-stakes testing at high fees: Notes and queries on the international English proficiency assessment market. *Journal for Critical Education Policy Studies*, 2(1), pp.1-8. Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal*, 42(3), 454-476. <https://doi.org/10.1002/berj.3215>
- White, M. (2023, September 9). The most insane things ChatGPT Has ever done. *Springboard Blog*. <https://www.springboard.com/blog/news/chatgpt-revolution/>

Schier, M. (2025). Smart and safe use of AI tools for setting assessments. In S. Barker, S. Kelly, R. McInnes & S. Dinmore (Eds.), <i>Future Focussed. Educating in an era of continuous change</i> . Proceedings ASCILITE 2025. Adelaide (pp. 1-8) https://doi.org/10.65106/apubs.2025.2621

ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Schier, M. 2025