

Does ChatGPT-Permitted Assessments Help Students Generate Better Answers and Learn More?

Michelle LF Cheong^a and Jean Y-C Chen^b

*School of Computing & Information Systems, Singapore Management University,
80 Stamford Road, Singapore 178902, Singapore*


Keywords: ChatGPT 3.5, Permitted Use, Assessment, Higher Education, Empirical Study.


Abstract: We discuss our methodology and implementation of ChatGPT-permitted assessments for a university-level spreadsheets modelling module. Through our quantitative data analysis, our students rated ChatGPT's answers to be incorrect on average and thus will not help them generate better answers directly, representing low "Perceived usefulness" (PU), while they rated ChatGPT 3.5 with relatively high "Perceived ease of use" (PE). They gave a good "Behavioural intention" (BI) rating indicating that they were motivated to use it in future as they could still learn more about this module by using ChatGPT 3.5. We found that both PU and PE affected BI positively, with PU being the stronger predictor, suggesting that developers should focus on improving ChatGPT's accuracy to improve PU, which will in turn have a higher positive impact on BI. Through our qualitative analysis, our students indicated that they could learn positively from ChatGPT 3.5 in terms of getting an initial idea on how to approach the problem, providing a first cut solution, learning the execution steps for complex Excel functions, providing an active learning opportunity through identifying and correcting the mistakes, and gaining the awareness of not committing such mistakes in the future.

1 INTRODUCTION

Since the launch of ChatGPT 3.5 in November 2022, many stakeholders including students, instructors, and institutions were amazed by its mostly accurate human-like responses in a myriad of domain areas including medical (Kung et al., 2023), journalism (Pavlik, 2023) and programming (Anagnostopoulos, 2023). However, at the same time, many have expressed concerns of academic integrity when students submit ChatGPT responses as their own, and excessive dependence may erode students' writing and critical thinking skills (Lim et al., 2023; O'Connor and ChatGPT, 2023; Stokel-Walker, 2022). In response, institutions around the world rushed to establish guidelines and policies on how to handle this sudden "invasion" of GAI tools to soften its negative impact on compromised education quality and how to exploit it to achieve positive impact on education (Haleem et al., 2022; Moorhouse et al., 2023).

Chan (2023) conducted a survey with Hongkong universities and proposed an AI Ecological Education Policy Framework which is organized into three dimensions: Pedagogical, Governance, and Operational. In terms of pedagogical dimension, they recommended "teachers to design assessments that allow AI technologies to enhance learning outcomes, rather than solely producing outputs" and to "focus on students' understanding, critical thinking, and analysis to prevent AI-generated content from compromising the assessment process". Moorhouse et al. (2023) examined the guidelines of 23 of the top 50 universities and they covered three common areas: academic integrity, advice on assessment design, and communication with students. For advice on assessment design, they include testing the assessments using GAI tools to understand the abilities and limitations, redesigning assessment tasks, focus on process and staged assessment design, incorporating GAI tools in the assessment process, and use them during in-class assessments.

^a  <https://orcid.org/0000-0002-3192-3452>

^b  <https://orcid.org/0000-0003-2338-4265>

The idea to incorporate GAI tools in the assessment process where GAI tool usage is permitted to aid with assessment completion is a fairly new approach. To do so will require careful design of the assessment, including the implementation process detailing what the students need to do and allowed to do when using GAI tools, how to separate GAI versus student generated answers, and therefore what will be graded and what will not. In this paper, we proposed a methodology to incorporate ChatGPT permitted use in assessment completion for a university-level spreadsheets modelling module. In this methodology, our students used ChatGPT 3.5 as an assistant to complete their assignments and created two separate spreadsheet models for each question, one based purely on answers suggested by ChatGPT and a second improved version based on students' own suggestions for improvements, and we determined if ChatGPT helped them generate better answers and whether they learn more from using it, through analysing the quantitative and qualitative survey data collected.

2 LITERATURE REVIEW

To address the concerns of academic integrity and dishonesty, and decline in students' writing and critical thinking skills, many articles (Cheong, 2023; Lim et al., 2023; O'Connor and ChatGPT, 2023; Zhai, 2022) suggested instructors to redesign assessments to focus on higher order thinking skills which are more "GAI-resistant". Instructors are also encouraged to include ChatGPT permitted use as part of the assessment process since GAI will likely become a "reality of today's educational and job landscape" (Halaweh, 2023; Markauskaite et al., 2022; Moorhouse et al., 2023). Specifically, Halaweh (2023) presented arguments in favour of it and proposed five strategies and techniques to ensure responsible and successful implementation of ChatGPT in teaching and research. These include setting clear policy for ChatGPT usage; requiring students to document the steps in the usage including contradictory findings, judgments and improvements made, and complete audit trails of questions asked, and corresponding answers received; use AI detector tools to inform contents similarity and plagiarism; and finally, instructor and student to swap roles to evaluate the authenticity of learning and critical thinking.

However, there are fewer articles on actual implementation of redesigned assessments and GAI-incorporated assessments with students. In French et al., (2023), the authors also commented that there are

many commentaries on advantages and limitations of GAI in education and proposed recommendations, but there are "few examples of actual practice with students". They conducted a study to integrate ChatGPT and Dall-E into a research and development assignment with their games programming students. Their students evaluated the tools in the context of game development, demonstrated working prototypes integrating ChatGPT and Dall-E, and reported on their findings. They discussed five student outputs in detail, highlighting the students' learnings, frustrations they had, and what ChatGPT and Dall-E can and cannot do well in games development. However, no statistical data analysis of student feedback was performed.

Another actual implementation is by Polasik (2023) who asked her students to use ChatGPT in their learning process in an introductory materials science course. These include getting the students to ask ChatGPT to generate a list of conceptual questions to ask among themselves to discover gaps in understanding, ask ChatGPT to explain the main themes in their own reports, and ask ChatGPT to write MATLAB scripts to accomplish small tasks. Again, no statistical data analysis of student feedback was performed.

Ngo (2023) conducted a survey with 200 students, and a semi-structured interview with 30 students to investigate how Vietnamese university students perceived the use of ChatGPT in education and provided recommendations to overcome the usage challenges. Descriptive statistics and one-sample t-tests were conducted, and it was found that students' positive perception of use and educational advantage were both higher than average. The students also had above average awareness of the challenges associated with the use of ChatGPT. It is worthwhile to note that the survey results were related to general usage of ChatGPT in education, and not specifically in assessment completion.

Our work contributes towards the limited empirical research in actual implementation of ChatGPT-permitted assessments in higher education. Our work is different from earlier works in three aspects. Firstly, our work was based on assessments in a spreadsheets modelling module where students will build mathematical models and perform complex calculations to answer business questions. To our best knowledge, we could only find one recent work by Cheong (2023) that discussed the performance of ChatGPT 3.5 on spreadsheets modelling assessments, to determine how well ChatGPT 3.5 can solve questions of different cognitive levels based on the revised Bloom's Taxonomy.

Secondly, as suggested by Moorhouse et al., (2023), “allowing or even requiring students to use GAI at various stages of the assessment process would, in fact, enhance the authenticity of assessments”. We required our students to use ChatGPT 3.5 to generate answers to the assignment questions and prepare a complete audit trail of the questions asked and answers received as recommended by Halaweh (2023). In addition, students must identify and document down mistakes committed by ChatGPT 3.5, following the recommendation by Chan (2023) to use “assessments and activities where students can by themselves discover the limits of such techniques”, and by Halaweh (2023) to get students to document down the contradictions. After the mistakes were identified, the students must generate an improved version of the answers based on their own suggestions for final submission and grading. Our proposed methodology of incorporating ChatGPT as an assistant in assessment completion with full audit trail and identification of mistakes and learnings at each step, and then create an improved solution model, presents an active learning model where students build the ability to critically judge the quality of responses generated by ChatGPT to develop evaluative judgement skill (Bearman et al., 2024).

Thirdly, our actual implementation allowed us to collect both quantitative and qualitative survey data from our students on the entire assessment completion process, to assess the efficacy of using ChatGPT 3.5 in assisting them to complete their assignments and determine what they learn from using it. With our data analysis, we aim to answer our research questions:

- RQ1: Does the usage of ChatGPT 3.5 help students to generate better answers for the assignment questions?
- RQ2: Does the usage of ChatGPT 3.5 help students to learn more in this module?
- RQ3: What are the learnings in using ChatGPT 3.5 to complete the assignment questions?

3 MODULE AND ASSESSMENT

3.1 Spreadsheets Modelling Module

The spreadsheets modelling module teaches students how to translate business problems into mathematical representations, and to build the spreadsheet models from scratch to perform mathematical calculations and data analysis to obtain insights for decision making. The module covers six main topics, namely:

- (i) basic modeling techniques, (ii) spreadsheets engineering skills, (iii) financial calculations, (iv) data lookups and optimization, (v) Monte Carlo simulation, and (vi) time-based discrete event simulation.

3.2 Assessment Design and Implementation

We designed two take-home assignment questions (denoted as AQ1 and AQ2) for the fall term of academic year 2023/24. As the assignment questions were original and did not exist before September 2021, they will not form part of the ChatGPT’s training corpus, hopefully making them more “GAI-resistant”. Each question described a business scenario and contained multiple consecutively linked parts for end-to-end analysis.

For each question, students will complete it following the steps below as depicted in Figure 1.

- 1) Feed the question part by part as prompts to ChatGPT 3.5 in a continuous conversation, to obtain suggested answers to each part. Create the spreadsheet model (Model A) according to the suggested answers. No further prompting permitted to solicit any improved responses by ChatGPT 3.5, as students were supposed to identify any mistakes and suggest improvements themselves.
- 2) Base on Model A created in step 1, take note of areas which the student thinks that the answer could be wrong, and areas which student thinks he/she would learn positively from the answer. Model A will not be graded due to differing responses of varied quality from ChatGPT 3.5.
- 3) Fill out the Questionnaire Form provided to document the audit trail for step 1, and the mistakes and learnings for step 2.
- 4) Create an improved version of the spreadsheet model (Model B) based on students’ own suggestions for improvements. Model B is submitted for final grading.

3.3 Questionnaire Design

The questionnaire was designed with three portions: audit trail, qualitative questions, and quantitative questions. One questionnaire was designed for each assignment question, denoted as Questionnaire 1 for AQ1 and Questionnaire 2 for AQ2.

- 1) For the audit trail portion, students will input the question as prompt into ChatGPT 3.5 and copy and paste the response, part by part.

- 2) For the qualitative questions portion, the students were asked to describe the wrong answers provided by ChatGPT 3.5, and which area(s) they have learnt positively from the responses provided.
- 3) For the quantitative questions portion, the students were asked to rate how good the ChatGPT's answers were, how much ChatGPT 3.5 enhanced their ability to generate better answers, how easy to use and understand ChatGPT's answers, did they learn more and will they be engaged to use it for future learning.

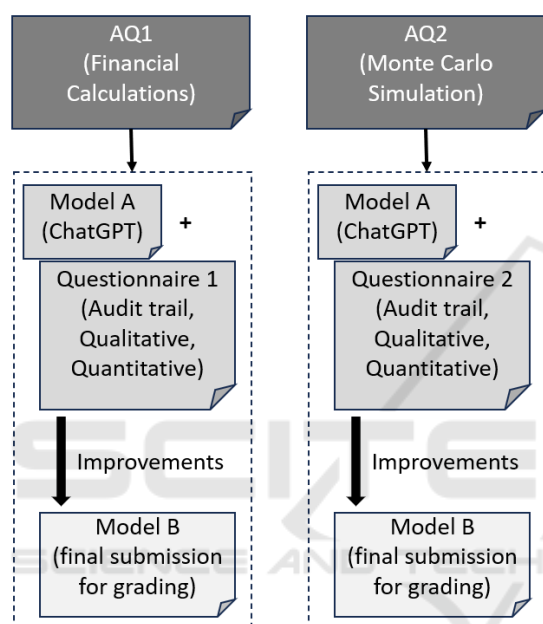


Figure 1: Assessment design and implementation.

3.4 Data Collection

In this study, the participants were master level students pursuing their degree in data science. A total of 146 students took the module and completed the two assignment questions according to the implementation process described in Section 3.2. Consent for data collection and analysis in the study were completely voluntary, and 138 students consented according to IRB approval process. For students who did not consent, their Model A and Questionnaires for both AQ1 and AQ2 were excluded from the study data set.

Students who gave consent initially were permitted to withdraw their consent at any point in time by contacting the Research Assistant who was not involved in the module delivery. The students' decisions to take part in the study would in no way impact their learning and final grades. The module

instructor was the Principal Investigator of the study would know which student consented or not, only until after the grades for the module were finalized and released to ensure no biasness.

4 DATA ANALYSIS AND RESULTS

For quantitative data analysis, we referred to past works by other authors who performed statistical analysis on students' acceptance and use of ChatGPT in education, based on the Technology Acceptance Model (TAM). In Lai et al. (2023), they investigated "Intrinsic motivation" and factors that would influence ChatGPT acceptance for active learning among Hongkong undergraduate students. They found that "Intrinsic motivation" was the strongest motivator, and "Perceived usefulness" was a strong predictor for ChatGPT use. However, there was no significant relationship between "Perceived ease of use" and "Behavioural intention". Both "Perceived usefulness" and "Perceived ease of use" were not significant mediators in the relationship between "Intrinsic motivation" and "Behavioural intent". Thus, to increase students' acceptance to use ChatGPT, developers should spend more effort to improve the subjective experience with injected humour and empathy to increase "Intrinsic motivation", and to improve the accuracy of the answers to increase "Perceived usefulness", instead of spending effort to improve the "Perceived ease of use".

Strzelecki (2023) investigated Polish students' acceptance and use of ChatGPT in higher education by adopting a modified version of the Unified Theory of Acceptance and Use of Technology. They considered seven predictors and found that "Habit", "Performance expectancy" and "Hedonic motivation" were the top three strongest predictors of "Behavioural intention". This finding was similar to Lai et al. (2023) where "Performance expectancy" can be mapped to "Perceived usefulness", while "Hedonic motivation" mapped to "Intrinsic motivation".

Duong et al. (2023) also adopted the Unified Theory of Acceptance and Use of Technology to investigate the how the two predictors, "Effort expectancy" and "Performance expectancy", affected students' intention to use and actual use of ChatGPT for their learning. They found that both "Effort expectancy" and "Performance expectancy" had direct positive impact on students' intention to use

ChatGPT, which in turn promoted them to use it. Also, the higher the incongruence between the two predictors, the lower would be the intention and actual use. Their findings differed from both Lai et al. (2023) and Strzelecki (2023), where “Effort expectancy” or equivalently “Perceived ease of use” was not found to be a strong predictor of intention.

We adopted a similar approach using two factors, “Perceived usefulness” (PU) and “Perceived ease of use” (PE), and evaluated their impact on “Behavioural intention” (BI) by mapping to our quantitative survey questions for data analysis. In addition, we performed qualitative data analysis to understand what our students learn from ChatGPT 3.5 in assessment completion. To better understand the data analysis results, the brief descriptions of the two questions are provided below:

- 1) AQ1: Financial Calculations. This question has four parts (a to d) to test the students’ ability to build models to compute present value, periodic payments, net cashflow, and return rate; to use iterative method to determine new dividend values based on a desired return rate; and to use Data Table to compute the present values for different return rates.
- 2) AQ2: Monte Carlo Simulation. This question has three parts (a to c) to test the students’ ability to build a simulation model using probability distributions and random generators; to simulate a boat race between two race teams; and to repeat the same simulation using Data Table for 30 races to determine the winning probability for one of the teams.

4.1 Quantitative Data Analysis

We collected students’ ratings on six quantitative survey questions, using rating scale of 1 (low) to 10 (high) for survey questions 1 and 2, and rating scale of 1 (Strongly disagree) to 5 (Strongly agree) for survey questions 3 to 6, as given in Table 1. For survey questions 1 and 2, we have used a 10-point scale as they are questions related to the correctness of ChatGPT 3.5’s responses and whether students generated better answers, to capture a more granular level of responses. While 138 students consented to data collection and analysis, only 126 valid responses ($n = 126$) were included for each questionnaire due to missing data.

Table 1: Six quantitative survey questions.

	Description	Factor
1	Rate ChatGPT 3.5’s performance in generating the correct answer for each part. (a to d) for AQ1, (a to c) for AQ2.	PU
2	Rate how much does the usage of ChatGPT 3.5 enhance their ability to generate better answer for each part. (a to d) for AQ1, (a to c) for AQ2.	PU
3	Rate if ChatGPT 3.5 is easy to use.	PE
4	Rate if it is easy to understand the suggested answers generated by ChatGPT 3.5.	PE
5	Rate if the student can learn more about this module by using ChatGPT 3.5 in the learning process.	BI
6	Rate if the student is engaged and motivated to use ChatGPT 3.5 in future learning in this module.	BI

We computed Cronbach’s alpha (Taber, 2018) to determine our students’ overall consistency in terms of their responses across all three factors for each questionnaire, and obtained the value of 0.8724 for Questionnaire 1, and 0.8389 for Questionnaire 2, indicating that both are in the good range. As the number of items in the questionnaire was different for each factor, we also computed Spearman-Brown Coefficient (Eisinga et al., 2013) for each factor for both questionnaires in Table 2. Most of the coefficients can be interpreted as excellent, except for the factor PE which were Fair for Questionnaire 1, and Good for Questionnaire 2. Overall, these values indicated that our students’ responses were consistent, and we could proceed to analyse their responses to gain insights.

Table 2: Spearman-Brown Coefficients.

Factor	Spearman-Brown Coeff	Interpretation
Q1: PU	0.940	Excellent
Q1: PE	0.625	Fair
Q1: BI	0.917	Excellent
Q2: PU	0.912	Excellent
Q2: PE	0.737	Good
Q2: BI	0.917	Excellent

We performed one-tail test for mean scores above 5.0 (out of 10) for questions 1 and 2, and mean scores above 2.5 (out of 5) for questions 3 to 6, to obtain the p-value to determine significance.

For Questionnaire 1, the mean scores for Questions 1a to 1d, and 2a to 2d given in Table 3 were mostly below the average of 5.0, except for Question 1a with mean score of 5.238 but was found to be

insignificant with p-value of 0.15. This indicates that on average, students did not rate ChatGPT's answers to be correct and thus did not help them generate better answers directly, representing low "Perceived usefulness" (PU). For questions 3 and 4, the mean scores were 3.905 and 2.968 respectively, and both had very low p-values. This indicates that on average, students rated ChatGPT 3.5 with relatively high "Perceived ease of use" (PE). Finally, for questions 5 and 6, the mean scores were 2.984 and 3.056 respectively, and both with very low p-values. This indicates that on average, students felt that they can learn more from ChatGPT 3.5 and were engaged and motivated to use it for future learning in this module, representing good "Behavioural intention" (BI).

Table 3: Questionnaire 1 ratings.

	Factor	Mean score	SD	p-value
1a	PU	5.238	2.559	0.15
1b	PU	3.579	2.405	-
1c	PU	3.413	2.338	-
1d	PU	3.127	2.257	-
2a	PU	4.865	2.521	-
2b	PU	3.690	2.425	-
2c	PU	3.500	2.510	-
2d	PU	3.317	2.493	-
3	PE	3.905	0.971	9.89×10^{-33}
4	PE	2.968	1.069	1.46×10^{-6}
5	BI	2.984	1.016	2.22×10^{-7}
6	BI	3.056	1.086	3.73×10^{-8}

For Questionnaire 2, the results given in Table 4 were similar to that of Questionnaire 1, with a slight difference for question 1a where the mean score of 5.794 was found to be significant. The overall results for Questionnaire 2 were low PU, relatively high PE, and good BI, identical to that of Questionnaire 1.

Table 4: Questionnaire 2 ratings.

	Factor	Mean score	SD	p-value
1a	PU	5.794	3.252	0.0036
1b	PU	3.548	2.080	-
1c	PU	2.587	1.912	-
2a	PU	4.841	3.030	-
2b	PU	3.579	2.220	-
2c	PU	2.563	2.026	-
3	PE	3.563	1.137	4.58×10^{-19}
4	PE	2.921	0.997	3.14×10^{-6}
5	BI	2.897	1.060	2.67×10^{-5}
6	BI	2.944	1.064	3.81×10^{-6}

This is an interesting finding, as it implies that while our students did not find ChatGPT's answers accurate and useful most of the time, they were still able to learn something from it and were motivated to use it in the future. This could be due to a few possible reasons. One, the relatively high PE drives BI positively; two, the process of identifying and correcting the mistakes enhance learning and students found such active learning useful for them; and three, the sheer intrinsic motivation to use new and exciting tools.

We performed a multiple linear regression to understand the impact of PU and PE on BI, by combining the survey responses of both questionnaires. The relationship obtained was $BI = 0.5261*PU + 0.2463*PE + 0.2294$, and the p-values were all very low and thus the coefficients and constant were all significant, as given in Table 5. Our results were similar to that of Duong et al. (2023) where both PE and PU had positive impact on BI. In addition, PU was a stronger predictor than PE, suggesting that if the developers can focus on improving ChatGPT's accuracy to improve PU, it will in turn have a higher positive impact on BI, which was also recommended by Lai et al. (2023).

Table 5: Multiple linear regression of PU and PE on BI.

	Coeff	Std Error	t	p> t
Constant	0.2294	0.070	3.275	0.001
PU	0.5261	0.102	5.135	0.000
PE	0.2463	0.101	2.444	0.016

4.2 Qualitative Data Analysis

Based on our students' qualitative descriptions of the wrong answers and the areas which they have learnt positively from the responses provided by ChatGPT 3.5, we performed text mining and created word cloud and topic modelling using Latent Dirichlet Allocation (LDA) for each part of AQ1 and AQ2.

For example, for AQ1 part a, the question required the students to compute the present value worth of an investment which pays dividends at the end of 5th, 6th, 7th, 8th and 9th year, using a return rate of 5%. The word cloud (Figure 2) for the feedback on the wrong answers provided by ChatGPT 3.5 highlighted that it understood the year wrongly as it misinterpreted the information that no dividends will be returned in the first four years and assumed year 5 to be year 1 thus leading to wrong values entered, plus the suggested Excel model had confusing cell locations.

The corresponding topic model is given as $0.053*year + 0.028*cell + 0.025*value$, representing

the most committed mistake by ChatGPT 3.5 involving year, cell and value. Specific student comments include:

“Did not understand that the dividend is in year 5-9, so the calculations are all incorrect.”

“Wrong location of year cells which led to wrong input of dividends.”

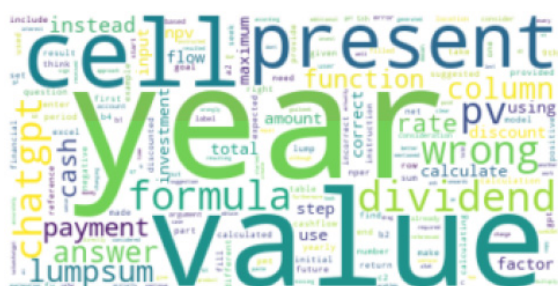


Figure 2: Word cloud for wrong answers to AQ1 part a.

Using another example, in AQ2 part b, the question required the students to create an Excel model to simulate a boat race to determine the winner between two fictitious teams, Oxfort and Cambrick, using random generators and probability distribution functions. The word cloud (Figure 3) for the feedback on the wrong answers provided by ChatGPT 3.5 highlighted that it was not able to apply the correct formula to perform Monte Carlo simulation, especially for team Oxfort, and thus unable to determine the winner.



Figure 3: Word cloud for wrong answers to AQ2 part b.

The corresponding topic model is given as $0.042 \times \text{formula} + 0.018 \times \text{winner} + 0.018 \times \text{oxfort}$, representing the most committed mistake by ChatGPT 3.5. Specific student comments include:

“The formula to calculate Oxfort position for each minute did not work.”

“Winner of the race is not calculated properly, and the winner column is entirely wrong.”

From these two examples, we can see that ChatGPT 3.5 can misinterpret the information

provided in the question and was unable to apply the correct formula to perform some calculations. Such an outcome was in fact an expected one, given that ChatGPT 3.5 is primarily a language model and does not fare as well in mathematical calculations, even more so in spreadsheets modelling. This phenomenon was similarly highlighted in Cheong (2023) particularly for questions of higher cognitive levels.

Albeit these mistakes committed by ChatGPT 3.5, the students provided comments on what they have learnt positively from it. For example, for AQ1 part c, the question required students to use iterative method to determine the new dividend value for a desired return rate. The word cloud (Figure 4) shows that students learnt how to use Goal Seek to perform the iterative calculations following the steps suggested by ChatGPT 3.5 to find the answer.

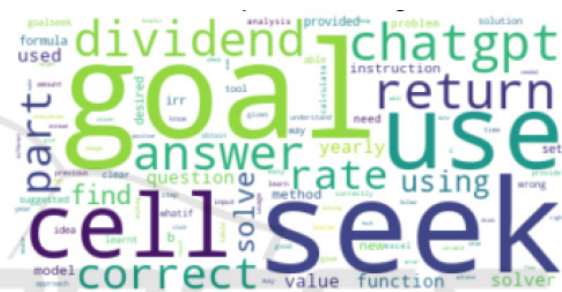


Figure 4: Word cloud for positive learning to AQ1 part c.

The corresponding topic model is given as $0.061 \times \text{seek} + 0.059 \times \text{goal} + 0.035 \times \text{use}$, representing the positive learning from ChatGPT 3.5. Specific student comments include:

“Using of the goal seek function from the instructions provided by GPT.”

“ChatGPT 3.5 was correct in pointing out that the Goal Seek function can be used to find the dividend amount based on the return rate of 8%.”

Quoting another example, in AQ2 part c, the question required the student to use Data Table to manage the simulation of 30 such races and compute the winning probability of team Oxfort from the 30 simulated races. The word cloud (Figure 5) shows that students learnt how to use Data Table to repeat the simulation 30 times following the steps suggested by ChatGPT 3.5.

The corresponding topic model is given as $0.065 \times \text{table} + 0.064 \times \text{data} + 0.029 \times \text{use}$, representing the positive learning from ChatGPT 3.5. Specific student comments include:

“How to navigate Data Table. The instructions were very clear.”

“It realizes that it needs to set up the format of a data table and it gives the steps very accurately just

that it does not specify which cell or which formula, so it gives a high level idea but it may not be totally accurate.”

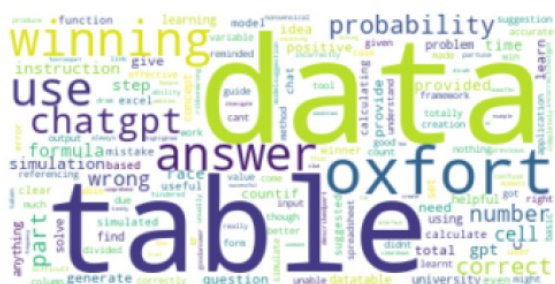


Figure 5: Word cloud for positive learning to AQ2 part c.

Such an outcome was similarly highlighted in Cheong (2023) that ChatGPT 3.5 was very capable in explaining how to use Excel functions correctly by providing clear step-by-step instructions, while it may not be able apply them correctly all the time.

However, not all students provided comments on positive learning from ChatGPT 3.5. Some students explicitly stated that they learnt nothing from ChatGPT 3.5 (6.0% for AQ1 and 15.4% for AQ2). If we include students who indicated “NA” for their responses, we obtained 21.2% for AQ1 and 32.6% for AQ2. These percentages were not high, in comparison with students who explicitly described their positive learnings.

4.3 Discussion of Overall Results

Overall, our students’ responses highlighted that while ChatGPT 3.5 was very easy to use (high PE), the quality of its generated responses to the spreadsheets modelling questions were low (low PU). However, the students were still motivated to use ChatGPT 3.5 in future for learning this module (good BI), as there were some “nuggets of wisdom” which they could glean from its responses. Some of the positive students’ comments include:

“Provided some ideas of how to approach the problem.”

“Give me the framework on how to solve the problem, although there are different corrections that need to be modified to get the correct answer.”

“ChatGPT’s model gave me a starting point to structure the model, as well as alerted me to mistakes/inputs that I should not make for the calculations to be valid.”

With our data analysis results, we attempt to answer our research questions. For RQ1, since the quality of its generated responses to the spreadsheets

modelling questions were low (low PU), ChatGPT 3.5 will not be able to help students generate better answers directly. However, it does provide an initial idea of how to approach the question and provide a first cut solution for the students to improve upon, which will assist the students indirectly.

For RQ2, our students felt that they can learn more from ChatGPT 3.5 and were engaged and motivated to use it for future learning (good BI). For RQ3, the learnings from ChatGPT 3.5 include the execution steps for complex Excel functions, active learning through identifying and correcting the mistakes committed by ChatGPT 3.5, and gaining the awareness of not committing such mistakes in the future. Such learnings form part of developing evaluative judgement skill in the students which aligns with the learning outcomes of the module.

5 CONCLUSIONS AND FUTURE WORK

It is expected that the quality of the GAI-generated responses will improve over time, and GAI tools will become embedded in common software tools such as MS office tools. Allowing and incorporating GAI tools usage in educational assessments will become inevitable. To be better prepared, instructors must engage in purposeful assessment designs that allow students to exploit the GAI tools to generate better solutions to real world problems, and at the same be cognizant about the strengths and weaknesses of such tools, and to use them responsibly.

Our methodology to incorporate ChatGPT in assessment completion with complete audit trails, identifying and documenting the mistakes made and positive learnings, and suggesting improvements, allows students to learn actively from the entire process developing better evaluative judgment skill and achieving better learning outcomes. Other instructors can adopt our methodology in their own course assessments to obtain their specific findings.

The proposed future work that should follow immediately would be to identify and quantify students’ skills, abilities and knowledge gain in using GAI tools in learning, by comparing with a control group that does not use GAI tool. In addition, we can also collect and analyse data on students’ and instructors’ perception of whether using such tools in learning will indeed result in the decline of students’ critical thinking skills or increase students’ reliance on these tools. The results will serve to validate some of the common concerns that were raised in many articles.

For course instructors, one future work would be to establish a methodology on testing their assessments using GAI tools to understand the abilities and limitations to inform assessment redesign to construct more effective assessments that can develop the desired skills for the next-generation workforce.

REFERENCES

- Anagnostopoulos, C-N. (2023). ChatGPT impacts in programming education: A recent literature overview that debates ChatGPT responses. *F1000Research* 2023. <https://doi.org/10.12688/f1000research.141958.1>
- Bearman, M., Tai, J., Dawson, P., Boud, D., Ajjaw, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 49:6, 893-905. <https://doi.org/10.1080/02602938.2024.2335321>
- Chan, C.Y.K. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20:38. <https://doi.org/10.1186/s41239-023-00408-3>
- Cheong, M. (2023). ChatGPT's Performance in Spreadsheets Modelling Assessments based on Revised Bloom's Taxonomy. In *Proceedings of the 31st International Conference on Computers in Education (ICCE 2023)*. Asia-Pacific Society for Computers in Education.
- Duong, C.D., Bui, D.T., Pham, H.T., Vu, A.T., Nguyen, V.H. (2023). How effort expectancy and performance expectancy interact to trigger higher education students' uses of ChatGPT for learning. *Interactive Technology and Smart Education*, 1741-5659. <https://doi.org/10.1108/ITSE-05-2023-0096>
- Eisinga, R., Grotenhuis, M., Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58(4), 637-642. <https://doi.org/10.1007/s00038-012-0416-3>
- French, F., Levi, D., Maczo, C., Simonaityle, A., Triantafyllidis, S., Varda, G. (2023). Creative Use of OpenAI in Education: Case Studies from Game Development. *Multi-modal Technologies and Interaction*. 7, 81.
- Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, 15(2), ep421. <https://doi.org/10.30935/cedtech/13036>
- Haleem, A., Javaid, M., Singh, R.P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 2:100089. <https://doi.org/10.1016/j.tbench.2023.100089>
- Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*, 2(2): e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
- Lai, C.Y., Cheung, K.Y., Chan, C.S. (2023). Exploring the role of intrinsic motivation in ChatGPT adoption to support active learning: An extension of the technology acceptance model. *Computers and Education: Artificial Intelligence*, 5:100178. <https://doi.org/10.1016/j.caeai.2023.100178>
- Lim, W.M., Gunasekara, A., Pallant, J.L., Pallant, J.I., Pechenkina, E. (2023). Generative AI and the future of education: Ragnarok or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21, 100790. <https://doi.org/10.1016/j.ijme.2023.100790>
- Markauskaite L., Marrone R., Poquet O., Knight S., Martinez-Maldonado R., Howard S., Tondeur J., De Laat M., Buckingham S.S., Gasevic D., Siemens G. (2022). Rethinking the entwinement between artificial intelligence and human learning: what capabilities do learners need for a world with AI? *Computers & Education: Artificial Intelligence*. 3:100056. <https://doi.org/10.1016/j.caeai.2022.100056>
- Moorhouse, B.L., Yeo, M.A., Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the top-ranking universities. *Computers and Education Open* 5:100151. <https://doi.org/10.1016/j.caeo.2023.100151>
- Ngo, T.T.A. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning (iJET)*, 18(17), 4–19. <https://doi.org/10.3991/ijet.v18i17.39019>
- O'Connor, S., ChatGPT. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, 66.
- Pavlik, J.V. (2023). Collaborating with ChatGPT: Considering the implications of Generative Artificial Intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78(1), 84-93. <https://doi.org/10.1177/10776958221149577>
- Polasik, A.K. (2023). Hey ChatGPT: Can you help me learn? *The Journal of The Minerals, Metals & Materials Society (TMS)*, 75(7), 2089-2090. <https://doi.org/10.1007/s11837-023-05924-1>
- Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays — should professors worry? *Nature*. <https://doi.org/10.1038/d41586-022-04397-7>
- Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2023.2209881>
- Taber, K.S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in Science education. *Research in Science Education*, Vol 48, 1273-1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Zhai, X. (2022). ChatGPT user experience: Implications for education. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4312418.