THE CONVERGENT VALIDITY OF MOBILE LEARNING APPS' USABILITY EVALUATION BY POPULAR GENERATIVE ARTIFICIAL INTELLIGENCE (AI) ROBOTS

Victor K. Y. Chan
Macao Polytechnic University
Rua de Luís Gonzaga Gomes, Macao SAR, China

ABSTRACT

This article seeks to explore the convergent validity of (and thus the consistency between) a few popular generative artificial intelligence (AI) robots in evaluating popular mobile learning apps' usability. The three robots adopted in the study were Microsoft Copilot, Google PaLM, and Meta Llama, which were individually instructed to accord rating scores to the eight major usability dimensions, namely, (1) content/course quality, (2) pedagogical design, (3) learner support, (4) technology infrastructure, (5) social interaction, (6) learner engagement, (7) instructor support, and (8) cost-effectiveness of 17 currently most popular mobile learning apps. For each of the three robots, the minimum, the maximum, the range, and the standard deviation of the rating scores for each of the eight dimensions were computed across all the mobile learning apps. The rating score difference for each of the eight dimensions between any pair of the above three robots was calculated for each app. The mean of the absolute value, the minimum, the maximum, the range, and the standard deviation of the differences for each dimensions between each pair of robots were calculated across all the apps. A paired sample *t*-test was then applied to each dimension for the rating score difference between each robot pair over all the apps. Finally, Cronbach's coefficient alpha of the rating scores was computed for each of the eight dimensions between all the three robots across all the apps. The computational results were to reveal whether the three robots awarded discrimination in evaluating each dimension across the apps, whether each robot, with respect to any other robot, erratically and/or systematically overrate or underrate any dimension over the apps, and whether there was high convergent validity of (and thus consistency between) the three robots in evaluating each dimension across the apps. Among other auxiliary results, it was revealed that the convergent validity of (and the consistency between) the three robots was marginally acceptable only in evaluating mobile learning apps' dimension of (1) content/course quality but not at all in the dimensions (2) pedagogical design, (3) learner support, (4) technology infrastructure, (5) social interaction, (6) learner engagement, (7) instructor support, and (8) cost-effectiveness.

KEYWORDS

Convergent Validity, Mobile Learning Apps, Usability, Generative Artificial Intelligence (AI)

1. INTRODUCTION

Mobile learning apps have become increasingly popular in recent years, with the number of adoptions ever soaring (Camilleri and Camilleri, 2020; Fan and Wang, 2020). As a result, there is a growing need to evaluate the usability of mobile learning apps to ensure that they are effective and user-friendly (Nielsen, 1994) and for users to select mobile learning apps from the market. For such evaluation, traditionally, usability testing has been conducted through manual methods such as questionnaire surveys, interviews, and focus groups (Maramba et al, 2019). However, these methods may be time-consuming, expensive, and subjective.

Generative artificial intelligence (AI) may be a promising alternative to traditional methods of evaluating mobile learning apps' usability. Generative AI refers to a category of AI algorithms that generate new outputs based on the data that they have been trained on. Unlike traditional AI systems that are designed to recognize patterns and make predictions, generative AI creates new content in the form of images, text, audio, and more. (Baidoo-Anu and Ansah, 2023; Gartner, 2023; World Economic Forum, 2023)

Albeit nascent, generative AI, and AI in general, in the context of education has already been extensively examined in extant literature. Macroscopically, for example, Gligorea et al (2023) is a compendious literature review of major AI categories' application to education and learning. More specifically, Leiker et al (2023) addressed the gap concerning the impact of using generative AI to create learning videos with synthetic virtual instructors. In the experiment with $n = 83$ adult learners, those in both the experimental group (adopting an AI-generated learning video with a synthetic virtual instructor) and the control group (employing a traditionally produced instructor video) demonstrated significant improvement from pre- to post-learning ($p < .001$), with no significant differences in gains between the two groups ($p = .80$), and with no qualitative differences in the perceived learning experience. These findings suggest that AI-generated learning videos have the potential to be a viable substitute for videos produced via traditional methods in online educational settings, making high quality educational content more accessible across the globe.

Ilieva et al (2023) spearheaded another application domain. The authors explored and compared the main characteristics of existing educational chatbots leveraging generative AI. They further proposed a new theoretical framework for blended learning with intelligent chatbots integration enabling students to interact online and instructors to create and manage their courses using generative AI tools. The advantages of the proposed framework are as follows: (1) it provides a comprehensive understanding of the transformative potential of AI chatbots in education and facilitates their effective implementation; (2) it offers a holistic methodology to enhance the overall educational experience; and (3) it unifies the applications of intelligent chatbots in teaching–learning activities within universities.

Regrettably, the author is not aware of any existing literature head-on evaluating mobile learning apps' usability by means of generative AI robots. This is exactly the gap that this article is to fill. In particular, this article is to utilize multiple popular generative AI robots to evaluate the various usability dimensions or perspectives of an appreciable number of popular mobile learning apps on the market, drawing on global users' comments on the apps' dimensions as appear on the web. The ultimate purpose is to determine whether the evaluations by the robots are consistent as gauged by the convergent validity of such evaluations. In fact, it is almost impractical to define the superiority of any particular robot's evaluation over others in view of the non-existence of any "yardstick" of evaluation as the paramount reference for any robot's evaluation to benchmark against. All one can do to decide which evaluation to be regarded as reliable is to measure the consistency between multiple evaluations. If a set of evaluations turn out to be consistent, it will be likely, if not absolutely, that all of them are reliable. This is exactly the concept of convergent validity of an operationalized instrument or scale to measure an abstract construct in most social or behavioral sciences. This article seeks to explore such convergent validity of and thus consistency between a few popular generative AI robots in evaluating popular mobile learning apps' usability.

2. METHODOLOGY

2.1 Data and Materials

The present study experimented with three very popular generative AI robots, namely Microsoft Copilot (Cambon et al, 2023), Google PaLM, (Anil, 2023), and Meta Llama (Oxford Analytica, 2023) as candidates for the evaluation of mobile learning apps' usability, the first being bundled with the Microsoft Edge browser and the other two being accessible through the AI portal poe.com. Eight major dimensions to evaluate a mobile learning app's usability were adopted (Albelbisi, 2020; Hew and Cheung, 2014; Khalil and Ebner, 2014; Kizilcec et al., 2013; Liyanagunawardena et al., 2013), they being (1) content/course quality, (2) pedagogical design, (3) learner support, (4) technology infrastructure, (5) social interaction, (6) learner engagement, (7) instructor support, and (8) cost-effectiveness, which were rated by each of the above robots. Content/course quality measures the overall quality and relevance of the course content, including the course design, instructional strategies, and assessment methods. It is essential to ensure that the course content is up-to-date, accurate, and relevant to the learners' needs. The quality of the mobile learning app's content is a critical factor that affects learners' satisfaction with the course. Pedagogical design refers to the design of the courses, including the teaching methods, assessment strategies, and learning outcomes. It is essential to ensure that the courses are designed in a way that encourages active learning and promotes learner engagement. The pedagogical design of a mobile learning app's courses is a critical factor that affects learners' engagement and motivation. Learner support includes the support provided to learners throughout the courses on the mobile learning app. It is essential to ensure that learners have access to adequate support, including technical support and academic support. Learner support is a critical factor that affects learners' completion rates and overall satisfaction with a mobile learning app's courses. Technology infrastructure measures the technological capabilities of the mobile learning app, including its ability to deliver course content, interact with learners, and manage learner data as well as how user-friendly, accessible, and reliable it is. Technology infrastructure may be assessed by metrics like a mobile learning app's uptime, speed of content delivery, compatibility with different devices, and security measures. Social interaction refers to the opportunities for learners to interact with each other and with the instructors. It is essential to ensure that learners have opportunities to collaborate, discuss, and share ideas with each other. Social interaction is a critical factor that affects learners' engagement and satisfaction with a mobile learning app's courses. Learner engagement measures the level of interaction between learners and the course content. Engagement is a crucial factor in determining the effectiveness of a mobile learning app as it affects the learning outcomes of learners. It can be appraised by such metrics as the average time spent on the course content, the number of interactions with the course materials, the number of forum posts and comments by learners, and the average completion rate of the mobile learning app's courses. Instructor support measures the level of support provided to learners by the instructors. Instructor support is important because it fosters a sense of community and increases learner engagement. It can be gauged by metrics such as the response time to learners' queries, the quality of responses to learners' queries, the availability of instructors during course hours, and the frequency of instructor-led sessions. Cost-effectiveness measures the cost of delivering the course content and the benefits derived from it. Cost-effectiveness is important because it determines the viability of a mobile learning app as a mode of delivering education. It can be translated into metrics, namely, the cost per learner, the return on investment of the mobile learning app, the cost savings compared to traditional modes of education delivery, and the revenue generated by the mobile learning app. (Chan, 2023)

The data collection commenced by searching for some popular mobile learning apps through specifying the keywords "mobile learning apps" for the Google search engine, which enumerated 18 apps below:

"Khan Academy, Quizlet, Photomath, EdApp, Kahoot!, iSpring Learn, Adobe Learning Manager, Connecteam-All-In-One, Dayforce, 360Learning, TalentCards, Adobe Connect, Moodle, Socrative Student, TalentLMS, LearnUpon, Trainual, BrainPOP Jr," referring to them as "Applications / M-Learning: From sources across the web." It is noteworthy that the search above effectively "commissioned" the Google search engine to shortlist currently popular mobile learning apps.

Then, the following request, explicitly spelling out the above 18 apps, was submitted to Copilot, PaLM, and Llama individually:

> "For each of the eight dimensions (1) content/course quality, (2) pedagogical design, (3) learner support, (4) technology infrastructure, (5) social interaction, (6) learner engagement, (7) instructor support, and (8) cost-effectiveness, please give a rating score to each of the popular mobile learning apps (namely, Khan Academy, Quizlet, Photomath, EdApp, Kahoot!, iSpring Learn, Adobe Learning Manager, Connecteam-All-In-One, Dayforce, 360Learning, TalentCards, Adobe Connect, Moodle, Socrative Student, TalentLMS, LearnUpon, Trainual, BrainPOP Jr. or as large a subset of them as you like) based on a scale of 1 to 10 (1 being the worst and 10 the best). Please derive your scores from global users' textual comments on these eight dimensions of these platforms as appear all around the web. It would be nice if you put your scores in a table form."

All the three robots replied with the rating scores in all the eight dimensions for all the 18 apps above. It is worth noting that the request above accentuated "…derive your scores from global users' textual comments on these eight dimensions of these platforms as appear all around the web." In order words, the robots were instructed to derive their scores from global users' textual comments appearing all around the web instead of echoing any analogous scores already published somewhere on the web or elsewhere. Also, in view of BrianPOP Jr. targeting the children's market (BrainPOP Educators, 2023) as opposed to the general education market of the remaining 17 apps, the former was excluded from further analysis.

It is noteworthy that all generative AI robots' outputs are dependent on the data on which they were trained, and such training data were inevitably updated up to a certain cutoff date. Any outputs are thus reflective of what the world was as of the cutoff date. By submitting to the three robots a simple question about their training data cutoff dates, it was revealed that Copilot, PaLM, and Llama of this study were trained on data up to somewhere around the end of 2023, September 2021, and September 2022, respectively. In other words, these three robots' rating scores relate to the 17 apps as of these three dates respectively. Equally noteworthy is that this study aimed to examine the convergent validity of (and thus the consistency between) these three robots in evaluating these 17 apps' eight usability dimensions, so its original scope did not include (and, as a matter of fact, its research resources were far insufficient to support) microscopically and technically discerning the way these three robots interpreted, comprehended, and measured each dimension of these 17 apps when being evaluated. Rather, such microscopic and technical (or even algorithmic) details were treated as black boxes such that this study focused on the ultimate evaluation results in the form of rating scores as what they were and on the convergent validity of (and thus the consistency between) these three robots' rating scores awarded to each dimension of the 17 apps irrespective of such microscopic and technical details.

2.2 Analysis

For each of the three robots, the minimum, the maximum, the range, and the standard deviation of the rating scores awarded by the particular robot for each of the eight dimensions were computed across all the 17 mobile learning apps. An appreciable range and standard deviation for a particular dimension signifies that the robot concerned accords discrimination in rating the dimension across the apps.

Then, the rating score difference for each of the eight dimensions between any pair of robots was calculated for each app. The mean of the absolute values, the minimum, the maximum, the range, and the standard deviation of the differences for each dimension between each pair of robots were calculated across all the 17 apps. If the mean of the absolute values, the range, and the standard deviation are sufficiently small for a particular dimension, it is indicated that the robots in the pair neither overrate nor underrate erratically with respect to each other the dimension across the apps. A paired sample *t*-test was then applied to each dimension for the rating score differences between each robot pair over all the 17 apps. If the *t*-test is significant for a particular dimension and the corresponding mean difference is positive (negative), it is implied that the first robot in the pair systematically overrates (underrates) the dimension with respect to the second robot.

Finally, for more statistically rigorous confirmation of the consistency between all the three robots' evaluation, Cronbach's coefficient alpha (DeVellis, 2005) of the rating scores was computed for each of the eight dimensions between all the three robots across all the apps. If Cronbach's coefficient alpha is high, for instance, over 0.5 or 0.6 (Ling et al, 2021; Nunnally, 1967) for a particular dimension, it is revealed that there is consistency between all the three robots in rating the dimension across the apps. Stated differently, the corresponding convergent validity of all the three robots in rating the dimension across the apps is high.

3. RESULTS

Table 1 enumerates the minimum, the maximum, the range, and the standard deviation of the rating scores as rated by each of the three robots for each of the eight dimensions across all the 17 mobile learning apps. Whereas all the three robots rated with considerable discrimination, Copilot did more so than the other two robots, especially, in the three dimensions learner support, social interaction, and instructor support as manifested by the disparity between the ranges and the standard deviations of these three dimensions' scores as rated by Copilot and those of other dimensions as also rated by Copilot and between the ranges and the standard deviations of most dimensions' scores as rated by Copilot and those as rated by the other two robots. By the same token, Llama rated the dimension learner support with less discrimination than it rated other dimensions and than the other two robots rated all the eight dimensions.

Table 1. The minimum, the maximum, the range, and the standard deviation of the rating scores as rated by each of the three robots for each of the eight dimensions across all the 17 mobile learning apps

Robot (sample size *n*)	Minimum/ maximum/ range/ standard deviation	Content/ course quality	Pedagogical design	Learner support	Technology infrastructure	Social interaction	Learner engagement	Instructor support	Cost-effectiveness
Copilot (*n* = 17)	Minimum	6	6	5	6	4	6	5	7
	Maximum	9	9	9	9	9	9	9	10
	Range	3	3	4	3	5	3	4	3
	Standard	0.9583	0.9785	1.2274	0.9984	1.4246	0.9583	1.1991	1.0847

	deviation								
PaLM (*n* = 17)	Minimum	7	7	7	7	6	7	7	8
	Maximum	9	9	9	9	8	9	9	10
	Range	2	2	2	2	2	2	2	2
	Standard deviation	0.6157	0.6157	0.6157	0.6157	0.6860	0.6077	0.6157	0.6157
Llama (*n* = 17)	Minimum	6	5	7	6	5	4	3	6
	Maximum	9	8	8	9	8	7	6	8
	Range	3	3	1	3	3	3	3	2
	Standard deviation	0.6691	0.6691	0.4609	0.6468	0.7838	0.6691	0.6691	0.6391

Table 2 lists the mean of the absolute values, the minimum, the maximum, the range, and the standard deviation of the rating score differences for each of the eight dimensions across all the 17 mobile learning apps between any pair of robots. Relative to PaLM, Copilot appeared to have overrated or underrated erratically the dimension social interaction in view of the corresponding mean of the absolute values, the corresponding range, and the corresponding standard deviation of the differences being greater than or equal to those for all the other seven dimensions. Likewise, in comparison with Llama, Copilot seemed to have overrated or underrated erratically the dimension social interaction as demonstrated by the corresponding range and standard deviation of the differences being greater than those for the remaining seven dimension.

Table 2. The mean of the absolute values, the minimum, the maximum, the range, and the standard deviation of the rating score differences for each of the eight dimension across all the 17 mobile learning apps between each pair of robots

Robot pair (sample size *n*)	Mean of the absolute values/ minimum/ maximum/ range/ standard deviation of the differences	Content/ course quality	Pedagogical design	Learner support	Technology infrastructure	Social interaction	Learner engagement	Instructor support	Cost-effectiveness
Copilot – PaLM (*n* = 17)	Mean of the absolute values	0.6111	0.9444	1.0556	1.0556	1.1667	0.8889	1.1111	0.8889
	Minimum	-2	-2	-3	-2	-2	-2	-2	-2
	Maximum	2	2	2	2	3	2	2	2
	Range	4	4	5	4	5	4	4	4
	Standard deviation	1.0432	1.1618	1.4061	1.2433	1.4818	1.0626	1.4142	1.2042
Copilot – Llama (*n* = 17)	Mean of the absolute values	0.8889	1.5556	1.1111	1.2778	1.3889	2.4444	3.2778	1.6111
	Minimum	-2	-1	-3	-2	-3	0	0	-1
	Maximum	2	3	2	2	3	4	5	4
	Range	4	4	5	4	6	4	5	5
	Standard deviation	1.0966	1.3284	1.4552	1.3198	1.8875	1.3525	1.7017	1.4127
PaLM – Llama (*n* = 17)	Mean of the absolute values	0.3889	1.2778	0.3889	0.4444	0.5556	2.3333	3.2778	1.6111
	Minimum	-1	0	-1	-1	-1	1	2	1
	Maximum	1	2	2	1	2	3	4	3
	Range	2	2	3	2	3	2	2	2
	Standard deviation	0.5745	0.5745	0.6691	0.5941	0.8085	0.5941	0.5745	0.6978

Table 3 depicts the paired sample *t*-tests of the rating score differences for each of the eight dimensions between each pair of robots over all the 17 mobile learning apps. Vis-à-vis Llama, Copilot tended to systematically overrate the five dimensions pedagogical design (at the 1% significance level or $p < 0.01$), technology infrastructure ($p < 0.05$), learner engagement ($p < 0.01$), instructor support ($p < 0.01$), and cost-effectiveness ($p < 0.01$) whilst PaLM inclined to systematically overrate the four dimensions pedagogical design, learner engagement, instructor support, and cost-effectiveness, all at the 1% significance level

($p < 0.01$). Otherwise, with respect to each other, the three robots neither overrated nor underrated systematically any other dimensions.

Table 3. The paired sample *t*-test of the rating score differences for each of the eight dimensions between each pair of robots over all the 17 mobile learning apps

Differences (sample size *n*)	Dimension	Mean difference / [95% confidence interval]	*t* (*p*-value) / degrees of freedom
Copilot – PaLM ($n = 17$)	Content/course quality	.059 / [-.438, .556]	.251 (.805) / 16
	Pedagogical design	.000 / [-.603, .603]	.000 (1.000) / 16
	Learner support	-.294 / [-1.038, .450]	-.838 (.415) / 16
	Technology infrastructure	.353 / [-.301, 1.007]	1.144 (.269) / 16
	Social interaction	.118 / [-.629, .865]	.334 (.743) / 16
	Learner engagement	.000 / [-.545, .545]	.000 (1.000) / 16
	Instructor support	.059 / [-.656, .774]	.174 (.864) / 16
	Cost-effectiveness	-.176 / [-.786, .433]	-.614 (.548) / 16
Copilot – Llama ($n = 17$)	Content/course quality	.353 / [-.191, .896]	1.376 (.188) / 16
	Pedagogical design	1.294 / [.596, 1.993]	3.928 (.001**) / 16
	Learner support	.000 / [-.771, .771]	.000 (1.000) / 16
	Technology infrastructure	.706 / [.007, 1.404]	2.142 (.048*) / 16
	Social interaction	.353 / [-.591, 1.296]	.793 (.439) / 16
	Learner engagement	2.353 / [1.674, 3.032]	7.349 (.000**) / 16
	Instructor support	3.353 / [2.501, 4.204]	8.348 (.000**) / 16
	Cost-effectiveness	1.471 / [.741, 2.200]	4.272 (.001**) / 16
PaLM – Llama ($n = 17$)	Content/course quality	.294 / [-.008, .596]	2.063 (.056) / 16
	Pedagogical design	1.294 / [.992, 1.596]	9.077 (.000**) / 16
	Learner support	.294 / [-.059, .647]	1.768 (.096) / 16
	Technology infrastructure	.353 / [.041, .665]	2.400 (.029) / 16
	Social interaction	.235 / [-.192, .663]	1.167 (.260) / 16
	Learner engagement	2.353 / [2.041, 2.665]	16.000 (.000**) / 16
	Instructor support	3.294 / [2.992, 3.596]	23.104 (.000**) / 16
	Cost-effectiveness	1.647 / [1.286, 2.008]	9.675 (.000**) / 16

** $p < 0.05$; ** $p < 0.01$

Table 4 delineates Cronbach's coefficient alpha of the rating scores for each of the eight dimensions between all the three robots over all the 17 mobile learning apps. Of all the eight dimensions, only the dimension content/course quality rendered a value of Cronbach's coefficient alpha marginally high enough (Nunnally, 1967) to indicate consistency between the three robots in evaluating the dimension. The values of Cronbach's coefficient alpha for the dimensions learner support, technology infrastructure, instructor support, and cost-effectiveness were found unavailable, implying sublimely inconsistency between the three robots. Likewise, those for the dimensions pedagogical design, social interaction, and learner engagement were all less than .15, uncovering similar inconsistency probably to a lesser extent. In summary, on the one hand, the convergent validity of the three robots was marginally acceptable for the dimension content/course quality, and thus the three robots may be rather reliable in evaluating this dimension of mobile learning apps' usability. On the other hand, the convergent validity of all the remaining seven dimensions was far from practicality, and thus one is better off refraining from evaluating these dimensions of mobile learning apps' usability by these three robots at least in the way adopted in this study.

Table 4. Cronbach's coefficient alpha of the rating scores for each of the eight dimensions between all the three robots over all the 17 mobile learning apps

Sample size *n*	Content/ course quality	Pedagogical design	Learner support	Technology infrastructure	Social interaction	Learner engagement	Instructor support	Cost-effectiveness
17	.566	.115	Nil [a]	Nil [a]	.149	.141	Nil [a]	Nil [a]

[a] In violation of the assumptions underlying Cronbach's coefficient alpha due to a negative average covariance among the rating scores accorded by the three robots.

4. CONCLUSION

There are quite some factors underlying inconsistency between generative AI robots in the evaluation of mobile learning apps or electronic learning platforms in general (or, in fact, anything under the sun). Albeit generative AI robots are promising as a new-fangled method to incisively analyze global users' textual comments at scale and to rate the multifaceted dimensions of each mobile learning app based on such comments, robots are beset by a number weaknesses. Inconsistency between different robots may be ascribed to the weaknesses as illuminated by Chan's (2023) study on MOOC platforms, which are tangentially comparable to mobile learning apps, and adaptively outlined below:

1. Textual user comments on mobile learning apps hinge on the content/courses experienced by the users concerned. Even for the same app, user comments may differ owning to the different content/courses studied.
2. Textual user comments are subjective and susceptible to bias or variation to the extent that even for the same mobile learning app, user comments may vary substantially across particular users.
3. Given profuse disparate user comments, a particular robot's ratings for a particular mobile learning app are very specific to the sample of user comments included in the robot's training. Therefore, it comes as no surprise to uncover discrepancy between two or more robots' ratings for the same app while the robots were presumably trained on different samples of user comments.
4. On the one hand, generative AI robots (inclusive of the three in this study) are powered by language models of considerably varied technologies and scales (Cambon et al, 2023; Anil, 2023; Oxford Analytica, 2023). For example, the number of parameters in the robots and the volumes of training data may differ tremendously. On the other hand, mobile learning apps (for example, the 17 ones in this study) may target disparate market niches and thus feature differently functionalities. In particular, the user interface, the gamification level, the multimedia content support, etc. may differ drastically across the apps. When different robots evaluate different apps, there may be a second degree of variation, precipitating inconsistency between the robots. The inconsistencies manifested in Tables 1 to 3 may have resulted as such.

One intriguing point in contrast with Chan's (2023) study on MOOC platforms is that the consistency found in the current study is far lower than that in Chan's. Whether this is due to the inherent nature of mobile learning apps versus that of MOOC platforms or due to the algorithms in the robots is beyond the scope of the current study and could be a subject of further research.

Also, this study itself is not without its critics. First, only three generative AI robots Copilot, PaLM, and Llama were experimented with in this study against the backdrop of myriad robots in operation worldwide. Second, these three robots were trained on data up to some cutoff dates, so even the rating scores generated by them today cannot catch up with the latest mobile learning apps and their versions. Therefore, it is invaluable to further extend the range of generative AI robots, in particular, those having incorporated the most current data in their training. Third, the disparity between the 17 apps regarding the volumes of global users' textual comments on them may have biased the evaluation by the three robots. Whereas this disparity is beyond the control of the author, this study tended to "absorb" such disparity and focused on the ultimate evaluation results in the form of rating scores as what they were and on the convergent validity of (and thus the consistency between) these three robots' rating scores awarded to each dimension of the 17 apps. In case of high convergent validity, all the three robots would theoretically be trustworthy to an extent even if such disparity existed.

Notwithstanding the relatively low convergent validity identified in this study and the other limitations above, generative AI robots are undeniably poised to be a major means of evaluation of opinions whether in academia or industry and whether in the domain of mobile learning apps or otherwise. Such evaluation is way less time-consuming, less expensive, less subjective, and broader in the coverage of more opinions from more users of more geographic locales worldwide than evaluation by humans.

REFERENCES

Albelbisi, N. A., (2020). Development and Validation of the MOOC Success Scale (MOOC-SS). *In Education and Information Technologies*, Vol. 25, No. 5, pp. 4535-4555. Accessed December 29, 2023 at https://doi.org/10.1007/s10639-020-10186-4

Anil, R., (2023). PaLM 2 Technical Report. *arXiv:2305.10403v3*. Accessed December 13, 2023 at https://doi.org/10.48550/arXiv.2305.10403

Baidoo-Anu, D. and Ansah, L. O., (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*, Vol.:7, No. 1, pp. 52-62. Accessed December 12, 2023 at https://dergipark.org.tr/en/pub/jai/issue/77844/1337500.

BrainPOP Educators, (2023). BrianPOP Jr. Accessed December 14, 2023 at https://educators.brainpop.com/contact-us

Cambon, A. et al., (2023). Early LLM-based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. Accessed December 13, 2023 at https://www.microsoft.com/en-us/research/uploads/prod/2023/12/AI-and-Productivity-Report-First-Edition.pdf

Camilleri, M.A. and Camilleri, A.C., (2020). The Students' Readiness to Engage with Mobile Learning Apps. *In Interactive Technology and Smart Education*, Vol. 17, No. 1, pp. 28-38. Accessed December 12, 2023 at https://doi.org/10.1108/ITSE-06-2019-0027

Chan, V. K. Y., (2023). Evaluating Popular MOOC Platforms by Generative Artificial Intelligence (AI) Robots: How Consistent are the Robots? *Proceedings of 20th International Conference on Cognition and Exploratory Learning in Digital Age 2023 (CELDA 2023)*. Madeira Island, Portugal, pp. 329-336.

DeVellis, R. F., (2005). Inter-rater reliability. In Kempf-Leonard, K., *Encyclopedia of Social Measurement*. Elesvier.

Fan, J. and Wang, Z., (2020). The Impact of Gamified Interaction on Mobile Learning APP Users' Learning Performance: the Moderating Effect of Users' Learning Style. *In Behaviour & Information Technology*, pp. 1-14.

Gartner, (2023). Gartner Experts Answer the Top Generative AI Questions for Your Enterprise: Generative AI isn't just a Technology or a Business Case — it is a Key Part of a Society in Which People and Machines Work Together. Accessed December 12, 2023 at https://www.gartner.com/en/topics/generative-ai

Gligorea, I. et al, (2023). Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review. *Education Sciences*, Vol 13, No. 12, pp. 1216. Accessed December 12, 2023 at https://doi.org/10.3390/educsci13121216

Hew, K. F. and Cheung, W. S., (2014). Students' and Instructors' Use of Massive Open Online Courses (MOOCs): Motivations and Challenges. *In Educational Research Review*, Vol. 12, pp. 45-58. doi: 10.1016/j.edurev.2014.05.001

Ilieva, G. et al, (2023). Effects of Generative Chatbots in Higher Education. *Information*, Vol. 14, No. 9, pp. 492. Accessed December 12, 2023 at https://doi.org/10.3390/info14090492

Khalil, H. and Ebner, M., (2014). MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review. *Proceedings of EdMedia 2014--World Conference on Educational Media and Technology*. pp. 1305-1313.

Kizilcec, R. F. et al, (2013). Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. *Proceedings of the Third ACM International Conference on Learning Analytics and Knowledge*. pp. 170-179. doi: 10.1145/2460296.2460330

Leiker, D. et al, (2023). Generative AI for Learning: Investigating the Potential of Learning Videos with Synthetic Virtual Instructors. In Wang, N. et al (eds), Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. AIED 2023. *Communications in Computer and Information Science*, Vol 1831. Accessed December 12, 2023 at https://doi.org/10.1007/978-3-031-36336-8_81

Ling, H.-C. et al, (2021). Exploring the factors affecting customers' intention to purchase a smart speaker. *In Journal of Retailing and Consumer Services*, Vol. 59.

Liyanagunawardena, T. R. et al, (2013). MOOCs: A Systematic Study of the Published Literature 2008-2012. *In The International Review of Research in Open and Distributed Learning*, Vol. 14, No. 3, pp. 202-227. doi: 10.19173/irrodl.v14i3.1455

Maramba, I. et al, (2019). Methods of Usability Testing in the Development of eHealth Applications: A Scoping Review. *In International Journal of Medical Informatics*, Vol. 126, pp. 95-104. Accessed December 12, 2023 at https://doi.org/10.1016/j.ijmedinf.2019.03.018.

Nielsen, J., (1994). *Usability Engineering*. Morgan Kaufmann Publishers, San Francisco, USA.

Nunnally, J. C., (1967). *Psychometric Theory*. McGraw-Hill, New York, USA.

Oxford Analytica, (2023). Meta LLaMa leak raises risk of AI-linked harms. *Expert Briefings*. Accessed December 13, 2023 at https://doi.org/10.1108/OXAN-ES276597

World Economic Forum, (2023). What is Generative AI? An AI Explains Accessed December 12, 2023 at https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/