

A COMPARATIVE ANALYSIS OF USER REVIEWS AND AI-BASED ASSESSMENTS OF EDUCATIONAL APPLICATIONS

ANA MIRKOVIĆ MOGUŠ

Department of Natural Sciences, Faculty of Education,
University of Josip Juraj Strossmayer in Osijek
Cara Hadrijana 10, 31000 Osijek

E-mail address: amirkovic@foozos.hr

ORCID: <https://orcid.org/0009-0002-0081-1754>

ABSTRACT

Evaluation of early childhood learning applications requires a multifaceted approach that considers factors such as educational value, appropriateness for developmental age, and the integration of technology. Recent research emphasises the necessity for clear assessment instruments and performance scales with the goal of setting the criteria by which educational apps and programmes are evaluated to determine whether their design aligns with the developmental needs of young children. This study compares Google Play user reviews with AI-generated assessments of educational applications for children aged 3–5. User-generated content was analysed through sentiment analysis and topic modelling. The results are then compared with rubric-based assessment that have been generated by AI model which rates educational value, usability, and developmental appropriateness. Results indicate that there is a weak correlation between public perception and AI evaluations, where users mainly focus on aspects such as entertainment and interface design, whereas AI places more emphasis on pedagogical criteria. The research emphasises the methodological value of combining user data from the internet with automated analysis tools for the evaluation of digital learning environments. It contributes by illustrating the strengths and limitations of AI-assisted evaluation in educational technology research.

Keywords: educational technology, app evaluation, user reviews, artificial intelligence, early childhood

INTRODUCTION

Quality assessment and evaluation of effectiveness play an increasingly important role with the growing numbers of educational applications being integrated into formal and informal learning settings. Although user reviews can provide genuine feedback about usability and satisfaction of educational applications, they are subjective and not consistent. By contrast, AI-oriented evaluation tools offer a structured mechanism for evaluating user interfaces according to explicit pedagogical and usability criteria. In this study, 30 of the most popular educational preschool applications were compared using their user reviews found in app stores with the same apps. The aim was to identify convergence and divergence in human and algorithmic evaluations. Previous research shows that the majority of the chosen educational apps that have been assessed have above average rubric scores, allowing children to use and benefit from them to enhance their learning experiences and foster more critical thinking on subjects that interest them (Chatzopoulus et al., 2023). However, the findings of the study Strataki (2022) demonstrated that the majority of purportedly educational applications lacked actual educational value.

The rapid development of educational technology means that there is an increased presence of educational mobile/web-apps both in classrooms and informal settings. These applications facilitate interactivity, personalisation, and gamification with the intention of improving outcomes. However, due to the massive number and variety of such apps, it becomes increasingly difficult for educators, parents, and institutions to identify reliable, efficacious, and engaging tools. There are two prevalent methodologies for evaluating educational applications: user reviews generated within app store platforms and online review sites, and expert and algorithm driven evaluations. User-generated reviews are beneficial for understanding the real-world experience of end-users such as how usable, engaging, and functional the apps are; however, they do not provide consistency, and tend to be heavily laden with emotion, compared to expert- and algorithm- based methods. An essential part of examining user opinion on educational platforms is sentiment analysis (Ahammad et al., 2024). Unlike user-generated reviews, AI-based assessments apply consistent and structured frameworks to measure app quality in terms of instructional design, user interface, accessibility, and data security. Using algorithms for evaluations eliminates biases associated with humans that might negatively impact scoring metrics but some subjective user experiences that pertain to context could be ignored. Despite the growing interest in both approaches, few studies have systematically compared them in the context of educational applications. The aim of this study is to close that gap by carrying out a comparative analysis of user reviews and AI-based evaluations of a sample of educational apps. Mapping the agreement and divergence between human and machine assessments provides a basis for more complex, multi-dimensional app evaluation strategies that combine the advantages of both views.

Recent studies regarding analysis of user reviews and AI-bases evaluations of educational apps, emphasise the rising role of AI in the development of educational tools and methods. The primary objectives of these studies are to understand the users' experience, to automatise the review analysis process, and to enhance educational assessments by means of AI. Use of AI in educational applications is considered a means of solving usability issues, making the assessment process simpler, and giving to the users an opportunity of individually tailored learning experiences. AI enabled educational software have numerous difficulties concerning their use which raise issues of their functionalities, performance, bias, and ineffective features, which influence users's satisfaction and learning outcomes. In order to advance user experience and make the learning methods adaptable for different types of users, it is important to deal with these issues (Alsanousi et al., 2023). The use of machine learning techniques for the classification of educational app reviews has yielded promising results. Research recorded 97% accuracy in sentiment identification and 94% in significant issue classification, thus showing the power of AI in making the whole process of user feedback analysis automated and efficient (Zahoor & Bawany, 2023).

AI plays a major role in various assessment methods for students which, in turn, is utilised in large part for formative evaluation as well as automatic grading. Such approach has the potential to reduce disadvantages of human grading in the area of bias and time-consuming tasks thus resulting in a more objective and efficient assessment system (González-Calatayud et al., 2021) (Parikh et al., 2023). A bibliometric analysis can show the main trends of AI usage in the education, such as the development of an intelligent tutoring system and learning analytics, which facilitate instructors and students alike, and, therefore, enable AI to be at the forefront of the changes in the education (Nguyen et al., 2024). Through systematic review Zhao et al. (2024) provide evidence of integrating GenAI in assessment design. AI technologies can be beneficial for education, but there are still issues to be resolved before these new tools can be utilised to their full potential. For example, it is necessary to ensure that AI driven test designs are suitable for students' learning and find solutions in case of technical or ethical issues. To do this, it is important to continue supporting technological competencies and helping students learn, rather than just forcing them to learn more.

METHODOLOGY

This study adopted a mixed methods comparative design to investigate the alignment and differences between user-generated reviews and AI-based assessments of educational applications. This study combined qualitative content analysis of user reviews with quantitative, rule-based assessments through an AI-powered evaluation framework. A purposive sample of 30 educational applications was chosen from the Google Play store, that target

preschools as a sample, focusing on the category of applications marked as educational by the developers themselves. The inclusion criteria were: more than 1,000 downloads, at least 100 user reviews, explicit description of educational goal set forth by developers, as well as whether it is available in English.

The sample included a mix of free apps across subjects such as language learning, maths, science, and general study skills. A total of 1,500 user reviews (approximately 50 per app) were collected. Reviews were scraped from app stores using custom scripts, then anonymised and cleaned for analysis. Reviews were filtered to include only those posted within the last 12 months and with at least 50 characters to ensure substantive content. Each app was evaluated using an AI-driven framework based on criteria derived from established models such as the REVEAC rubric (Papadakis et al., 2017). The REVEAC rubric served as the theoretical and methodological basis for assessing preschool educational applications, implemented through an AI-based scoring system that mirrors the rubric's 19 items rated from 1 to 4 (overall score 19-76).

Reviews were coded using thematic analysis with inductive and deductive coding techniques. NVivo software was used to identify patterns across dimensions such as ease of use, engagement, technical issues, and learning effectiveness. Sentiment analysis was conducted to gauge overall tone (positive, negative, mixed). AI assessments generated a score for each dimension in four domains: contents, design, functionality, and technical quality, each having multiple subdomains. Results from both data sources were compared using cross-tabulation, correlation analysis, and qualitative juxtaposition. Discrepancies and overlaps were interpreted to identify where human experience aligns or diverges from algorithmic evaluation. All data used were publicly available. No personally identifiable information was collected.

RESULTS AND DISCUSSION

Sentiment analysis and topic modeling of user-generated content are analysed and compared to rubric-based assessments of an AI model that rates educational value, usability, and developmental appropriateness.

Table 1. Thematic and sentiment analysis – five applications extracted

<i>Application</i>	<i>Positive themes</i>	<i>Negative themes</i>	<i>Overall sentiment</i>
App 1	Users praise the educational content (learning to sew and colour), the visuals, and the overall appeal for children.	Most complaints relate to the game's slowness, problems with coins (they disappear), and the lack of volume control for sound and music. Some users described the app as boring and illogical.	Mixed
App 2	Fun and realistic gameplay – users appreciate the creative roles (e.g. firefighter). Easy and visually appealing – graphics and design are well-rated. Interesting and simple for children – younger audiences especially enjoy playing.	Repetitive and quickly boring – some users feel the full version does not justify the price. Technical issues – broken buttons, crashes on different devices. Limited free content and unwanted purchases cause frustration. Basic level design and lack of challenge reduce user engagement.	Mixed
App 3	A fun and colourful game with a wide variety of mazes and characters. Intuitive controls and improvements like floating arrows. Offline mode, no ads, and the app's price are all praised. Suitable for children of all ages and even used as a relaxation tool for anxiety and PTSD. Developers respond to feedback and fix bugs.	Limited content in the free version and repetitive elements. Controls sometimes interfere with visibility, especially for younger users. Problems with offline functionality and incompatibility with devices with notch screens. The ability to purchase and access to external links is a concern for parents.	Positive
App 4	The application was described as a great fun, interesting and popular game, often used throughout the day. Users highlight the educational aspect of the game, as children learn while playing. The ability to watch videos to earn stars is considered useful. The game has a positive effect on the mood of users and brings them joy. It is easy to play and accessible to children of different ages.	Boredom after prolonged use is mentioned by a small number of users. Ads and in-app purchases cause occasional frustration. Some users note that children sometimes play the game too much, neglecting schoolwork. Lack of coins and minor confusion about functionality are also mentioned.	Positive

App 5	Educational content useful for children with special needs and early learning. Some users have praised the variety of content for young children and the fun interface. Some families have found value in the app during promotional periods.	Subscription issues: unwanted charges after trial period, lack of transparency, and inability to get a refund. Poor customer support: slow or automated response, unresolved complaints. Technical issues: freezes, being kicked out of the app, difficulty logging in. Content is often poor quality, recycled, and not aligned with educational goals. Difficult to navigate the app, not child-friendly.	Negative
-------	---	---	----------

Source: own research.

Rubric-based assessments of an AI model were conducted according to REVEAC rubric (Papadakis et al., 2017). Positive features and limitations of analysed applications are shown in table 2.

Table 2. AI assessment of selected applications according to REVEAC rubric

<i>Positive features</i>	<i>Limitations</i>
• high educational value, strong curriculum alignment, usability, and privacy standards • encourages socio-emotional and imaginative development • engaging, accessible, safe • engaging design, phonics, vocabulary building, child-safe • high entertainment value, interactive and motivational • visual appeal, theme-based exploration • well-designed, safe, and engaging	• needs improvement in personalisation and assessment analytics • lacks curriculum alignment, structured learning, and feedback systems • lacks goal-oriented tasks, adaptivity, and progress feedback • no adaptive learning or tracking tools • poor adaptivity, limited assessment, subscription barriers • weak educational structure, no adaptivity, minimal feedback • no tracking or adaptive learning features

Source: own research.

Furthermore, rating scores of applications set by users were analysed as well as ratings of the same set of applications obtained by AI model based on a REVEAC rubric. Also, user ratings and AI ratings were examined for a correlation. Results are shown in table 3.

Table 3. User ratings vs AI ratings

Applica- tions	User ratings						AI scores				
	N	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
30		4.03	0.59	3.95	3.40	4.70	3.27	0.67	3.05	2.60	4.70

Source: own research.

The Pearson correlation coefficient between the two dataset was approximately 0.388, indicating a weak positive correlation. However, the p-value is 0.268, which is not statistically significant at the common $\alpha = 0.05$ threshold. This means that there is no strong evidence of a linear correlation between the two datasets. There is a weak positive correlation, but it is not statistically significant. Any apparent relationship may be due to random variation. There was a relatively weak correlation between public perception and AI evaluations. User were primarily interested in entertainment and the design of the interface, while the AI focused on pedagogical aspects of applications. Limitation of the research reflects in biases related to potential review manipulation in app stores as well as potential AI model bias due to rubric interpretation.

CONCLUSION

This study shows that users and AI often see educational apps differently. While users tend to focus on fun, ease of use, or accessibility, AI assessments are more concerned with how educationally or technically reliable the app is. Because of these differences, relying only on one side, either people's opinions or algorithmic scores, can give an incomplete picture. This implies the importance of combining both perspectives. By bringing together real user experiences with structured AI evaluations, better support for app developers, teachers, and parents in choosing and improving educational tools for children can be provided.

REFERENCES

- Ahammad, M. S., Sinthia, S. A., Ahmed, M. R., Mou, F. Y., Ikram, M., & Ghosh, M. (2024). Sentiment analysis of user-generated reviews of online education applications using deep learning and transformer learning algorithms. *Proceedings of the 2024 International Conference on Computing, Communication, and Networking Technologies (ICCCNT)*, 1-7. <https://doi.org/10.1109/icccnt61001.2024.10724110>
- Alsanousi, B., Albeshier, A. S., Do, H., & Ludi, S. (2023). Investigating the user experience and evaluating usability issues in AI-enabled learning mobile apps: An analysis of user reviews. *International Journal of Advanced Computer Science and Applications*, 14(6). <https://doi.org/10.14569/ijacsa.2023.0140602>
- Artificial intelligence-based assessment in education (pp. 485-504). (2023). Cheltenham, UK: Edward Elgar Publishing. <https://doi.org/10.4337/9781800375413.00033>
- Chatzopoulos, A., Karafliis, A., Kalogiannakis, M., Tzerachoglou, A., Cheirchanteri, G., Sfiroera, E., & Sklavounou, E.-O. (2023). Evaluation of Google Play educational apps for early child-

- hood education. *Advances in Mobile Learning Educational Research*, 3(2), 770–778. <https://doi.org/10.25082/AMLER.2023.02.004>
- González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12), 5467. <https://doi.org/10.3390/app11125467>
- Nguyen, V. T., Phung, T.-N., & Cuong, D. (2024). A bibliometric and thematic analysis of systematic reviews of artificial intelligence in education (pp. 337–351). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-50818-9_37
- Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2017). Designing and creating an educational app rubric for preschool teachers. *Education and Information Technologies*, 22, 3147–3165. <https://doi.org/10.1007/s10639-017-9579-0>
- Parikh, R., Nimonkar, H., Gandhi, R., Budhrani, T., Dalvi, A., & Siddavatam, I. (2023). Streamlining educational assessment: A user-centric analysis of an AI-powered examination app. In *Proceedings of the 6th International Conference on Advances in Science and Technology (ICAST)* (pp. 341–345). Mumbai, India. <https://doi.org/10.1109/ICAST59062.2023.10454906>
- Strataki, A. (2022). An evaluation of educational apps for preschool-age children in Android and iOS. *Advances in Mobile Learning Educational Research*, 2(1), 278–288. <https://doi.org/10.25082/AMLER.2022.01.012>
- Zahoor, K., & Bawany, N. Z. (2023). Explainable artificial intelligence approach towards classifying educational Android app reviews using deep learning. *Interactive Learning Environments*, 32(9), 5227–5252. <https://doi.org/10.1080/10494820.2023.2212708>
- Zhao, J., Chapman, E., & Sabet, P. G. S. (2024). Generative AI and educational assessments: A systematic review. *Education Research and Perspectives*, 51, 124–155. <https://doi.org/10.70953/ERPv51.2412006>