Vol.6 No.1 (2023)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

Content Available at : http://journals.sfu.ca/jalt/index.php/jalt/index

Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance

Abdolvahab Khademi[A]

[A] Postdoctoral Fellow, University of Maryland, Baltimore, USA

Keywords

Artificial intelligence;
automated item generation;
ChatGPT;
educational technology;
Google Bard;
Large Language Models (LLMs);
natural language processing (NLP).

Correspondence

vahab.khademi@gmail.com[A]

Article Info

Received 2 May 2023
Received in revised form 10 May 2023
Accepted 10 May 2023
Available online 10 May 2023

DOI: https://doi.org/10.37074/jalt.2023.6.1.28

Abstract

ChatGPT and Bard are AI chatbots based on Large Language Models (LLM) that are slated to promise different applications in diverse areas. In education, these AI technologies have been tested for applications in assessment and teaching. In assessment, AI has long been used in automated essay scoring and automated item generation. One psychometric property that these tools must have to assist or replace humans in assessment is high reliability in terms of agreement between AI scores and human raters. In this paper, the reliability of OpenAI's ChatGPT and Google's Bard LLMs tools against experienced and trained humans in perceiving and rating the complexity of writing prompts is measured. Intraclass correlation (ICC) as a performance metric showed that the reliability of both ChatGPT and Bard was low against the gold standard of human ratings.

Introduction

Advances in artificial intelligence (AI) and computing hardware (e.g., graphics processing unit (GPU) and high performance computing) have brought significant progress and power to deep neural network learning and natural language processing (NLP) and their applications. In particular, generative AI has recently increased the power of NLP tools in terms of precision in understanding and predictive power. The public release of ChatGPT (based on generative pretrained transformer, GPT) by OpenAI and Bard (Experiment) by Google took different industry sectors by storm, inasmuch as earning the interest of industry leaders in integrating these tools in daily operations, such as content creation, code generation, mathematical proofs, healthcare analytics (Iftikhar, 2023), calculations, and translation. ChatGPT uses both supervised and reinforcement learning machine learning algorithms. Since the public release of ChatGPT, several studies have investigated its use, benefits, and harms in different endeavors. For example, Pavlik (2023) discusses the benefits and weaknesses of using ChatGPT for text generation in media and journalism. Some studies have shown that ChatGPT performs so well that it can complete some examinations with satisfactory results, such as the bar exam (Choi et al., 2023; Katz et al., 2023), the United States Medical Licensing Examination (USMLE) (Gilson et al., 2022; Kung et al., 2023) and the GRE, though some have shown otherwise (Huh, 2023). In a study comparing the quality of short essays on physics open-ended questions, Yeadon et al. (2023) report that ChatGPT was able to generate first-grade essays comparable to student essays achieving a very similar mean score. As such, further research is needed to explore the applications, benefits, and potential detriments of advanced AI technologies in different areas, especially in education.

AI tools have long been applied in learning analytics and educational technologies, dating back to the 1970's and researched ever since in academic and industry forums (Rudolph et al., 2023a). In particular, AI tools based on NLP have extensively been used in automated essay scoring (AES) and automated item generation (AIG) in areas such as languages, arts, mathematics, and sciences. AES technologies in educational assessment have enabled educators and education systems to go beyond discrete-choice assessment items through faster and reliable scoring and reporting methods. In this regard, one can categorize AI as an educational technology (Rudolph et al., 2023a; Tate et al., 2023) that can be integrated in the learning process as in intelligent tutoring systems (ITS).

One promising area that AI can be of great assistance to learning and assessment is automatic item generation for summative and particularly formative assessment, especially in self-assessment contexts and personalized learning through continuous feedback into the learning processes (Cope et al., 2021). For instance, language learning applications such as Duolingo provide a self-paced and personalized language learning path for the users, with numerous practice items and quizzes. In addition, with the widespread adoption of computer-based testing (CBT) and online delivery platforms and the need for the development of items at scale, AIG technologies can prove crucial and efficient (Gierl et al., 2021). Writing items for practice and evaluation by human item writers is costly and time-consuming. NLP tools integrated into AIG pipelines can significantly lower the costs in item generation if they are trained to match the performance of human item writers. Because item generation and mapping need to be at the level of the current ability or performance of the learners, NLP tools must be able to recognize the appropriacy of item contents in terms of their difficulty and complexity in accordance with the ability of the user. For instance, in mathematics learning, an NLP-based app must be able to generate mathematics practice items at the level of a fifth grader given the current performance of the learner or the expected learning outcomes. In language education applications, an NLP-based item generator must be able to produce vocabulary, grammar, reading, and writing items that correspond to the language proficiency or the grade level of the learner. If the generated items do not match the appropriate level of the learner, assessment estimates will not be accurate to evaluate the performance of the learner. Hence, the current AI tools must be trained to a degree that they should match a lower bound of human performance.

One metric to ensure the utility of AI tools in education and assessment is the degree of agreement between the AI tools and the human raters on a performance task, such as scoring essays or understanding the appropriacy of item complexity with a perspective on the current proficiency level of the learners. Although numerous studies have been conducted to ensure the reliability of AI tools in automated essay scoring, few studies have reported on the reliability of AI tools for the purpose of generating level-appropriate items. Hence, in the present study, I aim to evaluate the reliability of AI tools in understanding and rating the difficulty or complexity of topics for writing assessment. In particular, I am interested in evaluating the reliability of ChatGPT-3.5 and Bard (Experiment) in their ability to perceive and measure the complexity of writing prompts as an application of AI in automated item generation. I choose OpenAI ChatGPT-3.5 and Google Bard because they are the most well-known LLM-based generative AI tools and have been embraced positively by the general public and scrutinized by researchers. At the time of writing, Bard is in the experimental stage and this paper uses the free experimental version. In addition, I used OpenAI's ChatGPT-3.5 version for the present study.

Method

The present study aims to evaluate the reliability of ChatGPT-3.5 and Bard with regard to their perception and numerical rating of the complexity of writing prompts for writing assignments. Adoption of AI tools in automatic item generation (AIG) requires a reliability as high as the minimum acceptable performance of trained humans in order for the results obtained by the AI tools to be reliable and scalable. Reliability can be defined as the degree of agreement between two or more judges or raters measuring the same trait or object. Such agreement can be quantified through several statistical and mathematical methods, such as Spearman rho correlation, the Cohen's kappa, Kendall's tau, and the intraclass correlation (ICC). In the present study, I use ICC to quantify the degree of agreement among human raters as the benchmark and between the human raters and ChatGPT-3.5 and Bard.

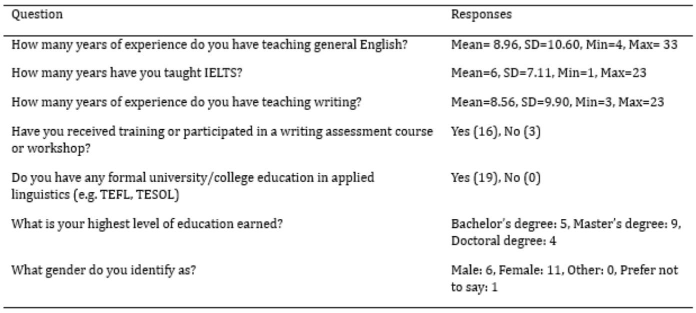
Data

The data were collected through an online questionnaire in which 20 IELTS Academic Writing Task II prompts were randomly selected from the pool of official past examination papers published by the Cambridge University Press in years 1996 to 2022 (except years 2012 and 2014, where the researcher was not able to find published official past examinations). For each administration year, two writing prompts were randomly selected. The selected prompts were placed in an online questionnaire in which the cognitive complexity of each prompt would be measured on a 1-8-point Likert scale by randomly selected human raters. In addition to the 20 writing prompts as the main questionnaire items, the researcher also included several questions about the demographic and professional information and background of the human raters. The questionnaire was designed and administered online through the Qualtrics survey platform. The questionnaire did not include any personally identifying items, and all responders consented to participate in the study. A rating guideline along sample rating was presented to the participants at the beginning of the questionnaire. The human raters in this task were required to rate the complexity of the writing prompts on a scale of 1 to 8 with unit interval, with 1 being the lowest possible complexity score and 8 the maximum. Data from the responses of participants were collected over several days. The questionnaire was not timed.

Human raters

After arranging the 20 randomly selected writing prompts in a questionnaire, participants were sought to rate the prompts in the questionnaire through the Qualtrics survey platform. Participants in this study were invited through an announcement on one professional forum platform (LTEST-L) and several teachers and professional group pages on social media. Participants in this study included 19 professionals with formal education, training, and experience in teaching writing to a diverse population of students. The human raters in this study had on average about nine years of experience teaching English at different proficiency levels. In addition, the human raters had an average of 8.5 years of experience teaching academic and general writing to students. All participants had received formal education in the areas of applied linguistics and additionally 84% of the participants had received extra training in workshops on writing assessment. Participants were educated at the undergraduate (26%), master's (47%), and doctoral (21%) levels in applied linguistics. The demographic and professional data of the human participants are presented in Table 1.

Table 1. Demographic and professional information of the human raters.



Question	Responses
How many years of experience do you have teaching general English?	Mean= 8.96, SD=10.60, Min=4, Max= 33
How many years have you taught IELTS?	Mean=6, SD=7.11, Min=1, Max=23
How many years of experience do you have teaching writing?	Mean=8.56, SD=9.90, Min=3, Max=23
Have you received training or participated in a writing assessment course or workshop?	Yes (16), No (3)
Do you have any formal university/college education in applied linguistics (e.g. TEFL, TESOL)	Yes (19), No (0)
What is your highest level of education earned?	Bachelor's degree: 5, Master's degree: 9, Doctoral degree: 4
What gender do you identify as?	Male: 6, Female: 11, Other: 0, Prefer not to say: 1

Machine raters

The focus of the present study was on the rating performance of artificial intelligence tools. I selected the ChatGPT-3.5 because it is the most referenced AI language model in the public domain and technology forums. In addition, I included Bard as a competitor. I used ChatGPT-3.5 on March 31, 2023 and Bard on April 1, 2023 through free personal sign-up. Performance of the AI tools refers to their latest development on these dates, as these tools are ever-developing and being updated with new training data. Therefore, the results of the study are to be interpreted based on the current versions of these tools at the time of the experiment. ChatGPT-3.5 and Bard both received the writing prompts manually and in the same order but on two different days (one day apart).

Analysis

In this experiment, I asked both the human raters and the AI raters to rate on a 1-8 scale (1= barely complex and 8 = highly complex) the complexity of the presented writing prompts as a writing homework assignment for students. The goal was to compare the performance of ChatGPT-3.5 and Bard as candidate technologies for item generation in writing assessment where prompts are measured for their complexity or difficulty to match the ability or grade level of the learners. The writing prompts in this experiment were randomly selected from IELTS Academic Task II writing components (Cambridge University Press). The 20 randomly selected prompts were placed on a questionnaire and sent via email to human participants to respond on the Qualtrics survey platform. At the beginning of the questionnaire, a written guideline was introduced to explain the purpose of the study and data collection and how to rate a prompt through a sample demonstration. In addition, some questions asked the human raters to provide demographic information, such as experience in assessing writing, education level, and native language. The data was collected over several days. The same writing prompts were manually presented through the dialog box to both ChatGPT-3.5 and Bard in the same order and with the same instruction (the instruction read, "On a scale of 1-8, how complex is this prompt for a student writing assignment homework? The prompt is: [prompt]"). Both ChatGPT-3.5 and Bard provided a numerical value and explanations justifying their judgement[1].

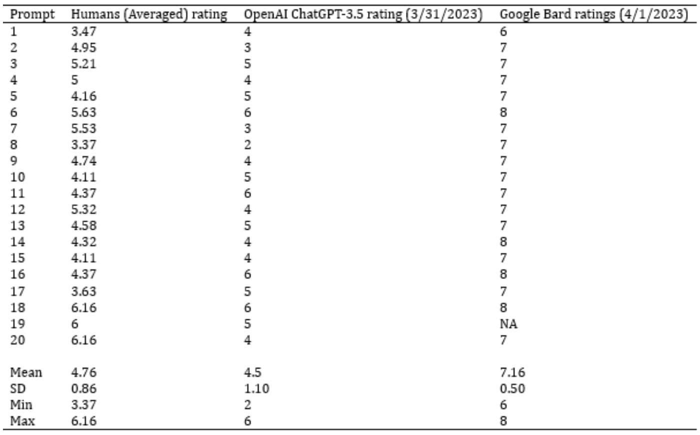
[1] The text of the prompts used, the numerical values of the complexity of the prompts justified by the AI tools, and the detailed justification for the complexity value by both ChatGPT-3.5 and Bard (ChatGPT-3.5 did not provide an answer to one prompt) are available on request by emailing the author.

The quality of rating by ChatGPT-3.5 and Bard was compared with the averaged ratings of the 19 human raters. The metric used was the intraclass correlation (ICC) which measures the degree of agreement between two or more judges or raters on ordinal measurements of the same objects. ICC is one of several measures of association or agreement used to quantify the intra-rater and the inter-rater reliability between judges when the ratings are on an ordinal scale. Four ICC values were computed for four inter-rater reliability measures: between human raters themselves, between human raters and ChatGPT-3.5, between human raters and Bard, and between ChatGPT-3.5 and Bard. The results are presented in the following sections. ICC estimates and confidence intervals were obtained.

Results and discussion

The data included 1-8 ratings (1 = barely complex, 8 = highly complex) on the complexity of writing prompts as homework assignments for students. The ratings by human raters were averaged over 19 raters and compared with the ratings produced by OpenAI ChatGPT-3.5 and the Bard. Table 2 below shows the numerical values and descriptive statistics for the complexity ratings of prompts produced by the human raters, the OpenAI ChatGPT-3.5, and Bard.

Table 2. Ratings on a 1-8 scale of the complexity of the writing prompts performed by human raters, ChatGPT-3.5, and Bard in response to, "On a scale of 1-8, how complex is this prompt for a student writing assignment homework? The prompt is: [prompt]."

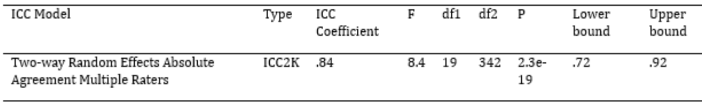


Prompt	Humans (Averaged) rating	OpenAI ChatGPT-3.5 rating (3/31/2023)	Google Bard ratings (4/1/2023)
1	3.47	4	6
2	4.95	3	7
3	5.21	5	7
4	5	4	7
5	4.16	5	7
6	5.63	6	8
7	5.53	3	7
8	3.37	2	7
9	4.74	4	7
10	4.11	5	7
11	4.37	6	7
12	5.32	4	7
13	4.58	5	7
14	4.32	4	8
15	4.11	4	7
16	4.37	6	8
17	3.63	5	7
18	6.16	6	8
19	6	5	NA
20	6.16	4	7
Mean	4.76	4.5	7.16
SD	0.86	1.10	0.50
Min	3.37	2	6
Max	6.16	6	8

The mean rating by the human raters is 4.76 (SD = 0.86) while those of ChatGPT-3.5 and Bard are 4.5 (SD = 1.10) and 7.16 (SD = 0.50). The mean rating by ChatGPT-3.5 seems to be similar to the averaged human ratings (and statistically similar, as shown by the Mann Whitney U test). However, I am more interested in knowing if the AI tools are as reliable as their human counterparts. To address this question, I calculated the intraclass correlation (ICC) as a measure of inter-rater reliability for multiple independent measurements on an ordinal scale produced by a random sample of judges. I computed two-way random effects intra-class correlation for four sets of ratings: between human raters themselves, between human raters and ChatGPT-3.5, between human raters and Bard, and between ChatGPT-3.5 and Bard. The reason I conducted an ICC among the human raters was to make sure that our benchmark or gold standard was reliable and could serve as a criterion (because I averaged the scores produced by human raters). I computed the ICC in the R statistical package (R Core Team) using the package psych (version 2.3.3). Inter-rater reliability measured by the intraclass correlation is formulated differently based on the model, type, and definition of the intended inference (McGraw & Wong, 1996). Because ICC is essentially based on analysis of variance (ANOVA), the output includes model statistics, such as the F value and the degrees of freedom for the F-distribution.

The inter-rater reliability for all human raters (the gold standard) was computed using two-way random effects absolute agreement multiple raters intraclass correlation (ICC2 in McGraw and Wong's (1996) classification and ICC (2,k) in Shrout and Fleiss's (1979) classification). Table 3 shows the results of the ICC analysis for human raters.

Table 3. Inter-rater reliability between human raters measured by intraclass correlation (ICC2K).

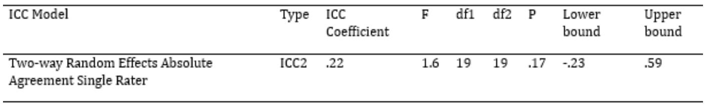


ICC Model	Type	ICC Coefficient	F	df1	df2	P	Lower bound	Upper bound
Two-way Random Effects Absolute Agreement Multiple Raters	ICC2K	.84	8.4	19	342	2.3e-19	.72	.92

As the 95% confidence interval indicates in Table 3 above, the inter-rater reliability for human raters is good to excellent (Koo & Lee, 2015). Now that I have verified the reliability of measures obtained by human raters, I compare the reliability of the AI tools with the human raters and between the AI tools using the ICC measure.

The inter-rater reliability between (mean) human ratings and the OpenAI ChatGPT-3.5 was measured using two-way random effects absolute agreement single rater intraclass correlation (ICC (2,1) in Shrout and Fleiss's (1979) classification). Table 4 shows the results of the ICC analysis for ChatGPT-3.5 and human raters' inter-rater reliability measure.

Table 4. Inter-rater reliability between ChatGPT-3.5 and human raters measured by intraclass correlation (ICC(2,1)).

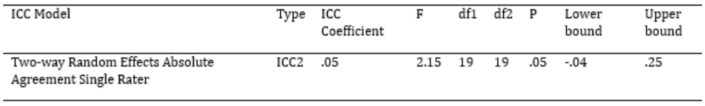


ICC Model	Type	ICC Coefficient	F	df1	df2	P	Lower bound	Upper bound
Two-way Random Effects Absolute Agreement Single Rater	ICC2	.22	1.6	19	19	.17	-.23	.59

As the 95% confidence interval indicates in Table 4 above, the inter-rater reliability between OpenAI ChatGPT-3.5 and human raters is poor to moderate and statistically nonsignificant.

Next, I measured the agreement between Bard and human raters. The inter-rater reliability between Google Bard and human raters was measured using two-way random effects absolute agreement single rater intraclass correlation (ICC (2,1) in Shrout and Fleiss's (1979) classification). Table 5 below shows the results of the ICC analysis between Bard and human raters.

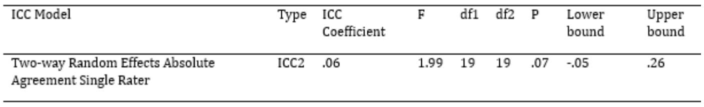
Table 5. Inter-rater reliability between Bard and human raters measured by intraclass correlation (ICC(2,1)).



ICC Model	Type	ICC Coefficient	F	df1	df2	P	Lower bound	Upper bound
Two-way Random Effects Absolute Agreement Single Rater	ICC2	.05	2.15	19	19	.05	-.04	.25

As the 95% confidence interval indicates in Table 5 above, the inter-rater reliability between human raters and Bard is poor, statistically nonsignificant, and lower in magnitude compared with that between ChatGPT-3.5 and human raters. Finally, I measure the inter-rater reliability between ChatGPT-3.5 and Bard using two-way random effects absolute agreement single rater intraclass correlation (ICC (2,1) in Shrout and Fleiss's (1979) classification). Table 6 shows the results of the ICC analysis between ChatGPT-3.5 and Bard.

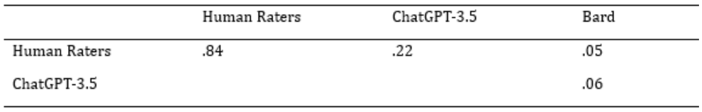
Table 6. Inter-rater reliability between ChatGPT-3.5 and Bard measured by intraclass correlation (ICC(2,1)).



ICC Model	Type	ICC Coefficient	F	df1	df2	P	Lower bound	Upper bound
Two-way Random Effects Absolute Agreement Single Rater	ICC2	.06	1.99	19	19	.07	-.05	.26

As the 95% confidence interval indicates in Table 6 above, the inter-rater reliability between the OpenAI ChatGPT-3.5 and Bard is poor and statistically nonsignificant. I have summarized the interrater reliability between human raters, ChatGPT-3.5, and Bard in a correlation matrix in Table 7 below.

Table 7: Interrater reliability between human raters, ChatGPT-3.5, and Bard in an ICC Matrix.



	Human Raters	ChatGPT-3.5	Bard
Human Raters	.84	.22	.05
ChatGPT-3.5			.06

As the summary ICC matrix shows in Table 7 above, the agreement between ChatGPT-3.5 and the human raters in rating the perceived complexity of writing prompts is low. Similarly, the agreement between Bard and the human raters is very low. However, the agreement between ChatGPT-3.5 and human raters is higher (r = .22) than that between Google Bard and human raters (r = .05).

Conclusion

Even in their early stages of development, Large Language Models (LLM) have found applications in a wide spectrum of industries, such as in content creation, code generation, graphics, and education, where humans have traditionally managed the operations. However, with current advances in computing, larger corpora, and more precise machine learning algorithms, LLM tools are closing their gap with the human performance. Nevertheless, in some applications, such as education and assessment, these AI tools need more finetuning and training to perform on par with their human counterparts. In the present study, I demonstrated with empirical data that ChatGPT-3.5 and Bard failed to achieve a performance comparable to human experts in rating the complexity of writing prompts. However, the difference in performance between the two LLM tools I tested in this experiment shows that there is some leeway in improving the models to close the gap with human performance. Our results in the present paper are in line with the findings by Rudolph et al. (2023b) who found that the performance of ChatGPT (both the free version and the commercial version) was much better than Google Bard (74 and 78 vs. 51) on an experiment where fifteen questions from different fields were asked from both AI tools, placing ChatGPT as a C-student and Bard as an F-student.

Natural language processing (NLP) has long been researched in the computer science field and has produced promising applications such as machine translation and expert systems which have tremendously helped task automation traditionally performed by humans. One aspect of language that most machine learning algorithms find challenging is the semantic and pragmatic aspects of language. Such aspects are still outperformed by human experts, as seen in machine translation, automated essay scoring, and automated item generation. The present study also supports this hypothesis that machines still are behind in performance compared to the human workforce in certain areas where tasks are more human-specific, such as translation and language comprehension due to semantic and pragmatic nuances. Therefore, at this stage of their development, tools such as ChatGPT-3.5 and Google Bard can only be trusted with some human supervision.

References

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). ChatGPT-3.5 goes to law school. *SSRN*. http://dx.doi.org/10.2139/ssrn.4335905.

Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory, 53*(12), 1229-1245.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation.* Routledge.

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2022). How well does ChatGPT-3.5 do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment. *medRxiv,* 1-9.

Huh, S. (2023). Are ChatGPT-3.5's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study. *Journal of Educational Evaluation for Health Professions, 20*, 1-5.

Iftikhar, L. (2023). Docgpt: Impact of ChatGPT-3.5-3 on health services as a virtual doctor. *EC Paediatrics, 12*(1), 45-55.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). Gpt-4 passes the bar exam. *SSRN 4389233*. http://dx.doi.org/10.2139/ssrn.4389233.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT-3.5 on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health, 2*(2), e0000198.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods, 1*(1), 30-46.

Pavlik, J. V. (2023). Collaborating with ChatGPT-3.5: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator, 78*(1). 10776958221149577.

Rudolph, J., Tan, S., & Tan, S. (2023a). ChatGPT-3.5: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9.

Rudolph, J., Tan, S., & Tan, S. (2023b). War of the chatbots: Bard, Bing Chat, ChatGPT-3.5, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364-389. https://doi.org/10.37074/jalt.2023.6.1.23.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Tate, T. P., Doroudi, S., Ritchie, D., Xu, Y., & Uci, M. W. (2023, January 10). *Educational research and AI-Generated writing: Confronting the coming tsunami*. https://doi.org/10.35542/osf.io/4mec3

Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education, 58*(3), 035027.