Journal of Information Technology Education: Research

An Official Publication of the Informing Science Institute InformingScience.org

JITEResearch.org

Volume 21, 2022

A SYSTEMATIC LITERATURE REVIEW OF LEARNING APPS EVALUATION

Shahjad *	Jamia Millia Islamia New Delhi India, South Delhi, India	shahjad11089@gmail.com
Khurram Mustafa	Jamia Millia Islamia New Delhi India, South Delhi, India	kmustafa@jmi.ac.in

* Corresponding author

ABSTRACT

Aim/Purpose	The goal of this writing was not to promote any particular assessment tool. We aimed to critically explore the numerous assessment techniques that are accessible to app stakeholders with an emphasis on their strengths, shortcomings, and trustworthiness. We underline the importance of a relatively good and research-based tool that can readily assess the existing Learning Apps (LAs).
Background	A thorough and comprehensive literature review of LAs and their assessment tools was the primary goal of reporting the state of the art through this SLR (Systematic Literature Review) writing.
Methodology	We restricted our search space to ten databases and covered the most relevant studies from 2008 to 2022. To accomplish this predefined research interest, we divided our whole SLR methodology into four pertinent steps.
Contribution	The primary goal of the current writing is to know the state of the art regarding LAs' appraising instruments so that we can clearly reveal a list of essential research gaps on the same problem. Accordingly, app designers gain valuable insight from these forms of texts in order to develop better LA(s).
Findings	After careful examination of included studies (114), we found a total of 70 studies that discussed at least one evaluation tool in their research design, and the remaining studies were useful for theoretical support in writing this review. Although we discovered a large list of evaluation tools on LAs, the majority are suffering from some serious flaws. This emphasizes the need for a concise, comprehensive, and concrete theoretical evaluation tool for LAs.

Accepting Editor Athanassios Jimoyiannis │ Received: July 30, 2022 │ Revised: October 24, November 20, November 22, 2022 │ Accepted: November 26, 2022.

Cite as: Shahjad & Mustafa, K. (2022). A systematic literature review on learning apps evaluation. *Journal of Information Technology Education: Research, 21,* 663-700. https://doi.org/10.28945/5042

Recommendations for Practitioners If practitioners incorporate the summarized research findings into their app design process, it may be possible to produce high-quality educational apps that could significantly improve our current educational system.

Recommendations for Researchers We analyzed a large amount of relevant literature on LAs assessment. As a result, we have represented the current state of the art as well as some other key research discoveries in a clear and concise manner on the same research design. Thus, novices may easily gain a theoretical understanding by reading this research article, rather than having to read many individual pieces of literature, which may be a time-consuming process.

Impact on Society Education is a crucial component of our society. In light of this, we did a thorough literature review of LAs and discovered a number of deficiencies. If we research and answer these flaws scientifically, it may be possible to create high-quality apps that could considerably improve our current educational system.

Future Research Future research may focus on developing a sound framework or model for evaluating educational apps and being tested on our self-designed LA.

Keywords learning apps, educational mobile app, evaluation, assessment, framework, content, pedagogy, technology

INTRODUCTION

Learning is defined as a relatively permanent, but demonstrable, change in a person's knowledge or behavior due to their experience. Moreover, learning is a never-ending process that begins at birth and continues throughout life. As a process, it can be formal, informal, direct, indirect, synchronous, and asynchronous (Gupta, 2022). There are numerous favorable benefits of these learning characteristics, including the learners' ability to effortlessly acquire new knowledge, abilities, habits, attitudes, and aptitudes during the learning process (McGrath, 2011). The proof of such learning features has been seen in various learning and teaching models proposed by worldwide educationists, including the lecture model, demonstration model, d-Learning models (used by Isaac Pitman; Brodsky, 2021), e-Learning based models (used by Elliott Masie; Cross, 2004), and blended learning models (used by Bonk and Graham; Graham, 2009).

As the name suggests, the meaning of blended is mixed combining traditional teaching methods with modern ones. We have many examples of such models, for example, the integration of eLearning (learning through electronic media) with traditional teaching models, using m-Learning (learning through portable mobile devices) with old teaching methods and integrating Learning Apps with existing learning strategies. The blended learning paradigm has dominated in recent years due to its extensive coverage or assimilation of learning characteristics. As a result, the uses of such models in today's virtual classroom are becoming increasingly popular for several circumstantial and other good reasons. Such reasons often include the modalities for enhancement, accessibility (anytime, anyplace, any device), user-friendliness, cost-effectiveness, mass-customizability, learner-centeredness, availability of global expertise, availability of effective collaboration, and overall utilization of the better learning experience (Bhatasana, 2020).

Existing literature indicates that learning through mobile apps has grown in popularity over the last few years. In line with that, researcher McGrath (2011) stated, approximately 50% of searching for learning purposes had been done with the help of smartphones. The working of these devices depends on their small and light-weight software called 'Apps' (Sanromà-Giménez et al., 2021; Tu et al., 2020). Of them, some came with smartphones and we called them inbuilt apps, and some are installed from some special online sources (Apps Stores). Due to their wide operability, everyone is trying to use them and wants to integrate them for their individual purpose. The m-App (Mobile App) industry is vast and continuously proliferating, with no clear boundaries defined yet for its development process. Their growth is now at an all-time high and reaching unprecedented heights that have been never seen before (Agarwal, 2021; Glomack, 2021). When these apps are used for educational purposes, they are referred to as Learning Apps (LAs), Educational Apps, Instructional Apps, Mobile Learning Apps, or eLearning Apps (eLAs).

LAs are a special kind of mobile app that have all the parameters needed for learning (Kalogiannakis & Papadakis, 2017). Furthermore, they are small, self-contained, light-weight software applications designed to perform a specific task and are runnable on multiple platforms including touch displays (Baran, 2014; Griffith et al., 2020). The primary goal of such apps is not to replace face-to-face learning mechanisms with virtual systems, but to augment existing teaching models beyond the predetermined physical classroom boundaries. However, the educational fraternity is continuously attempting to emphasize the unavoidable importance of these apps in today's classrooms. Meanwhile, they are trying to integrate the technology-based educational system into the conventional education system by incorporating these apps. As a result, anyone can use them without concern for social inequality or other such factors (Bentrop, 2014). Additionally, these apps are on the verge of revolutionizing the education system by doing away with the traditional classroom-based learning environment where a single teacher was in charge of instructing the entire classroom. In contrast, LAs have taken learning pedagogies to a new level where no one could have imagined ever before, and sometimes no need for teachers or instructors.

We are now the eyewitnesses of their numerous and undeniable educational contributions including effectively monitoring the classroom process, providing various feedback (Namukasa et al., 2016), tracking learners' progress, and providing virtual attendance facilities. Moreover, we can access them remotely and endlessly (Kay & Knaack, 2008; Taylor et al., 2022), and they are also supporting sustainable, situated, authentic, and connected learning-like features that have not been explored before. Most importantly, they are providing individualization (Shuler, C., 2009, 2012; SiteProNews, 2020) and multimodality (Neumann, 2018) like learning features that were not yet enabled by any other teaching paradigm. Now it has been verified that this learning software is extremely effective for children with disabilities as well (Bentrop, 2014; Bhatasana, 2020; Edsys, 2017; Gupta, 2022; InstructionalDesign.org., 2021; Mobile App Daily, 2021; Situated learning, n.d.).

Contrarily, the trends of this rapidly growing paradigm are multiplicatively increasing in our day-to-day lives due to their popularity and other factors. As of 2011, there were barely 40,000 LAs accessible in both Apple and Android app stores (Walker, 2010, 2011). However, in 2013, this figure surpassed 100,000, according to the data obtained from the Apple App Store. If we look at both app stores from the same year, the total number of apps is around 0.5 million. Similarly, the growth of these apps was predicted to be 500,000 in 2017 across both app stores (Kay, 2018a; Kolak et al., 2021; SiteProNews, 2020). Thus, we may predict the trends of these apps by observing these statistics, particularly in the various domains of information and communication technology (ICT). According to Shing and Yuan (2016), the market of educational apps, especially for preschoolers, will continue to grow in the foreseeable future as well. Despite their enormous importance in today's classrooms, they are confronted with some new emerging challenges (Flewitt et al., 2014; Kucirkova, 2019; Vincent, 2014; Walker, 2010).

In the meantime, we revealed some of the most relevant reviews on the same research domain, including Hussain et al. (2018), Mustaffa et al. (2016), and Papadakis (2021). Their studies included several prospective research-based arguments concerning the success factors of mobile apps, with a particular focus on LAs and their important assessment elements. Hussain et al.'s (2018) research was primarily concerned with the usability of learning software and addressed four emerging usability features. Mustaffa et al. (2016) wrote one of the most important articles on app rubrics, yet it only meant a single app evaluation tool, i.e., rubric(s), and left out other related ones. Following that, a single existing SLR article focused solely on LAs for preschoolers, with a small sample size of only 11 studies, and their included article might have a higher chance of self-biasness because it was written by a single author (Papadakis, 2021). With the aforementioned possibilities in mind, now is an admirable time to conduct a larger number of such studies. Teachers, parents, and other caretakers need a clear and reliable way to determine whether these apps are actually instructive and hence capable of enhancing students' learning experiences.

There appears to be a lot of research on LAs and their development that deal with design difficulties, but there appears to be relatively less research on their evaluation. That is the primary concern behind the construction of this study and our aim was to purposely cover the most relevant literature on such apps and provide a comprehensive overview of their assessment tools. Meanwhile, our aim was not to advertise any existing or upcoming appraising methods but to inform parents, teachers, software developers, and other stakeholders about the same, especially their strengths and shortcoming, so that we can transcend the existing design need for LAs to better our society's future. Results reveal that, despite the fact that the literature has only a small number of such tools, those that are available are not without severe problems, such as a lengthy list of their evaluation criteria and a lack of generalizability (Cherner et al., 2014), unavailability of research-based tools (Flewitt et al., 2014; Kucirkova, 2019; J. S. Lee & Kim, 2015; Papadakis, 2021; Walker, 2011), poor usability support (Lubniewski et al., 2017), and more.

In short, academicians have been far from developing evaluation tools for LAs that are both theoretically and practically sound and that will allow them to differentiate between apps that truly enhance learning and those that are merely entertaining (Kucirkova, 2019; Sanromà-Giménez et al., 2021; Walker, 2010). In addition, the entire educational community (especially those who believe in virtual learning) is eagerly anticipating the development of such scientific tools, but their eyes are aching to witness a miracle that has not yet occurred. This highlighted the urgent need for more clear, thorough, and evidence-based assessment frameworks for appraising the plethora of apps accessible in various app stores (Agarwal, 2021; Glomack, 2021; Kay, 2018a; SiteProNews, 2020). In order to improve the future of this new paradigm, it is important that educators, app developers, and other stakeholders have a solid grasp of what is now available in the way of app evaluation instruments and what else we produce by using this information.

Based on the aforementioned considerations, we formed a comprehensive review study with the help of 114 most relevant papers entitled 'An SLR on Learning Apps Evaluation'. Additionally, we compiled an exhaustive review methodology for conducting this review article by incorporating multiple stages, so that we can learn even more about the apps assessment instruments. In order to accomplish the main goal of this study, we thus focused on three important aspects: evaluation instruments, their assessment criteria, and the flaws associated with them. In order to address the first aspect, our team revealed five essential apps evaluation techniques (see the Results section) Then we examined a comprehensive list of current evaluation criteria and categorized them into three meaningful and manageable classes for easy interpretation, namely technology, pedagogy, and contents. The last theme concentrated on three types of rising concerns: general, evaluation, and design. Finally, we divided our research paper into several sections as follows: the first section is this introduction, the second describes the research background and the need for an SLR, the third section describes the SLR methodology, the fourth section describes the results for the research questions, the fifth section presents discussion, research directions, and some crucial findings, the sixth section describes prospective research implications, and the final section contains the conclusion and future work.

RESEARCH BACKGROUND

Researchers at Carnegie Mellon University began investigating mobile learning in 1994. In China, the concept of m-Learning came. Apps for mobile devices were first released by Apple in 2008 with the release of the iPhone. Apps had not yet become commonplace at that time. In the same year, technology had a profound impact on nearly every part of our lives, including education. This coincided with a rise in internet accessibility (Baran, 2014; Tu et al., 2020). To obtain necessary educational software, the same approach is employed (N. Sharma, 2021). Though, due to the consistent development of this software, everyone has been alarmed by the availability of their sheer volume. The majority of users and other stakeholders are unable to select an app that has true instructional content and is designed appropriately. With the tag of education, many apps may be found on different App Stores and are marketed as educational, but a majority of them have more game-like content than anything else. Students and parents alike have a difficult time to decide on a quality one. To address such kinds of emerging gaps, we require a team of competent academics who can create some theoretically sound evaluation tools for such applications. We, therefore, require more continuous research in this field.

There have been many studies completed on LAs design; however, there has been far less work reported on their evaluation domain. This, in turn, might have led to the uncontrolled development of LAs without many evidentiary theoretical bases and without realization of their especially valued 'behavior modification'. We hereby reveal several notable researchers on the topic of LAs assessment, including Baloh et al. (2015), Cherner et al. (2016), Falloon (2013), Handal et al. (2014), Hirsh-Pasek et al. (2015), Hussain et al. (2018), Kay (2018b), Kay and Knaack (2008), Leacock and Nesbit (2007), C. Y. Lee and Cherner (2015), McManis and Parks (2011), Mustaffa et al. (2016), Ok et al. (2016), Papadakis (2021), Shoukry et al. (2015), Stoyanov et al. (2015), Taylor et al. (2022), Vaala et al. (2015), Walker (2010), and Weng (2015). To date, there is only one SLR (Papadakis, 2021) published on educational apps evaluation tools, but the review study suffered from several key concerns. That is why writing a comprehensive review paper is needed. To do so, we examined some most suitable studies on learning app evaluation using our pre-proposed search strategy. The most prominent ones are as follows.

The history of learning evaluation can be traced back to very early times, although our primary focus here will be on virtual learning evaluation tools or other similar methodologies. That is why we begin our literature review with learning objects. We discovered some of the most important assessment tools for learning objects from the literature, such as the learning object review instrument (Leacock & Nesbit, 2007), learning object evaluation metrics (Kay & Knaack, 2008), and the objects evaluation tool for mathematics apps (Namukasa et al., 2016). Leacock and Nesbit (2007) carried out a review approach that comprised learning object users with nine critical evaluation metrics. The primary goal of their research was to balance the assessment validity with the efficiency of the assessment process. Their work, however, was intended to describe merely the theoretical concept of such evaluation instruments. Therefore, more rigorous efforts are needed.

Kay and Knaack (2008) used a similar concept but in a more extensive manner. They surveyed about 1,000 middle and secondary school students in order to create a multi-component assessment model for learning objects based on four major learning dimensions: interactivity, design, engagement, and usability. The primary goal of their research was to educate the educational community about the pedagogical impact of technology in today's educational setting. Furthermore, they explained the true features of excellent LAs that have been proven to actively satisfy their teachers and learners while being used. Their designed model's reliability, on the other hand, was examined on an ad hoc basis, with no standard measurement. As a result, the learning objects pertained to the mathematics and scientific disciplines, and the problem of generalization was identified. That is why, we must alleviate these potential learning concerns, by revealing the true root causes. It is time, then, to work out such factors.

Namukasa et al. (2016) designed a valid appraising tool for mathematical apps in terms of learning manipulatives, keeping the foundation of learning objects in mind. They did this by looking at four parameters commonly used in popular apps (curriculum, interaction, interactivity, and design elements). Their participants categorized 80 mathematics apps into three tiers based on these criteria (levels 1-3). The major purpose of their study was to aid educators and other caretakers in selecting high-quality apps and to advise app developers on how to make their products more engaging by going beyond merely imparting knowledge and instead creating an adaptive curriculum. Unfortunately, such situations still exist, and researchers are continuously working on such foremost considerations.

In the same domain (mathematics learning apps), Handal et al. (2014) proposed a framework (TPACK) for evaluating elementary and secondary LAs. Mishra and Koehler (2008) are credited with first proposing the TPACK model concept. Handal et al.'s model is one of several that have been published in the literature. However, most of them are not grounded in learning theory (Mishra & Koehler, 2008). Handal et al.'s model combines Technology, Pedagogy, and Content Knowledge (PK, TK, CK). The TPACK model resulted from the intersections of PK, TK, and CK. Finally, Handal et al. (2014) constructed app evaluation instruments for mathematics apps based on these four criteria and had them evaluated by educators. The evaluation process was completed by Tasks Structure, Cognitive Involvement, General Pedagogical, and Operational Factors. It should be noted that this was the first study that was used to evaluate mathematics apps and was based on a strong theoretical basis. While the bulk of mathematics apps available on the app store focus on drill and practice, we require a couple of LAs that target the same discipline but must be designed using appropriate pedagogies.

Ok et al. (2016) developed a rubric for appraising Learning Disabilities (LD) apps. Their evaluation process has three parameters: (i) identifying information, (ii) evaluation, and (iii) providing grading for an app. The validity of the recommended rubric was assessed by five experts by assigning 1 (minimum) to 3 (maximum) numbers to each evaluation element for evaluating each category. Their 13-element rubric was used to evaluate the simple math app. Afterward, the selected app received 36 out of 39 points, and the percentage of that app was calculated. Following that, they divided the apps into several categories based on their percentage. So, teachers, parents, and practitioners may try to use this rubric to assess the app quality but it was only applicable to LD students, and the rubric faced several other crucial flaws as well. Thus, the stakeholders need some generalized rubrics that may accommodate a variety of learning disciplines. Hence, we have summarized some pertinent apps assessment parameters that may become part of a new and generalized evaluation approach.

Weng (2015) presented a similar approach to Ok et al. (2016) that included only iPad LAs in his evaluation process, though the basic idea of his rubric was taken from Higgins et al. (2000). Weng's (2015) rubric included three specific evaluation criteria: (i) apps screening, (ii) formative evaluation, and (iii) summative evaluation. Then, they categorized all the evaluation criteria into four major dimensions (format, content, efficiency, and shared ability). Thus, the primary goal of their study was to evaluate the usability levels of special LAs. To do so, they included nine commercial apps targeting LD children (based on user reviews and other recommendations) in their usability test. Such tests had potential significance regarding the quality of the app selection. Additionally, the results of pre-test questionnaires reported that fewer users had utilized these evaluation tools for their apps evaluation purpose due to their lack of awareness regarding such tools. To fully inform teachers, parents, and app developers about such assessment approaches, we also shared their perceptions (from the literature) about the same tool so that app creators may readily learn the best-case scenario about these apps.

Green et al. (2014) evaluated science instructional apps using six quality criteria (accuracy, related-contents, sharing, feedback, inquiry and practice, and navigation). Meanwhile, they also compared MASS (mobile app selection for science learning) to ERMA (evaluation rubric for mobile applications). Collaboration, personalization, and authenticity parameters were used to evaluate the MASS rubric. Green et al.'s new rubric in the form of a framework was designed based on four design cycles. According to the researchers, the main conclusion was that accuracy and sharing were two novel quality factors that had not been investigated before the MASS rubric. Only four of the six quality criteria were successfully compared between the MASS and ERMA. In addition, pedagogical attributes played a major role in the designing of a rubric than technological ones. Therefore, we must emphasize these six apps evaluation factors. Furthermore, we need to design further new studies based on the aforementioned appraising criteria and establish empirical validation (J. S. Lee & Kim, 2015) of these elements.

Mustaffa et al. (2016) performed a meta-review of the app rubric and included seven pertinent rubrics after their comprehensive review process: (i) a rubric for LD that covers 13 learning dimensions (Ok et al., 2016); (ii) Evaluation rubric for ipod apps (Walker, 2010) rubric for mobile apps based on seven evaluation criteria; (iii) a rubric for preschool learners (REVEAC) based on 19 learning dimensions (Papadakis et al., 2017); (iv) Buckler's rubric (Buckler & Peterson, 2012) for assessing mobile applications design with six dimensions; (v) a rubric for early literacy learning having four evaluation criteria (Israelson, 2015); (vi) a rubric for language learning that was based on seven learning dimensions (X. Chen, 2016); and (vii) a rubric for educational apps that was based on the 24 learning dimensions (C. Y. Lee & Cherner, 2015). The primary objective of Mustaffa et al.'s (2016) study was to examine the effectiveness and comprehensiveness of existing rubrics for educators to make it easier to choose high-quality educational apps. Only the evaluation tools made by Walker (2011) and C. Y. Lee and Cherner (2015) were proven to be useful. However, the majority of educators believe that the tool proposed by C. Y. Lee and Cherner (2015) is superior to the other six. It is worth noting that, in general, all contemporary studies focusing on LAs assessment have adhered to the notions presented in these works. As a result, it has been emphasized that one must be aware of the strengths and weaknesses of the arguments and use some of the proceeds from current projects to fund other projects.

Tu et al. (2020) designed a prominent study on vocabulary LAs. With the help of a survey of 60 students, they identified the top 10 vocabulary apps. Based on these, they meticulously discovered the five most successful evaluation criteria for building a checklist for such vocabulary apps. These criteria include content quality, multimodal presentation, user engagement, personalization (Sanromà-Giménez et al., 2021), and usability. In the meantime, for the goal of validating their proposed checklist, they applied it to one of the most popular LAs, namely Vocabulary.com, and found that it was effective in terms of app features and app evaluation. However, their study had a small sample size (60), hence a larger sample space is required. Additionally, the selected app's evaluation parameters may be investigated further for a more practical and usable app evaluation framework. This stresses the requirement for additional follow-up research.

Cherner et al. (2016) proposed a rubric on teacher resource apps to assist teachers in completing their educational tasks. Their research aimed to identify teacher resource apps attributes and, based on such attributes, they proposed an assessment tool. Then, they discussed app rating systems and their flaws, such as when a teacher wants to download a LA from an app store, they normally see the available ratings given by others. Mostly, people who have used apps gave them good ratings. So, there might be a higher chance of their ratings being biased. Also, the rating service did not state why that particular app earned scores of 1, 2, 3, 4, and 5. In short, the ratings for these apps do not match with how well they teach (Harrison & Lee, 2018; Papadakis, 2021). In light of this, Cherner et al. (2016) developed their rubric in two steps: (i) a literature review of existing rubrics, and (ii) a categorization of reviewed dimensions into three broad categories. In addition, they supplied crucial rating criteria for the app selection. When deciding what the goal of an app review is, it is important to look at the app's rubric and size, look at an app carefully before downloading it, and rate it as if one were a real user. Despite their substantial significance, it has some flaws including the issues of rubric updating options and generalization.

Recently, the two most important tools for evaluating educational apps based on four important pillars (active learning, engagement, meaningful learning, and social interaction) have been designed by Hirsh-Pasek et al. (2015) and Meyer et al. (2021). Their primary concern was to investigate how educational these apps could be. Then they used a 0 (lowest) to 3 (highest) rating system for the apps selection process. After rigorous investigation, they claimed that all the learning pillars received low scores. That is the reason, behind the concern of educational specialists about the educational values of these ubiquitous apps. Such findings emphasized more improved versions of currently available apps (Meyer et al., 2021). Kolak et al. (2021) conducted a more detailed study on the same research subject. In this regard, their research focused on three emerging gaps in current app evaluation tools, as a long list of apps evaluation criteria, app-gaps or social disadvantages (Kolak et al., 2021; Vaala et al., 2015), and technical language.

Further, they offered a two-phase evaluation process by conducting their research and included both free and commercial apps in their design process. They found equal educational potential in both types of apps but differed in two ways (screen elements and object features). Paid apps include more screen elements and animations than free apps. Design and communication scored highest in all apps, whereas customized learning material scored lowest. Overall, this rating approach had a strong theoretical basis and satisfied the validity procedure as well; however, all three of them (Hirsh-Pasek et al., 2015; Kolak et al., 2021; Meyer et al., 2021) were designed to evaluate pre-school LAs and identified a lack of learning feature interaction. Thus, more research is needed to establish scientific evaluation methodologies for this under-development educational paradigm.

Several frameworks on LAs assessment have been published in the literature, with J. S. Lee and Kim (2015) being the most prominent one. Their frameworks evaluated educational apps using predefined criteria, including (i) teaching and learning, (ii) screen design, (iii) technology, (iv) economy and ethics. The 156 middle and high school students completed a survey based on 43 evaluation items. An exploratory factor analysis validated the framework. The results showed minimal evaluation of technical components (Walker, 2011). In terms of content and design features, the findings of this study are nearly identical to those of Falloon (2013). The findings of their study may assist students, teachers, and parents in selecting effective educational apps. However, their small sample size, issue of generalization, and targeting only game-based learning became obstructions in the way of this immature learning paradigm. Therefore, we require some sort of framework or method for evaluating various types of apps, and requisite validation with a large sample size.

Papadakis (2021) recently conducted a vital study on LAs evaluation. They looked at what had been already written about these apps for children aged 3 to 6 and discovered some interesting facts. According to them, very few effective LAs at app stores provide truly instructive information to their consumers (Vaala et al., 2015). Furthermore, even if such apps are available in app stores, it is unclear how to locate a decent one. Indeed, due to a dearth of open source and the fact that the majority of app developers lack a solid theoretical foundation, it is difficult to find such sound apps (Flewitt et al., 2014; Papadakis et al., 2017). So, with these concerns in mind, we consider that we need some quality based (Papadakis et al., 2020) and highly effective apps assessment approaches to make our lives simple and easy without any extra effort.

Kay (2018a) published an evaluation framework on quality mobile apps that was based on subject areas, classification, and categorization (Papadakis et al., 2017). Kay (2018a) analyzed the 2011 to 2017 literature based on three criteria (subject domain, classification, and categorization). They included the 11 most significant articles on LAs classification (Chergui et al., 2017; Cherner et al., 2014). They proposed their framework with eight emerging categories of LAs: (i) instructive, (ii) exercise-oriented, (iii) metacognitive, (iv) constructive, (v) productive, and (vi) communicative, (vii) collaborative, and (viii) game-oriented (Ebner, 2015; Falloon, 2013; Handal et al., 2014). Even though they did not provide any rubric or checklist for evaluating the quality of apps, their work was deemed valuable for this SLR writing.

Lubniewski et al. (2017) proposed an app evaluation tool called App Checklist Evaluators (ACE) based on four parameters, including (1) learner interest, (2) design characteristics, (3) curriculum connection, and (4) instructional characteristics. They included 151 teachers during their app evaluation process. Approximately, 133 teachers (out of 151) claimed that there was not even a single evaluation tool accessible in the market for legitimate educational apps evaluation objectives. Furthermore, they stated that for an app selection, the teachers generally used a random web search (35%), some used a recommendations strategy (35%), and some used other techniques for doing the same task. Finally, they proposed the ACE tool and evaluated it on three LAs. The majority of the evaluators gave the selected LAs a high rating based on the appraising procedure described above. However, their study had some serious flaws. Pertinent questions were raised about their evaluation parameters and the issue of their sample size. To put it briefly, it may now take many hours to build an app-rating tool that is understandable, succinct, and useful to a variety of LAs to enable an accurate and effective evaluation of each of the LA's components.

In addition to frameworks, we also compiled a list of some of the most important websites that are being used as assessment tools for the same purpose. The list of such websites includes Appitic, MindLeap, Best-Kid-App, Mac-App-Store, Fun-Educational-App, Smart-Apps-For Kids, Best-Apps-For Kids, Teachers-With-Apps, Apps in Education, and Educational Review (Pinkston, 2021). Taylor et al. (2022) did a more comprehensive study on LAs assessment by using the two most popular app rating websites (Good App Guide and Common Sense Media) based on eleven vital features. The major goal of Taylor et al.'s study, however, was to evaluate the learning potential of 0-4 year-old children by employing 10 high-rated apps and 10 low-rated apps from these websites. The results showed that higher app ratings have more potential educational content than lower ones. However, apps with higher ratings do not necessarily have actual learning values.

According to Meyer et al. (2021), around 60% of apps labeled as instructional were suffering from low quality in terms of educational potential. Furthermore, both types of apps (high and low ratings) have lacked age-developmentally appropriate learning content, adequate learning feedback, and more. Summarily, they contended that research is yet to be conducted by educators that can adequately measure the learning values of a certain LA. This reiterates the requirement of more rigorous, experimental, and improved versions of similar app rating websites or other similar approaches to successfully advise on future educational apps (Pinkston, 2021; Taylor et al., 2022).

Finally, we found and analyzed the only one (available) published SLR on learning app evaluation by Papadakis (2021) and conducted this study in line with that study. Papadakis investigated the literature from 2010 to 2020 for his review writing. Additionally, his review method comprised planning, researching relevant literature, analyzing, and interpreting the included studies. The majority of the included studies focused on rubrics as an app evaluation tool followed by checklists. Only three of the eleven apps assessment tools were utilized to evaluate preschoolers' apps whilst eight evaluation tools were intended for general purpose followed by special learners' apps. Moreover, the bases of approximately all evaluation tools were Walker's rubric (Walker, H., 2011), which is publicly available on the web (Walker, H., 2010).

However, these sorts of evaluation technologies are lacking in their complete set of evaluation criteria (ingredients) and peer reviews. To classify the research, Papadakis (2021) employed C. Y. Lee and Cherner's (2015) study (34-criteria evaluation method). The most stated evaluation criteria for scientific instruments were developmentally age-appropriate learning content, feedback, screen design, and a user-centered approach. Non-scientific evaluation methods favored content that was appropriate for the age of the learner. However, Papadakis' (2021) study had numerous flaws, including a small sample size of only 11 articles, and ignored those evaluation methodologies designed to assess adult learning apps. To get around the Papadakis study's drawbacks, we enlarged the sample size of pertinent research that does not just concentrate on preschoolers but covers a wide audience.

We envisage that assessment instruments or other similar techniques for digital learning are necessary for several reasons. These may include the need to know the pedagogical potential of digital learning tools (Kay & Knaack, 2008), the quick availability of desired learning material without investing any unnecessary time in searching desired one (Harrison & Lee, 2018; Kay, 2018a; Kolak et al., 2021; Namukasa et al., 2016; Papadakis, 2021; Rosell-Aguilar, 2017; Sanromà-Giménez et al., 2021; Taylor et al., 2022; Tu et al., 2020); to reuse them without further assessment (Kay & Knaack, 2008); and more.

However, designing such evaluation tool(s) that provides all the mentioned reasons to their apps caregivers or other stakeholders is not an easy task. Such evaluation instruments may be suffering from several serious issues. The most influential ones may be their long list of evaluation parameters of from 18 to 70 or more, which may make the assessment process impractical (Kolak et al., 2021), lack of their validation process (Walker, 2010), issue of generalization, the problem with their subjective criteria (Kolak et al., 2021; J. S. Lee & Kim, 2015; Shoukry et al., 2015), lack of adequate theoretical support (Hirsh-Pasek et al., 2015; Kolak et al., 2021), and many more. Thus, the primary concern of this study is to assess the most significant literature on educational apps and their related assessment tools without taking any proprietary assessment tools into account, but as a resource for normal stakeholders, particularly the producers and consumers. As a result, specialists of this immature learning paradigm (virtual learning) may require some significant precautions from this study, particularly when developing a solid, coherent, and viable model for true instructional mobile apps with a higher learning potential.

We proposed a list of significant research objectives for this review study, taking into account the aforementioned areas of interest. These included knowing the state of the art regarding current app assessment tools and identifying the ongoing research gaps in the same domain that are both imminent and pertinent in a setting that is already compelling and developing. Also on the agenda was investigating what pedagogical, technological, and contextual features a good learning app should have. So, it seemed like it was time for a more thorough research plan, with a focus on the pros and cons of the existing app evaluation tools, so that app developers could better understand the challenges of making good learning apps.

Research Questions

The first step in writing the SLR is to develop pertinent research questions, which serve as the foundation of the entire SLR approach. To include nearly all of the known dimensions of this specific research problem, i.e., learning apps evaluation, as well as to achieve some of the key research objectives that we mentioned at the early stage of this review writing we have designed a set of three research questions. Further, to achieve the desired objective(s), we carefully reviewed 114 of the most pertinent studies that were shortlisted for the proposed research problem. The vast majority of such studies concurred with the mentioned emerging issues. The majority of pertinent research indicates that there is much room for improvement in terms of the educational potential of children's touchscreen apps (Meyer et al., 2021; Taylor et al., 2022).

To that end, we have proposed a set of three research questions for systematically revealing what else is available regarding these apps and their evaluation tools:

RQ1: what are the pertinent evaluation methods, tools, and other techniques for assessing the LAs?
RQ2: what are the important evaluation criteria being used to address the evaluation of LAs?
RQ3: what are the emerging research gaps in the field of LAs?

RQ1 is concerned with the appropriate evaluation procedures, instruments, and other techniques for evaluating LAs such as rubrics, checklists, frameworks, and more; RQ2 is concerned with the crucial evaluation standards applied to LA evaluation; RQ3 is concerned with the present research challenges in the area of LAs. A detailed discussion of these questions is included in the Result section.

Methodology

SLR has a long history (Kraus et al., 2020). In the 18th century, Dr. James Lind noted the importance of a thorough and partial literature review on the Treatment of Scurvy medical sciences and had the first systematic literature (Hong & Pluye, 2018; O'Brien & McGuckin, 2016). Over the past 40 years, SLRs have been used in almost every scientific field. SLR's evolution has had three vital stages: foundation (1970-1989), institutionalization (1990-2000), and diversification (2001-today). An SLR begins with pre-planned research questions, identifies all relevant research papers on a specific research problem, evaluates qualitative results, and summarizes and synthesizes the included studies. Finally, it provides clear and targeted answers to proposed research questions and promotes study replication and transparency to eliminate research bias (Hannes & Claes, 2007).

We used a few review recommendations from Kitchenham and Charters (2007) and Okoli (2015) for writing this study. Afterward, we derived the inclusion/exclusion criteria process from Herodotou (2018) and Tahir and Wang (2017). Based on these researchers' review strategies, we completed our literature review process. Our review procedure followed well-defined steps from R1 to R5 (see Figure 1) and was structured in accordance with de Almeida Biolchini et al.'s (2007) review methodology. Then, the roles of both the authors and an additional expert (not an author of this study) came into existence (see the Quality Evaluation sub-section). The first author performed the primary search (R1) and filtering (R2) process. In the next stage, the authors of this article were involved in developing the inclusion and exclusion procedure (step R3). In the data evaluation process, we also included expert researchers. Step R4 is carried out by the first author of this article. Finally, all three persons were involved in the quality assessment process, i.e., step R5.

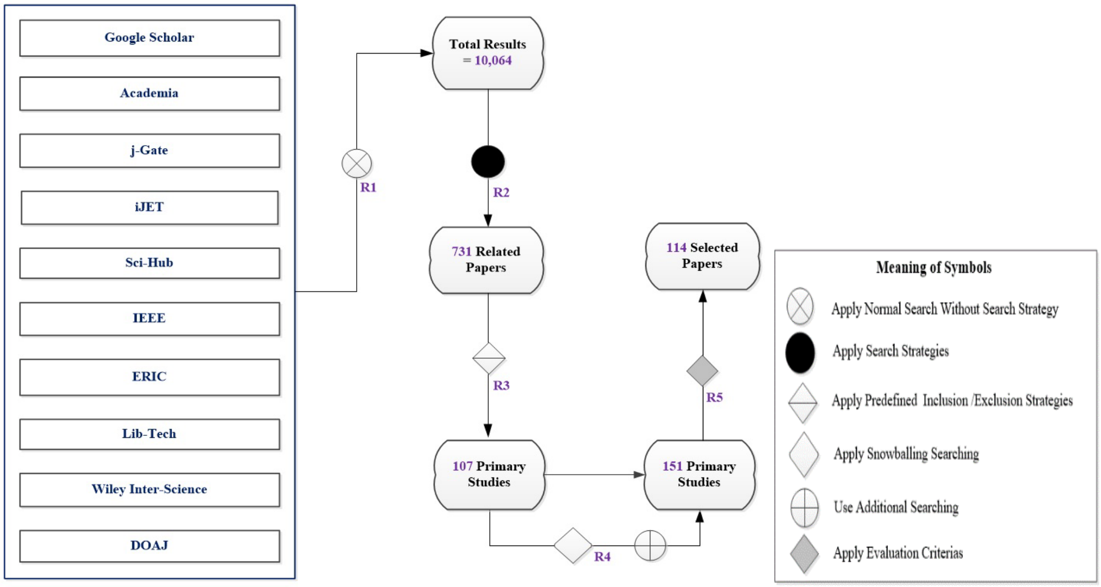

Figure 1. Search and selection procedure

The Search and Selection Process

This section discusses the search tactics we have used. We compiled a list of existing search techniques and created an operational search strategy that included numerous essential searching features; e.g., identifying the search period, constructing search strings, identifying prominent search terms, and identifying the searching databases.

Search period

From 2008, when Apple launched its first App Store, to July 2022, 'learning apps evaluation' were the major search keywords used for preliminary searches. We also carried out a manual search to find past study results using different search phrases, such as instructional software, instructional apps, learning apps, educational software, interactive applications, intelligent applications, mobile learning applications, eLearning Apps (eLAs), and assessment tools (Sheikh et al., 2019).

Searching protocols

We applied PICOC guidelines to design the search strings based on keyword identification (Griffith et al., 2020; Kitchenham & Charters, 2007; Mengist et al., 2020), detailed as follows.

- **P**opulation – educational applications, including LAs, educational software, interactive apps, mobile apps, eLAs, tablet apps, and instructional apps. The English language must be used in all studies, no matter where they are carried out.
- **I**ntervention – includes app appraising tools (rubrics, checklists, user reviews, frameworks, models, and other techniques).
- **C**omparison – after executing the first two steps, we employed the data evaluation process and compared it to the existing tools and techniques.
- **O**utcome – examine the current trends on LAs by conducting critical literature reviews on apps evaluation techniques and identify some emerging research gaps.
- **C**ontext – for easy understanding, we categorized all the included studies and assessment tools into two major dimensions (evaluation and design) based on critical analysis. However, our primary focus was on the evaluation dimension.

Identification of prominent keywords for search string construction

Before making the actual search strings, we roughly tested them on distinct search engines on a random basis. Then, we iteratively applied this process and examined the most frequently appearing keywords in our search process. A list of such keywords is evaluation, assessment, frameworks, educational-apps, learning-software, and other such combinations. A list of such major keywords is shown as follows:

LAs

LAs, mobile app, mLAs, educational apps, instructional apps, tablet apps, and some other software that are used for learning purposes.

Evaluation

Assessment, appraising, rubrics, checklist, user reviews, teachers-suggestions, websites-suggestions apps rating systems, and so on.

Frameworks

Frameworks, models, guidelines, and other appraising tools.

Construction of search strings and database selection

After conducting random tests on selected databases to identify a set of adequate search terms for the construction of viable search strings, we presumptively disqualified some of the search criteria, including design, pedagogies, and game-based learning. However, learning apps, intelligent apps, educational apps, instructional apps, mobile learning apps, literacy apps, tablets apps, evaluation, assessments, appraising, rubrics, checklists, models, modeling, framework, rating systems, end-users' reviews, suggestions, blogs, and websites were thought to be the major terms for search string construction. In the beginning, we used these terms on an individual basis for database searching. For instance, 'tools for LAs evaluation', 'LAs evaluation framework', and other such combinations. Then, based on these primary keywords, we produced a list of necessary synonyms to cover a comparatively large aspect of our problem definition.

After that, we used Boolean operators (OR, AND) to perform more advanced searches on the same search space with the help of Primary Search (PS) and Secondary Search (SS) strings. We applied the OR operator to combine the major search terms and their synonyms in an alternative fashion. It uses the either-or principle for an exhaustive search process. The AND operator was employed to link the primary keywords to narrow the scope of the search results. Boolean AND operators work only when both the inputs are set to a high at the same time. Using the aforementioned search methodology, we will demonstrate how to generate the search strings using two examples (ex.1, ex.2). Ex.1: (LAs) OR (mobile learning apps) OR (educational software) OR (intelligent apps) AND (evaluation) OR (rubrics) OR (framework). Ex.2: (eLAs OR tablets OR iPads OR touchscreen OR apps OR interactive media OR educational apps) AND (learning OR literacy OR mathematics) AND (toddlers OR young children OR preschoolers OR online learners OR e-Learners). A detailed discussion of complete search strings is shown in the Appendix.

Database selection was the next step after designing the search strings construction process. Based on a random search, we explored some of the prestigious databases for this SLR writing, such as Google Scholar, Academia, j-Gate, iJET, IEEE, ERIC, Lib-Tech, Wiley Inter-Science, DOAJ, and Sci-Hub. But we did not use some of the growing databases because of the following. At the beginning of this writing, we decided to only include the most well-known publications because their coverage could be summed up and was implicit. We made search strings for up to 10 pages of each database, which gave us over 200 results. But when we put our planned search terms into the Scopus database, we did not get many useful results. Most of them were the same or were not good enough for this review writing. We got a small number of results from the Web of Science database when we used the search terms we had already thought of. After giving them a close look, we found that they were already a part of other database studies we were doing. So, they were not able to become part of our writing. Also note that the Sci-Hub source was only used for article downloading purposes.

To select relevant literature, we performed a simple search, a random search, and a targeted search approach with the help of both simple and advanced search strategies (see Appendix). To find the maximum possible relevant studies, ten electronic databases were searched using several search terms in an iterative fashion. Each digital library has its own search characteristics, which was the reason behind conducting repeated experiments with slight tweaks to tailor the search strings for a large number of databases. In addition, for larger precision, we developed a set of supplementary search strings. Our search space was not limited to the aforementioned databases only; we also looked for technical reports and blogs.

Table 1. List of included databases

Database	Link	Selection procedure
Google Scholar	https://scholar.google.com/	Complete Search
Academia	https://www.academia.edu/	Complete Search
j-Gate	https://jgateplus.com/home/	Random Search
iJET	https://online-journals.org/index.php/i-jet	Complete Search
Sci-Hub	https://sci-hub.hkvisa.net/	Targeted Search
IEEE	https://ieeexplore.ieee.org/Xplore/home.jsp	Targeted Search
ERIC	https://eric.ed.gov/?journals	Targeted Search
Lib-Tech	https://www.learntechlib.org/	Complete Search
Wiley Inter-Science	https://onlinelibrary.wiley.com/	Random Search
DOAJ	https://doaj.org/	Targeted Search

Our search process began with the standard search process that was performed in step R1. At this step, we got a huge number of articles (N = 10,064). To refine our search results, we conducted 2,189 search processes at various phases of R2 to R5 (see Figure 1). We also filtered out the most relevant studies (731 from 2,189) during step R2 based on the pre-proposed search strategy. The filtering process in step R2 was further divided into two sub-steps: first, we did manual scanning based on the titles of the included papers (10,064); then, we examined the abstracts, keywords, and sometimes the whole text of some retrieved studies for additional filtering. As a result, we included only 731 relevant articles in our database pool at step (R2). A total of 107 most relevant articles were filtered out at step R3 by applying the predefined inclusion and exclusion protocols. Following that, we used the second stage of the search procedure, which included a reference search or snowballing search and an author-specific search in OBDB (online bibliographic database browsing). This complementary technique at step R4 enabled us to incorporate any possible work that was overlooked inadvertently. A total of 44 new articles were added after the additional searching, bringing our total number of articles to 151. Finally, after the quality assessment process (see Table 2) applied at step R5, we got a set of 114 highly relevant articles to answer our research questions that had been formulated in the early stages of this writing.

Management, Selection, and Inclusion-Exclusion Process

Management and selection of selected literature

We applied the literature selection methodology before the inclusion-exclusion process. So that we can more easily and purposefully complete the entire inclusion-exclusion process later on. The role of study selection came into play at the conclusion of the stage (R2). A total of 731 related research studies were incorporated following the completion of stage (R2). In order to manage this sheer number of studies (731), we have created 20 unique folders and subfolders on the D drive of our computer. Such as: 'apps evaluation' (separate folders for checklists, rubrics, guidelines, experts reviews, and user rating systems),' 'apps design', 'features and benefits for LAs', 'LAs ingredients', 'current state of the arts', 'research gaps', 'existing SLR', 'required SLR methodology' 'SLR related links and websites', 'e-Pedagogies', 'm-Learning', 'eLearning', 'latest papers', 'duplicate papers', 'most important papers', 'papers needed in future', 'glossary', 'summarized papers', and 'newly downloaded papers (without scanning). Note that, we have created additionally, 10 folders for all the databases as mentioned earlier. After that, we eliminated redundant studies from our database pool. Then we used the operational inclusion-exclusion criteria to optimize the remaining 731 studies (see upcoming section). Note that the data evaluation strategy was the more detailed version of the inclusion-exclusion process.

Inclusion exclusion criteria

It is essential to undertake a thorough evaluation of the included studies (731) in order to screen and select the most pertinent once. After applying inclusion-exclusion criteria at step (R3), we were only left with a set of 114 relevant studies for conducting this writing. Later on, we looked closely at each of the apps evaluation criteria we had included. Based on included studies, we put all of the important assessment criteria into three manageable categories: technology, pedagogy, and content. Afterward, we put each apps appraisal instrument in its right category based on its appraising parameters, strengths, and weaknesses. So that we may share these tools on an individual basis, which will allow us to do better study and come up with better findings in favor of developing educational apps. Do all LAs, for instance, employ the same evaluation standards? How they differ and how they are similar, etc. The inclusion-exclusion procedure is described in detail as follows:

Table 2. Inclusion-exclusion criteria

Inclusion Criteria (IC)	**Exclusion Criteria (EC)**
IC1: The study focused on the LAs (both preschoolers and beyond (adults, special learners, and others)).	**EC1:** The studies which do not focus on research findings.
IC2: The study focused on evaluation tools, techniques, methods, frameworks, rubrics, rating systems, guidelines, reviews, and others.	**EC2:** The studies which do not focus on learning or education dimensions.
IC3: The Study should be written in English language only.	**EC3:** The studies which focus on LAs design.
IC4: The study should be addressed at least one pre-proposed RQs.	**EC4:** The studies that are incomplete and have only abstract or presentation.
IC5: The period of included studies should be between 2008 to 2022.	**EC5:** The studies which were published based only on an opinion.
IC6: The study should be published in any conference, journal, and grey literature.	

Data management process

After executing the designed search strings on selected databases, we carried out the data evaluation process. It was used immediately after step R3 and before step R4. The data evaluation procedure consisted of six steps: title reading, abstract reading, diagonal reading, abstract + introduction, abstract + introduction + conclusion, and complete reading. It is worth noting that we used inclusion-exclusion criteria prior to the data evaluation process. It could be described as a more refined version of the inclusion-exclusion approach, in which the included studies must be rigorously analyzed and evaluated. Assume we were reading a specific research article, and there may be several possibilities that could be arrived at. For example, a specific research study can be identified solely by its title. Following that, if the study's title does not provide sufficient information about the research findings, we must read its abstract part also. If the abstract section contains sound knowledge parallels to our research discipline, it must be included, otherwise discarded. In some instances, it may not be necessary to read the article's title and abstract in order to receive sufficient information. When this is the case, we must read the study's introduction section as well. Moreover, even after reading the abstract, introduction, and conclusion sections of the selected paper, we may not be able to properly categorize it; therefore, we must read the selected paper thoroughly.

Data synthesis process

The primary goal of this phase is to determine how to extract the most important findings from the list of included studies. To accomplish this task, we read approximately 15 review papers and 20 SLR papers. Out of them, we are going to cite only the three important ones (Mengist et al., 2020; Okoli, 2015; Sheikh et al., 2019). The entire data synthesis procedure is depicted below:

Data Synthesis Process (SP):

- **SP1** – The R1, R2, R3, and R4 stages resulted in the compilation of a list of included studies. These studies were then divided into three categories: primary (A), secondary (B), and tertiary (C), and sub-categories (framework, design, tools, and design).
- **SP2** – Then we assigned each paper to its proper category or sub-category, considering that a single paper could be assigned to multiple categories.
- **SP3** – After evaluating each included paper, we ranked them as weak, average, or strong, according to their contribution.
- **SP4** – After assigning a rank to each of the included research studies, we organized them into folders or subfolders for citation purposes.
- **SP5** – Then, we examined and summarized each of the included studies in an MS Word file so that we could easily access the required article for this writing.
- **SP6** – Finally, all of the summarized studies written in MS Word were organized into a final table with relevant headings.

Data validity

Our relevant literature was validated using two criteria: internal validity and external validity. Internal validity was used to validate the respective folders and subfolders (this means that the contents of the created folders and subfolders were the same as we thought at the starting of step R3). If that was not the case, then we moved that particular research paper or group of papers to their appropriate folder or subfolder. We repeated this process until we had all the required contents in their respective folders or subfolders. The external validity of our included studies was focused on the primary papers (class A papers) that were also designed after step R3. Next, we did two things to evaluate the class A papers (which are closely related to or research problem): (i) if the study or studies met all of the necessary inclusion criteria as defined at the starting of this SLR writing, then we included them, and (ii) we included only those studies that were strongly targeted on our problem definition.

Quality evaluation

The quality evaluation procedure was initiated at step R5 to compile the final list of included literature by using the Conflict Resolution Table (CRT) that we created based on some of the quality assessment criteria, so that we can purposefully ascertain the evaluation, reliability, completeness, and significance of the included studies. In order to do this, we devised a scoring system. Accordingly, each quality evaluation received three possible scores from a total of five values (1 to 5) from both the authors and one more person. The overall quality score can be calculated by aggregating the total scores to obtain the quality control responses for a specific study. This is the place where the roles of both the authors and the third came into existence. Both authors were satisfied with the independent assessment of the included studies. Later on, we also included the company of that third person. Finally, we began evaluating the quality of the included studies with the help of CRT method. The complete procedure of this CRT technique is shown as follows:

- A study is either included or excluded without conflict if both authors concur on its inclusion or exclusion.
- A conflict would be resolved with the assistance of the second author, who is the most experienced member of our group, using the conflict resolution process if any of the participants (first author and third person) disagreed regarding a particular study.
- Additional upcoming conflicts were resolved using the same conflict table. Finally, the research studies that have been filtered out at this point may be considered potential candidates for writing this review.

Construction Procedure of CRT

- Step 1: After a healthy group discussion between both authors (first and second) and the third person, we decided on ‘5’ as the maximum scoring value that any group member can give for any conflict study AND the mathematical number ten (‘10’) was decided as a threshold value (the value gained after summation).
- Step 2: Next, we calculated the sum of each scoring value provided by all three group participants for a particular conflict study to accept or reject the purpose. Step 2 is further subdivided into two substeps:

```
Step 2.1: If (the sum of all the scoring values for a particular conflicted study is more than or
equal to 10) Then
          Accept it
Step 2.2: Else
          Reject it
```

- Step 3: **Repeat** Step 2 for each conflict study.
- Step 4: If no more conflicting studies remain, **Then**
- Step 5: **STOP** (end of the table construction process).

For demonstration purposes, we present a sample scoring technique for five conflicted papers in Table 3.

Table 3. Working of the conflict resolution table

Paper	Author-1	Author-2	3rd Member	Summation	Acceptance	Rejection
P1	3	2	1	6	✗	✓
P2	4	3	2	9	✓	✗
P3	1	4	6	11	✗	✓
P4	0	5	5	10	✓	✗
P5	5	1	3	9	✗	✓

Table 3 shows a sample of the scoring technique for five conflicting papers (P1 to P5). The table includes a sample of five papers for assessing the quality of included studies. P1 had a total summation of 6. Such values had been given by three participants as author-1 given 3, author-2 given 2, and

3rd member given 1 respectively (total summation was 6). Then, we compared the estimated value (6) to the threshold value (10). We found that the estimated number was less as compared to the threshold value (10). So, P1 was rejected. Using the same method, we summed P2 to get 9. This value was again below the value of 10. So, it was also rejected. We applied the same procedure for all the included studies. The same resolution method was applied to a few real studies and the chosen two articles to show the proposed table process.

For example, we eliminated Aloraini and Nagappan (2017) study on 'evaluating buffer errors in android mobile apps.' Furthermore, their planned study did not correspond to any of our pre-set research questions. As a result, their research did not even add to the theoretical support of this writing and was unable to provide apps assessment-related facts that may become supporting evidence for an app assessment tool. That is why we did not include it. Importantly, we did not generate any scoring value using the Conflict Resolution Table (CRT) because we already knew that none of our research aims was aligned with the cited paper. We followed the same strategy for each of these kinds of studies.

On the other hand, we agreed with Kolak et al. (2021) proposed study on both qualitative and quantitative results on the topic of 'evaluation of preschoolers app in light of educational potential.' In their investigation, we discovered a number of notable apps evaluation criteria that align with our research problem and readily fit into our three proposed classes for app evaluation parameters. That allowed us to gain a deeper understanding of all current apps assessment parameters. We also quantified the same research study using a resolution table (see Table 4). We (all participants) decided early on that if a study highlighted sound evaluation standards and likely LAs concerns, we would award it a 5 numerical value which would be the highest score offered by anyone. Therefore, the study we were evaluating got 15 points from all three participants and it was accepted with 15 points (averaged score of all the participants).

RESULTS

This section discusses recent developments in the field of learning apps evaluation, such as rubrics, checklists, frameworks, models, rating-systems, websites, and other such assessment tools. Meanwhile, we discuss the state of the arts, emerging research gaps, some existing apps appraising methods, and a few other recent research dimensions on the same research problem. Moreover, a sincere effort has been made to respond categorically to the venerable research questions in light of the findings from a systematic review of the methodically chosen 114 studies in the ways that are described as follows.

APPRAISING TECHNIQUES, INSTRUMENTS, AND OTHER TOOLS FOR THE LEARNING APPS EVALUATION

We investigated five important educational apps evaluation methods based on a comprehensive analysis of the relevant literature: (i) rubrics, (ii) checklists, (iii) frameworks, (iv) reviews and rating systems, and (v) random searches (see Figure 2). Approximately 70 of the 114 most relevant papers contained at least one evaluation instrument or method for educational apps appraising. The following is a description of each in detail.

We discovered a total of 31 rubrics on LAs, with the most rubrics proposed in 2012 (5), followed by 2013, 2014, and 2015 (4 rubrics contributed) (see Figure 3). The most well-known rubrics proposed by various authors are as follows: in 2011, the two most well-known rubrics were proposed by Harry Walker (2011) (a rubric for quality app evaluation and a rubric for iPad evaluation). The following year (2012), most rubrics were proposed as 'the rubric for app evaluation', 'rubric for mobile apps evaluation', 'rubric for iPad app evaluation' (Tolisano, 2012), 'rubric for educational app evaluation' (Vincent, n.d.b), and 'rubric for apps assessment.' In the next year (2013), 'rubric for iPad apps', 'a rubric for technological education', 'an evaluation rubric for mobile apps' (Walker, 2011), and 'rubric on special apps evaluation' (Malone & Peterson, 2013) were the most popular assessment instruments. In the same way, 'a rubric for special learners' (Bentrop, 2014), 'rubric for language learning' (Martín-Monje et al., 2014), 'MASS rubric' (Green et al., 2014), and rubric for selection of m-Apps were the most influenced rubrics of the year 2014.

Next, in 2015, there were four rubrics proposed by multiple authors including a *rubric for LAs evaluation* (C. Y. Lee & Cherner, 2015), *rubric for apps evaluation* (Weng, 2015), *rubric for systematic evaluation of literacy apps* (Israelson, 2015), and *rubric for educational apps* (J. S. Lee & Kim, 2015). Moreover, in 2016, three pertinent rubrics were proposed: *rubric for assessing the quality of teacher apps* (Cherner et al., 2016), *rubric for students with LD*, and *rubric on language learning evaluation* (OK et al., 2016). The *rubric for preschoolers educational app* (Kalogiannakis & Papadakis, 2017), *rubric for phonemic apps* (Lisenbee, n.d.), *REVEAC* (Papadakis et al., 2017), *rubric for educational m-Apps* were the main rubrics of the year 2017. In a similar fashion, we revealed the remaining ones.

From 2011 to 2022, the most relevant frameworks on LAs evaluation were: (i) m-app usability (Tahir & Arif, 2014), (ii) TPACK (Handal et al., 2014), (iii) systematic evaluation of literacy apps (Israelson, 2015), (iv) RETAIN (Campbell et al., 2015), (v) framework on a selection of m-Apps, (vi) evaluation framework on smart learning (J. S. Lee & Kim, 2015), (vii) assessment tool for m-Apps (Hassen, 2016), (viii) framework for quality m-Apps assessment (Baloh et al., 2015), (ix) teachers' app-evaluation criteria (Baran et al., 2017), (x) framework on evaluating language apps (Rosell-Aguilar, 2017), (xi) assessment for STEM educational apps, (xii) appraising framework for educational apps (Kay, 2018a), (xiii) evaluating mathematics apps (Kay, 2018a) (xiv) framework on evaluating thinking apps (T. Chen et al., 2019), (xv) framework on assessing paid mobile apps (Mneumann et al., 2019), (xvi) framework for evaluating LAs for young students (Papadakis, 2021), (xvii) apps evaluation tool for Kindergarten (Kalogiannakis & Papadakis, 2017), and (xviii) preschool evaluation tool for children (Kolak et al., 2021).

We found only six most relevant studies on checklists, including (i) evaluation of iPad apps (Schrock, 2011), (ii) Harry Walker's apps evaluation checklist, (iii) checklist on educational software (Boone & Higgins, 2012), (iv) great checklist for educational apps, and (v) app checklist for educators (Lubniewski et al., 2017). Consequently, we included approximately 18 studies on reviews and rating systems for app evaluation, as good app guides and common-sense media (Taylor et al., 2022), ways to evaluate educational apps (Vincent, 2014), APPitic website (Mobile App Daily, 2021), Apps in Education (Swanson, n.d.), common sense media (Common Sense Media, n.d.), app-ed review of educational apps, Children's Technology Review (http://www.childrenssoftware.com/). Mobile app rating system (Stoyanov et al., 2015), perfecto rating tool (Perforce, 2021), mobile app selection (Sarrab et al., 2015), apps classification (Maalej et al., 2016), finding quality apps, crowdsourced source evaluation process (Khan et al., 2017), apps evaluation factors (Zeng et al., 2017), were the best application for instructors and education (Dove & Revilla, 2021).

Finally, 10 studies were chosen for 'Random Searches' for educational apps evaluation purposes. Such studies are the top 10 learning apps in India on the Google app store (Mobile Action, n.d.), the 5 best online LAs (91mobiles.com., 2021), the best educational apps, the top ten educational apps in India (Harsh, 2018), the top ten learning apps for preschoolers (Udayan, 2022), the top 25 free learning apps (Team Leverage Edu, 2021), the top 21 education apps in India for online learning (S. Sharma, 2016), the top 10 educational apps of 2021 (Mobile App Daily, 2021), and the highest rating educational apps on app stores (https://www.educationalappstore.com/). The pictorial representation of the aforementioned tools is shown in Figure 2.

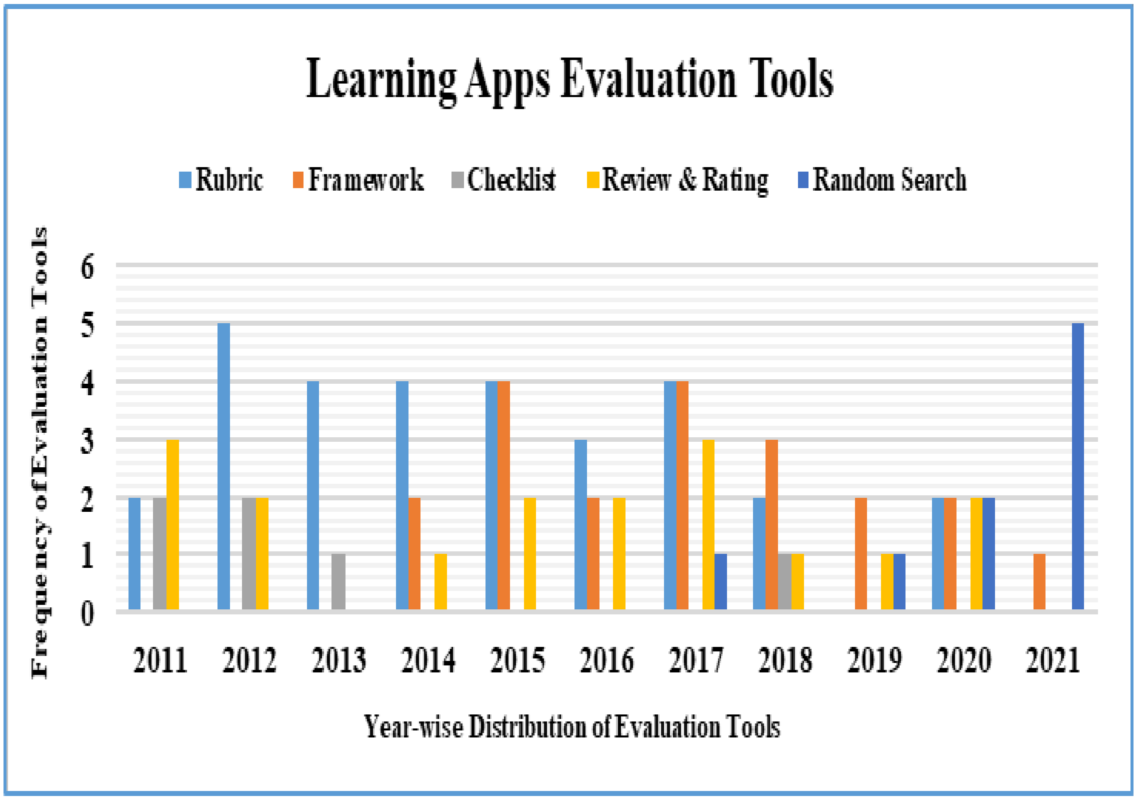

Figure 2. Distribution of learning apps evaluation tools from 2011 to 2021

Figure 2 shows a clear trend in how apps are being evaluated. In the graph, the horizontal line (x-axis) represents the year-wise distribution of learning apps evaluation tools while vertical bars represent their frequencies from 2011 to 2021. According to the research, most of them are rubrics (31), followed by frameworks (20), and then checklists, which have the least impact on app evaluation. The frequency of random searches was absent in the years 2011, 2012, 2013, 2014, 2015, 2016, and 2018, but the contribution of it was 9 in number. Despite a significant number of these tools being available, only a few of them are truly beneficial for app evaluation. In most cases, the methods used to evaluate apps are not scientific, such as those used in app reviews and ratings, or those based on random searches. Many unscientific rubrics also exist in the literature that do not effectively assess the apps. A list of most influencing apps evaluation mechanisms includes Walker (2011), C. Y. Lee and Cherner (2015), Israelson (2015), J. S. Lee and Kim (2015), Papadakis and Kalogiannakis (2017), and more. It is widely believed that Walker's (2010, 2011) tool forms the foundation for the majority of other app evaluation tools.

Based on the aforementioned pieces of evidence, we have conducted a more in-depth tabular analysis of the most important relevant studies, taking into account four key criteria, such as the types of tools utilized (rubric, checklist, framework, website review, or random search), as well as the primary criteria used to evaluate the apps instruments (such as what are the most important aspects of the study to evaluate). The second criterion was a tool's 'major characteristic' which highlighted its most salient features. In addition, there are four criteria that make up the criterion dimension: page length (how long an app's assessment tool is), sub-criteria (whether or not the assessment tool has additional criteria beyond the primary ones), scoring criteria (which instruments they used; if unavailable, we deemed it to be Not Available (N/A)), and target audience (focusing domain of selected tool). The last one was a set of extremely important findings or remarks (Table 4). Note that we have also referenced a number of pertinent observations in other sections of this writing.

Table 4. Critical research findings of existing assessment tools for learning apps

Tools Type	Major Criteria	Major Characteristics	Main Findings/Remark(s)
Rubric Walker, 2010, 2011	Curriculum links, authenticity, feedback, differentiation, user-friendliness, and motivation.	Page length: 1 Sub Criteria: No Scoring Criteria: 0 (lowest) to 4 (highest) Targeting: N/A	Validation was absent from the proposed Rubric. Nonetheless, a majority of app evaluators viewed it as the foundation for any other product.
Rubric Vincent, n.d.b	Relatedness, individualization, feedback, usability, thinking skills, engagement, and sharing.	Page length: 1 Sub Criteria: No Scoring Criteria: 1 (lowest) to 4 (highest) Targeting: N/A	They recommended evaluation criteria that were nearly identical to those provided by Walker (2011). Nonetheless, they highlighted shareability and critical thinking as two new app features.
Rubric Tolisano, 2012	Considerations, contents, logistics, fluency, substitution, evidence	Page length: 1 Sub Criteria: 21st century skills, Bloom's taxonomy, multiple intelligence skills, differentiation, authenticity, curriculum connection, inappropriate content, user-friendliness, advertisements, and some more. Scoring Criteria: N/A Targeting: N/A	They listed a vast number of assessment criteria and did not explain their validation approach.
Rubric C. Y. Lee & Cherner, 2015	They created a rubric based on 24 assessment criteria (see literature review), focusing on three broad instructional dimensions (instruction, design, and engagement).	Page length: 8 Sub Criteria: each dimension has approximately 8 such criteria (see literature review section). Scoring Criteria: 5 (1 to 5) point Likert Targeting: general	C. Y. Lee and Cherner (2015) created their rubric with a strong theoretical foundation and a set of widely used dimensions. However, their rubric is plagued by two unresolved issues: generalization and usability.

Tools Type	Major Criteria	Major Characteristics	Main Findings/Remark(s)
Rubric Papadakis et al., 2017	Framed by four broad evaluation parameters: content, design, functionality, and technical.	Page length: 3 Sub-Criteria: their study has approximately 18 sub-criteria (cited in the literature review section). Scoring Criteria: 4 (1 to 4) point Likert Scale. Targeting: preschools	The REVEC rubric has been deemed effective, though it does have certain drawbacks in the context of inter-rating, reliability, and sample size.
Framework Israelson, 2015	Framed by four prominent evaluation parameters: multimodal, learning content, navigation, and engagement.	Page length: 1 Sub Criteria: N/A Scoring Criteria: 4 (1 to 4) point Likert Scale. Targeting: elementary	The proposed framework was conceptualized on a theoretical foundation, and then applied to educator literacy apps. Despite this, the proposed assessment tool was extremely time-consuming and had usability challenges.
Framework J. S. Lee & Kim, 2015	Teaching and learning, design, technology, and economy and ethics.	Page length: 2 Sub Criteria: contained 8 sub-criteria (see literature review section). Scoring Criteria: 4 (1 to 4) point Likert Scale. Targeting: general (smart learning)	The proposed study revealed four theoretical parameters to ascertain smart learning apps. Of them, the technology dimension bare minimum and more focused on instructional parameters. For assessment purposes, they have included only games targeted to LAs and included a short sample size. Therefore, further adjutant is needed.
Framework Rosell-Aguilar, 2017	Covered four pertinent dimensions (technology, pedagogy, user interface, targeted subject).	Page length: 1 Sub Criteria: contained a large list of sub-criteria (see literature review section) Scoring Criteria: N/A Targeting: general (language learning)	Rosell-Aguilar (2017) designed his evaluation framework based on two things: apps taxonomy and four quality assessment criteria. There was no detailed definition of the proposed sub-criteria, and they did not appraise any apps themselves.

Tools Type	Major Criteria	Major Characteristics	Main Findings/Remark(s)
Framework Kay, 2018a	Worked on eight parameters: learning value, quality, goal, engagement, usability, individualized, feedback, and teamwork.	Page length: N/A Sub Criteria: N/A Scoring Criteria: N/A Targeting: general	First and foremost, researchers advise identifying the types of apps that a specific learner wants. Following that, they evaluated that app based on the suggested app's features. They did not carry out any validation procedures.
Framework Kolak et al., 2021	Proposed 2 appraising tools containing 12 and 5 items respectively.	Page length: 2 Sub Criteria: available Scoring Criteria: 0 to 4 scoring technique Targeting: preschoolers (math and literacy apps)	This was the first research that attempted to address the app-gaps issue. Meanwhile, they conducted an experiment comparing free and paid apps and found no significant difference in their pedagogical efficacy.
Website Reviews Taylor et al., 2022	They have done their assessment based on 10 criteria: learning goal, meaningful instruction, problem-solving, feedback, interaction, exploration, storyline, language quality, individualization, and design features.	Page length: 1 Sub Criteria: N/A Scoring Criteria: 5-point Likert Scale Targeting: preschoolers	The primary goal of this writing was to evaluate 39 preschool apps (both low and high ratings). Furthermore, they discovered that the apps with higher ratings have greater educational potential than those with lower ratings. However, all of the mentioned apps had lack instructional potential in general. As a result, more research on what should be contained in true LAs is required.

We analyzed a small number of 'really-good' assessment tools that have been designed for educational app appraisal, if any exist at all, with unbalanced assessment criteria. Some of them included an excessively long list of criteria, while others preferred a fairly small number of such assessment characteristics. In a similar vein, Papadakis et al. (2017) stated that when an assessment tool has a large number of such items, it may become dysfunctional, and when it contains a small number of such things, there is a risk of insufficient evaluation of a specific object. Therefore, such claims are rather debatable (Rosell-Aguilar, 2017). Furthermore, the majority of previously employed assessment criteria are not scientifically matched with new ones (C. Y. Lee & Cherner, 2015) and are not validated at the research level (Kolak et al., 2021; Rosell-Aguilar, 2017). So yet, it is unknown how many parameters should be included in a sound app assessment tool. Another long-standing difficulty is determining which average rubric score should be considered ideal. We did not find any information on this; however, we explored a lot of research that stated that a solid average score should be calculated to identify an app as adequate. In addition, the educational brotherhood is persistently pressing for standardized testing procedures that adhere to established linguistic standards and share a consistent organizational framework for such assessment instruments. Therefore, we (software developers, teachers, parents, and other professionals) need to work side by side, evolve an optimum set for effective evaluation, and put in a more rigorous effort to ensure our students' bright future.

EXISTING CRITERIA FOR LEARNING APPS EVALUATION

To adequately address this research question, we conducted a thorough and critical examination of the most relevant studies on educational apps evaluation. We revealed numerous scientific and non-scientific appraising tools for educational apps. As previously stated, we identified a total of 70 studies on app evaluation (see Figure 2), and from these studies, we investigated a set of viable critical evaluation criteria for educational apps. We divided them into three broad evaluation categories: Technology, Pedagogy, and Contents. It should be noted that all of the sub-criteria for these three dimensions were not mutually exclusive. So, they could be part of more than one dimension. In the end, we deleted all duplicate sub-criteria from mentioned categories.

The Technology dimension included the following 18 sub-criteria from TC1 (Technological Criteria 1) to TC18 (Technological Criteria 18) as user support, accessibility, design, purpose, usability, stability, portability, multimodal options, functionality, communication, performance, gamification, interoperability, navigations, working mode, design elements, dependency on technology, and social interactions (Baran et al., 2017; Bentrop, 2014; T. Chen et al., 2019; Green et al., 2014; Israelson, 2015; R. Kay et al., 2019; Lubniewski et al., 2017; Martín-Monje et al., 2014; McQuiggan et al., 2015; Reeves, 1994; Rosell-Aguilar, 2017; Schrock, 2011; Tahir & Arif, 2014). The frequently cited evaluation criteria were technology accessibility, design, usability, functionality, gamification, and communication regarding category A.

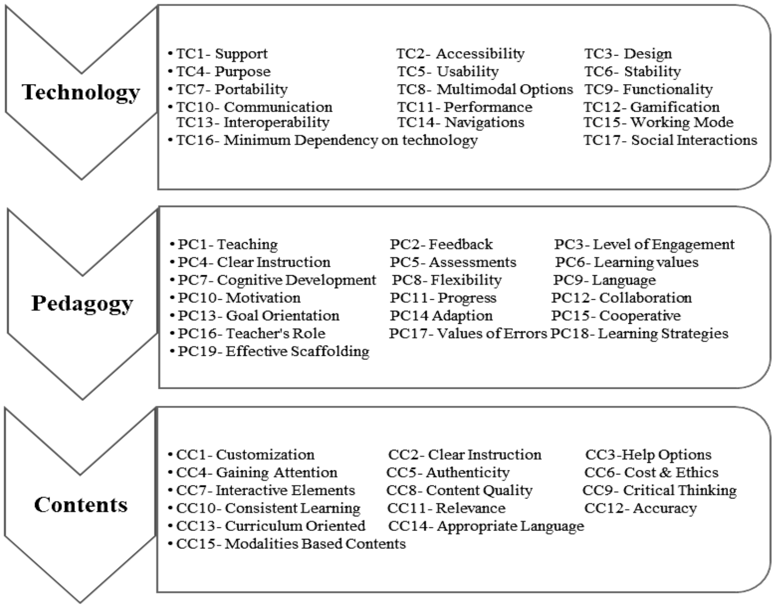

Figure 3. Categorization of evaluation criteria for learning apps

On the other hand, we refined some of the pertinent evaluation criteria for the Pedagogy category from PC1 (Pedagogy Criteria 1) to PC19 (Pedagogy Criteria 19) as follows: teaching, feedback, level of engagement, clear instruction, assessments, learning values, cognitive development, flexibility, language, motivation, progress, collaboration, goal orientation adaption, cooperative, teacher's role, values of errors, learning strategies, and effective scaffolding. Note that feedback, learning values, motivation, communication, and effective instruction were deemed the most significant evaluation criteria for the apps assessment as compared to others. In the last, we sum up some of the most effective appraising criteria concerning the Content dimension. The list of such criteria is shown as follows: customization, clear instruction, help options, gaining attention, authenticity, cost and ethics, interactive elements, content quality, critical thinking, consistent learning, relevance, accuracy, curriculum-oriented, appropriate language, and modalities based contents (Baran et al., 2017; Bentrop, 2014; X. Chen, 2016; Cherner et al., 2016; Handal et al., 2014; Israelson, 2015; Kay et al., 2019; J. S. Lee & Kim, 2015; McQuiggan et al., 2015; More & Travers, 2013; Papadakis et al., 2020; Wang et al., 2019).

The selection of evaluation criteria was found to be skewed in its uniformity. It all came down to what kinds of apps were being assessed. As a result, there was not an even single framework or another similar tool available that outlined consistent evaluation criteria for general educational apps. Therefore, we have identified some of the most important evaluation criteria that may become essential apps ingredients for good educational apps, such as accessibility, design, usability, functionality, gamification, communication, feedback, learning values, motivation, effective instruction, customization (Kolak et al., 2021), interactive approaches, curriculum-oriented learning elements, and support critical thinking. Here, we explained the basic meaning of these evaluation criteria in order to gain a better understanding of them. We defined app accessibility as being available at any required source to access it on our smartphone and allowing all required features on our mobile phone with the least amount of cognitive effort. There should be a greater emphasis on aesthetics and user-friendliness when it comes to app design. The app's usefulness should be evaluated by individual users in real-time, based on principles of total usability (Tahir & Arif, 2014). Understanding, learnability, operability, aesthetic-ability, and effectiveness are some of the characteristics of software components defined by certain academics (Baloh et al., 2015).

App functionality should be very effective during app usage periods. If we can use an app intuitively, it means we can take advantage of all of its features (Cherner et al., 2016). Gamification is the process of incorporating game-like features into mobile apps. Effective communication (to the point, clear, and focused) is required throughout the app's usage. Because of the paradigm shift, feedback attributes received the most citations in online learning. The highest priority in any learning environment is to motivate the learners. So, with the help of pedagogic feedback provided by their instructors to students, it is possible to foster motivation while improving student performance (Vincent, n.d.b). The majority of the apps claim to be in the educational domain, but they appear to have a true learning content problem. To address such a problem, our apps should include curriculum-based learning content that is goal-oriented, interactive, and up to date. Learner customization options are another important evaluation criterion. Every learner has a different learning strategy for their learning which is why there should be options for learner customization so that they can set their learning pace according to their individual needs. We have reviewed many educational apps but, unfortunately, they have not been developed in accordance with their curriculum or other learning strategies. Critical thinking features, such as raising questions, evaluating the given information, openness, etc., were missing in most of the apps.

Emerging Gaps Concerning Learning Apps Design and Evaluation

Even though a lot of research has been done in this field to illustrate the importance of virtual learning in today's world, there is still a lot of room for its improvement. The benefits of virtual learning include the provision of a user-friendly learning environment, self-paced learning, moveable learning, suitable for current era's learners, and more. In addition to all of these effects, the current research domain faces a number of concrete research challenges. For simplicity, we categorized all the research gaps into three broad categories: general, design, and evaluation. The list of the emerging research challenges is shown in Table 5.

Table 5. Emerging issues in the field of learning apps

General	Design	Evaluation
knowledge integration	app overwhelmed issue	evaluation parameters
digital tech-savviness	poor quality learning apps on app stores	in-depth evaluation
awareness of emerging LA paradigms	App-mashing problem	theoretical foundation
training on evaluation tools and techniques	self-proclamation	quantification features
	apps-gap	criterion subjectivity
		usability
		generalization
		classification
unregulated app market		

We have outlined three broad classes for current issues on learning apps: general issues contained the issue of knowledge integration, the issue of tech-savviness (over 50% of the population is tech-illiterate), and the involvement of government (T. Chen et al., 2019; Israelson, 2015). Moreover, the current apps market is considered an unregulated market (Taylor et al., 2022) because developers are just seeing their profit only and they are just pushing the newly developed apps on the app store in the educational category without appraising about their quality. This scenario is known as an unregulated market. An excellent learning app developer knows the subject matter well. Sometimes app developers met this knowledge criterion but even they could not appropriately incorporate the knowledge experience into desired apps; this is termed as an issue of 'knowledge integration.

Next, the design issues include universal design problems and issues of apps-gaps; it is how social disadvantage becomes a learning obstacle through free and premium learning apps (Kolak et al., 2021). But recently, it was found that free and paid educational apps are almost similar in terms of their learning potential (Kolak et al., 2021). Next, learners sometimes use a particular app to accomplish specific assignments or projects but, sometimes they need another resource(s) to complete that particular task (Rosell-Aguilar, 2017). In other words, how this software is utilized with other learning substantial remains an emerging research gap and is still under-researched (T. Chen et al., 2019; Goodwin & Kucirkova, 2012; Handal et al., 2014; Hirsh-Pasek et al., 2015; Israelson, 2015; Kay, 2018b, 2018a; Papadakis & Kalogiannakis, 2017; Rosell-Aguilar, 2017; Vaala et al., 2015).

From the early stage of this writing, our primary emphasis was on the evaluation dimension of educational apps. On the basis of this, we revealed a list of highly noticeable issues, some are listed as follows: the very first issue is regarding the long list of evaluation criteria (Kolak et al., 2021; C. Y. Lee & Cherner, 2015) of existing apps evaluation tools. As Kolak et al. (2021) stated, the majority of apps assessment instruments have approximately from 18 to 70 or more assessment parameters which may make apps assessment process impractical. On the same issue, Papadakis et al. (2017) stated that when an assessment tool has a large number of such items, it may become dysfunctional, and when it contains a small number of such things, there is a risk of insufficient evaluation of a specific object. Therefore, we need a consensus assessment tool that has optimum assessment parameters. However, teachers, parents, and other caregivers are eagerly anticipating the development of such scientific tools, but their eyes are aching to witness a miracle that has not yet occurred.

Moreover, we also noticed an issue with a lack of in-depth evaluation (Kolak et al., 2021). Numerous app evaluation methods are unable to evaluate the selected app in every respect. Next, the lack of theoretical support was a common theme that emerged from the research (Hirsh-Pasek et al., 2015; Kolak et al., 2021; J. S. Lee & Kim, 2015) suggesting that the vast majority of currently available apps are not developed in accordance with sound learning theories and failed to adequately incorporate child cognition strategies. During this analysis, the issue of quantifying app features is also identified (Kolak et al., 2021; J. S. Lee & Kim, 2015; Papadakis et al., 2020 This means that when individuals use a certain app, they are typically unaware of the number of times they have utilized a particular app feature(s) throughout their app usage journey. If this is possible, then it will be easy to develop a good app assessment instrument that will contain all the necessary apps features. Next, numerous studies have found that evaluating apps is more challenging due to their subjective criteria (J. S. Lee & Kim, 2015; Shoukry et al., 2015) provided by diverse app users, which may in turn raise concerns regarding user biases (Kolak et al., 2021).

Even more, we have analyzed several other assessment instruments that are currently available, but of them, some are unscientific (Vincent, n.d.a), whereas a few of them are researched-based (scientific) (Israelson, 2015; Papadakis & Kalogiannakis, 2017; Walker, 2011). However, their usage of the so-called generality issue is restricted (C. Y. Lee & Cherner, 2015). For example, some are only applicable to a few preschool applications, while others are responsible for LD apps, and still others are created to evaluate a few LAs. Meanwhile, we revealed a few assessment methods that have been proven across a variety of learning apps (Baloh et al., 2015; Baran et al., 2017; Kolak et al., 2021), such as REVEC (Papadakis et al., 2017), MASS (Green et al., 2014), rubric for mobile apps and rubric for special apps evaluation (Malone & Peterson, 2013). Furthermore, these investigations have been validated on 42, 22, 21, and 21 educational apps, respectively.

Another noticeable concern was identified, which is subjective or technical feedback offered by the app users after utilizing a specific app. Therefore, it is probable that other individuals who intend to use that app on their portable devices may find it difficult to comprehend the comments offered by these users. So, while reviewing an app for learning purpose, aim to provide clear, concise, and technically sound comments (feedback) so that anyone is able to understand the already provided ratings by app users in assisting desired education apps. Consequently, if we are unable to adhere to the proper and transparent review criteria, it may provide some additional research issues, such as the wrong categorization of educational apps. Therefore, we must organize these app classifications into a simple, concise, and manageable format to facilitate the placement of new apps in their proper place to improve the app selection process.

Thousands of new educational apps are added to the various app stores every day. To this point, developers have windthrown a set of apps from a specific app store and then uploaded the same set of apps under a different name. Some of the apps in this collection are free, while others need payment. However, currently, it is difficult to say that they can promote desirable learning content to their apps consumers and properly fit inside the true category of educational app stores. This is because there are fewer good apps appraising tools available, and if any somehow exist, they suffer from a flawed experimentation procedure. Furthermore, the length of their evaluation parameters is becoming a nuisance for researchers. As a result, there is an urgent need to develop well-designed assessment instruments with sufficient capability to measure the learning potential of these apps using an appropriate set of evaluation criteria so that teachers, parents, students, and other stakeholders can feel at ease with this awkward apps section procedure for the sake of our society's future.

DISCUSSION AND FINDINGS

After critically analyzing the relevant literature on learning apps evaluation, we have discovered compelling evidence regarding the overwhelming volume of such apps on the current app market (android, apple, and others). In addition, around half of them are targeting preschoolers. In terms of their specific educational discipline, language learning apps were reported to be the most popular ones, followed by math apps. The majority of included studies defined their learning apps as positively significance in terms of their learning outcomes and the need for their integration in current virtual learning settings but choosing the right one is still becoming an emerging issue for the researchers' fraternity. However, we observed some of the most pressing general concerns that parents, teachers, and other caregivers are facing in regard to the selection of future educational apps. Such concerns include app evaluators having adequate knowledge of their subject domain, being concerned about app rating scores (Taylor et al., 2022), their price because free apps are more affordable than paid ones, platform (whether selected apps are workable or not on a specific platform), and privacy (whether this app is free of any data leak problem or not) (Kay, 2018a).

Thus, the primary concern of the study was to identify (inform) useful assessment tools or techniques on the behalf of mobile learning apps. However, the ultimate purpose of the same research design was to familiarize educators, apps developers, and other related stakeholders with the current trends of these apps. In this context, we formulated three pertinent research questions designed to show the ongoing trends regarding apps assessment tools, their prominent assessment parameters, and a list of the most significant research concerns to whom educational apps are confronting.

Based on the analysis, we revealed five foremost apps appraising instruments: rubrics, frameworks, checklists, website reviews, and random searches. We also found a list of the most cited evaluation parameters used by these appraising instruments. Such parameters include learning objectives, feedback, adequate communication, individualization, progress monitoring, motivation, user-friendly design, options for user support, curriculum connection, cognitive development learning contents, usability, and more. We also observed their evaluation parameters cited patterns in several included studies and found that the majority of them have been designed with dissimilar assessment parameters. As Taylor et al. (2022) claimed, the majority of apps appraising methods have contained roughly of 18 to 70 or more appraising attributes. Based on this argument, we may guess the parameters divergence (mismatching) problem and found approximately 52 dissimilar apps appraising parameters.

Regarding this, C. Y. Lee and Cherner (2015) stated that the app assessment criteria of previously conducted studies are not scientifically matched with new ones. People do not have so much time to understand only the working procedure of such evaluation tools that have been designed by a large set of dissimilar attributes. Similar to previous arguments, Papadakis et al. (2017) investigated that when an assessment tool has a large number of appraising parameters, it may become dysfunctional, and when it contains a small number of such items, then there is a risk of insufficient evaluation. But researchers are continuously working on the same issue and knowing the optimal set of such appraising criteria becomes a debatable concerned (Rosell-Aguilar, 2017). In light of this, we have categorized all of the cited evaluation parameters into three concise and manageable categories: technology, pedagogy, and design. In recent years, a new concern is emerging which is 'learning ergonomics'. Experts rightly consider this an indispensable dimension of the virtual learning environment. But, to our surprise, hardly no one could include it as a necessary criterion (possible ergonomics parameters) as part of the app evaluation attributes in their designed assessment instruments. Hence, we need to design a concise (consider all necessary assessment criteria from all crucial dimensions - technology, pedagogy, design, and ergonomics), concrete (explainable), and theoretically sound appraisal tool for learning apps that can properly assist (guide) our youngsters, teachers, and other caregivers in selecting the appropriate one.

We have shown that the app development process is currently proceeding at a dizzying rate, and such apps are increasingly being published (added) in the education sector of selected app stores without any defined procedures. Apps that were originally designed for altogether a different purpose (not for education) may find their way into the app stores in the educational category. What we have here is an unchecked, uncontrolled, or unstandardized market for instructional apps. It was also brought to people's attention that the same app could be located in both the free and premium sections of the same app store or a distinct store. Because of this, it is not uncommon for the same app to be submitted multiple times under different names amongst the app stores. These unfavorable conditions render the apps selection process more ineffective. To avoid this undesirable scenario, app makers and educators alike should standardize the app-uploading process by which inappropriate apps are kept out of the educational app store and duplicate copies of the same app are prevented from being distributed. If it happened, hopefully, teachers and parents might feel more comfortable than earlier. However, we could not find a set of studies that have been designed to resolve such uncommon issues. Therefore, more research is needed in this direction.

Following thoroughly in this study, we uncovered some crucial information about the app validation process. The majority of studies, we discovered, have not been validated or gone through a specified assessment instrument at the research level (Kolak et al., 2021; Rosell-Aguilar, 2017). If some of them did, they have only tested (validated) it (assessment tool) on a limited number of apps to confirm their validity (Baloh et al., 2015; Baran et al., 2017; Kolak et al., 2021). This issue is still a concern because the bulk of validated studies had their findings verified by the appropriate teachers inside official classroom settings. However, a majority of real apps users (apps consumers) belong to informal groups, i.e., outside of the official classroom settings, meaning that they use the apps informally. Unfortunately, nobody has ever designed a research study that could focus on the validation of their apps targeting real apps consumers (Rosell-Aguilar, 2017). Therefore, in order to truly validate our designed learning apps, we need more in-depth systematic reviews that can uncover some insightful information from these real consumers of apps.

Irrespective of the above discussion, we have summarized a set of viable key insights based on deep analysis. A list of such findings is shown as follows:

- Numerous research studies have shown that learning apps can be integrated into the current educational system, because they are providing endless benefits to their consumers, such as providing options for constructive learning, supporting mobility, fostering more sustainable and authentic learning options, emphasizing individualization of learning options, and so on.
- The relevant literature identified three foremost frameworks regarding apps appraising in terms of their popularity as one proposed by Papadakis et al. (2020). They highlighted navigation, age-appropriate content, learning feedback, user interface, and learner-centered design as the most influential app evaluation criteria.
- Several researchers viewed Walker's Rubric (Walker, 2011) as the foundational model for the majority of app assessment tools, which has six pertinent evaluation criteria including curriculum connection, content authenticity, learner feedback, individualization learning approaches, learner-centered contents, and contained motivation-like elements.
- Based on a critical investigation of the relevant literature, we have divided all of the apps assessment criteria into three categories: (i) technology, (ii) pedagogy, and (iii) contents. The foremost parameters of the technology domain include accessibility, gamification, performance, communication, navigation, usability, and interaction. Similarly, the crucial evaluation criteria of the pedagogy category are clear instruction, goal orientation, effective scaffolding, learner feedback, collaboration, and learning assessment. The final category contained customization, learning assistant, content quality, interactive elements, and adequate language. However, highly pertinent evaluation criteria in this age of inevitable technologies related to learning ergonomics were missing.
- After a critical examination of the included studies, we found that only four categories of learning apps are predominantly evaluated: apps for preschoolers, language learning apps, early mathematics apps, and certain applications for special education.
- According to the literature, there are just a few proven methods for assessing the quality of learning apps. If any of them exist, the majority of stakeholders do not know; how to use them for apps assessment.
- In the literature, there exist tendencies of unscientific app assessment tools that are almost incapable of evaluating educational apps. Although a small number of scientific tools exist in the literature, they appear to suffer from several flaws.

In light of these research results, we can assume that there are negligible general models available for learning apps appraisal. So, the need is to make a sound tool for app evaluation, especially for educational apps. This kind of assessment model should be based on adequate learning theories and child development tendencies that provide clear instructions for each app assessment parameter. Additionally, they must be validated at the research level by real apps consumers.

PROSPECTIVE IMPLICATIONS

Like any systematic literature review, we not only looked at what research has already been done on a topic, but also evaluated, summarized, compared, contrasted, and correlated different scholarly gains and other relevant sources that are directly related to current research. There are three ways to examine the impact or implications of this study: from a research, theoretical, and practical standpoint. The wide-ranging impact of the key findings on decision makers and future research is the subject of research implications. While theoretical is about contextual validation through methods of support, denial, advancements, or potential influence on research/developments, implications for practice immediately relate to their efficacy and efficiency. Moreover, prospects from all standpoints are visible concerning the existing apps assessment domains technology, pedagogy, and contents, and has been mentioned earlier. This contribution is highly likely to provide valuable support towards all three standpoints and the three domains as well, which are depicted symbolically in Figure 4.

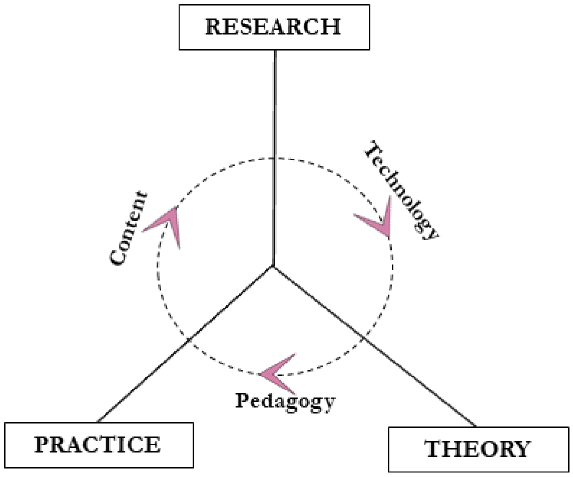

Figure 4. Three-perspective of learning apps research

It is envisaged that the contribution will support current and future researchers towards building quality LA and evaluation of the same leading to typical consolidations and standardizations. In turn, it may lead to constructing a cohesive, compact, and workable apps assessment model by taking necessary app assessment parameters while developing improved instrument(s) of evaluation; for example, which set of criteria is the most frequently stated (including essential ones), or which set of important criteria is missing (learning ergonomics). On the other side, app developers may benefit considerably from the current writing while designing and building LAs, using the basic concepts of the fundamental app ingredients (as shown in Figure 3) that might be part of a good LA. Furthermore, it may aid designers as well as educators in developing an effective or well-balanced app assessment instrument for these apps based on good theoretical underpinnings. Most importantly, our work is deemed significant for primary stakeholders (students and teachers), as well as for secondary ones (developers, designers, evaluators, parents, and other caregivers) in several ways including the following:

- Standardizations of learning apps research methodology, design, developmental, and evaluation practices.;
- Upfront research on identified gaps and limitations – with the state of the art presented;
- Support a detailed investigation of the challenges encountered by apps users through the examination of the existing learning apps;
- Cause a paradigm shift by constant confrontation with numerous invalidated research, theoretical, and implementation practices; and

- Overall and eventually meliorate LAs concerns in the evolving context, perspective, and emerging scenarios.

CONCLUSION AND FUTURE WORK

The dizzy development of learning apps has also witnessed their widespread popularity, and it has become an integral part of today's learning system known as the Virtual Learning System. The consequences of this informal development concerning LAs have impacted everyone, including students and other stakeholders. Consequently, selecting high-quality and pedagogically sound apps for their students is still an emerging area of research. In favor of such a scenario, we conducted a systematic review study with three relevant research questions. Importantly, our findings show how important it is for developmental psychologists to collaborate with app creators to further the pedagogical potential of touchscreen apps. Additionally, we established a new structure for app evaluation parameters used by many researchers in their app evaluation procedure and classified them into three manageable categories.

In short, we revealed approximately 70 studies on our proposed research design that discussed at least one evaluation tool in their research proposal. However, the majority of them have to face serious flaws, such as weak theoretical support followed by a lack of a true validation process. Additionally, several other studies have been designed in an unscientific way, and some of them are unsuitable for practical use due to their lengthy list of assessment criteria and sometimes their ambiguous language. Despite having a long list of assessment parameters, evaluation tools typically contained the Not Applicable (N/A) option. It is true, as Rosell-Aguilar (2017) pointed out, that there cannot be any universally applicable evaluative framework for all educational mobile apps. Because these apps are designed to accommodate a wide range of needs of their users. Because of this, not all the LAs will benefit from the same set of assessment factors. This means that all evaluation systems for learning apps should provide a N/A option for better assessment purposes.

The app store industry is now thought of as an unregulated marketplace, where anyone may publish any number of apps to any app store without having to deal with any protocols. The most obvious results of these are that everyone is always occupied with the task of choosing between various apps and is often left feeling overwhelmed and bewildered by the sheer number of options available. Now is the time to develop an assessment tool or framework for these apps that is clear, optimal to effect, thorough towards the three dimensions of learning, validated, and user friendly. We should also care about the learning ergonomics new mandatary dimension regarding these apps. But the advancement in learning app development is miraculous, and it is difficult to predict where software and hardware will go next.

We are not alone in having problems with our study; other studies have them too. The small size of the sample used (10 databases) may be the most easily identifiable limitation of this writing as of now but certainly may not undermine the findings from what we have analyzed. Some very important studies may have been left out. Despite this, our search construction method was thorough, but if we tried them on different search combos with new search terms, it might be able to cover more relevant results. However, following that, we divided all of the app evaluation criteria into three distinct groups. Some criteria may have been used for more than one group. Furthermore, accurate categorization should be necessary. Therefore, we require additional work regarding the accurate categorization of all the current evaluation factors of educational apps. With this in mind, it is important to include all the criteria for evaluation that are required. These criteria should be objective and fair. Last but not least, it was underlined that this research incorporated publications about LA evaluations. A comprehensive literature review on the topic of learning apps design was, thus, warranted and has been carried out. Future research could focus on developing a well-validated, trustworthy, prescriptive evaluation framework with sound theoretic foundations for instructional applications.

References

91mobiles.com. (2021). *5 best apps for online learning and education in India.* https://www.91mobiles.com/hub/best-online-learning-education-apps-in-india/

Aloraini, B., & Nagappan, M. (2017, September). Evaluating state-of-the-art free and open source static analysis tools against buffer errors in android apps. In *2017 IEEE International Conference on Software Maintenance and Evolution* (ICSME) (pp. 295-306). IEEE. https://doi.org/10.1109/ICSME.2017.77

Agarwal, N. (2021, August 9). How mobile apps are transforming the education system. *YourStory.com.* https://yourstory.com/2021/07/how-mobile-apps-are-transforming-the-education-sys/amp

Baloh, M., Zupanc, K., Kosir, D., Bosnić, Z., & Scepanović, S. (2015, June). A quality evaluation framework for mobile learning applications. *Proceedings of the 4th Mediterranean Conference on Embedded Computing, Budva, Montenegro*, 280–283. https://doi.org/10.1109/MECO.2015.7181923

Baran, E. (2014). A review of research on mobile learning in teacher education. *Educational Technology and Society*, *17*(4), 17–32.

Baran, E., Uygun, E., & Altan, T. (2017). Examining preservice teachers' criteria for evaluating educational mobile apps. *Journal of Educational Computing Research*, *54*(8), 1117–1141. https://doi.org/10.1177/0735633116649376

Bentrop, S. M. (2014). *Creating an educational app rubric for teachers of students who are deaf and hard of hearing* [Masters of Science, Washington University]. https://digitalcommons.wustl.edu/cgi/viewcontent.cgi?article=1681&context=pacs_capstones

Bhatasana, D. (2020). *Mobile applications: A growing trend in the education industry. eLearning industry.* https://elearningindustry.com/mobile-apps-in-student-learning-growing-trend-education-industry

Boone, R., & Higgins, K. (2012). The software √-list: Evaluating educational software for use by students with disabilities. *Journal of Special Education Technology*, *27*(1), 50–63. https://doi.org/10.1177/016264341202700105

Brodsky, J. (2021). How blended learning can work best. *Forbes Magazine.* https://www.forbes.com/sites/juliabrodsky/2021/01/17/how-blended-learning-can-work-best/?sh=682ce8e91dc4

Buckler, T., & Peterson, M. (2012). Is there an app for that? Developing an evaluation rubric for apps for use with adults with special needs. *The Journal of BSN Honors Research, 5*(1), 19–32. http://hdl.handle.net/2271/1095

Campbell, L. O., Gunter, G., & Braga, J. (2015). Utilizing the Retain Model to evaluate mobile learning applications. In D. Rutledge & D. Slykhuis (Eds.), *Proceedings of the Society for Information Technology & Teacher Education International Conference* (pp. 732–736). Association for the Advancement of Computing in Education.

Chen, T., Hsu, H. M., Stamm, S. W., & Yeh, R. (2019). Creating an instrument for evaluating critical thinking apps for college students. *E-Learning and Digital Media*, *16*(6), 433–454. https://doi.org/10.1177/2042753019860615

Chen, X. (2016). Evaluating language-learning mobile apps for second-language learners. *Journal of Educational Technology Development and Exchange*, *9*(2). https://doi.org/10.18785/jetde.0902.03

Chergui, O., Begdouri, A., & Groux-Leclet, D. (2017). A classification of educational mobile use for learners and teachers. *International Journal of Information and Education Technology*, *7*(5), 324–330. https://doi.org/10.18178/ijiet.2017.7.5.889

Cherner, T., Dix, J., & Lee, C. (2014). Cleaning up that mess: A framework for classifying educational apps. *Contemporary Issues in Technology and Teacher Education*, *14*(2), 158–193. https://citejournal.org/volume-14/issue-2-14/general/cleaning-up-that-mess-a-framework-for-classifying-educational-apps/

Cherner, T., Fegely, A., Lee, C. Y., & Santaniello, L. (2016). A detailed rubric for assessing the quality of teacher resource apps. *Journal of Information Technology Education: Innovations in Practice*, *15*(1), 117–143. https://doi.org/10.28945/3527

Common Sense Media. (n.d.). *App Reviews.* https://www.commonsensemedia.org/reviews/category/app/genre/education-58

Cross, J. (2004). An informal history of eLearning. *On the Horizon*, *12*(3), 103-110. https://doi.org/10.1108/10748120410555340

de Almeida Biolchini, J. C., Mian, P. G., Natali, A. C. C., Conte, T. U., & Travassos, G. H. (2007). Scientific research ontology to support systematic review in software engineering. *Advanced Engineering Informatics*, *21*(2), 133–151. https://doi.org/10.1016/j.aei.2006.11.006

Dove, J., & Revilla, A. (2021). *The best apps for teachers and educators.* Digital Trends. https://www.digitaltrends.com/mobile/best-apps-for-teachers-education

Ebner, M. (2015). Mobile applications for math education – How should they be done? In H. Crompton & J. Traxler (Eds.), *Mobile learning and mathematics: Foundations, design and case studies* (pp. 20–32). Taylor & Francis. https://www.academia.edu/en/50444276/Mobile_Applications_for_Math_Education_How_Should_They_Be_Done

Edsys. (2017). *12 benefits of using apps in education.* https://www.edsys.in/12-benefits-of-using-apps-in-education

Falloon, G. (2013). Young students using iPads: App design and content influences on their learning pathways. *Computers and Education*, *68*, 505–521. https://doi.org/10.1016/j.compedu.2013.06.006

Flewitt, R., Kucirkova, N., & Messer, D. (2014). Touching the virtual, touching the real: iPads and enabling literacy for students experiencing disability. *Australian Journal of Language & Literacy*, *37*(2), 107–116.

Glomack, S. (2021). Current trends and future prospects of the mobile app market. *Smashing Magazine.* https://www.smashingmagazine.com/2017/02/current-trends-future-prospects-mobile-app-market/

Goodwin, K., & Kucirkova, N. (2012, March). *iTouch and iLearn: An examination of educational apps.* Paper presented at the Early Education and Technology for Children Conference, Salt Lake City, Utah, USA.

Graham, C. R. (2009). Blended learning models. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (2nd ed), pp. 375-382. IGI Global. https://doi.org/10.4018/978-1-60566-026-4.ch063

Green, L. S., Hechter, R. P., Tysinger, P. D., & Chassereau, K. D. (2014). Mobile app selection for 5th through 12th grade science: The development of the MASS rubric. *Computers and Education*, *75*, 65–71. https://doi.org/10.1016/j.compedu.2014.02.007

Griffith, S. F., Hagan, M. B., Heymann, P., Heflin, B. H., & Bagner, D. M. (2020). Apps as learning tools: A systematic review. *Pediatrics*, *145*(1), e20191579. https://doi.org/10.1542/peds.2019-1579

Gupta, D. (2022, August 19). Are mobile apps the future of the education industry? *Appinventiv.* https://appinventiv.com/blog/mobile-apps-a-growing-trend-in-education-industry/

Handal, B., Campbell, C., Cavanagh, M., & Dave, K. (2014). Appraising mobile maths apps: The TPACK model. In T. Sweeney, & S. Urban (Eds.), *Proceedings of the 26th Australian Computers in Education Conference* (pp. 251–269). Australian Council for Computers in Education.

Hannes, K., & Claes, L. (2007). Learn to read and write systematic reviews: The Belgian Campbell Group. *Research on Social Work Practice*, *17*(6), 748–753. https://doi.org/10.1177/1049731507303106

Harrison, T. R., & Lee, H. S. (2018). iPads in the mathematics classroom: Developing criteria for selecting appropriate learning apps. *International Journal of Education in Mathematics, Science and Technology*, *6*(2), 155–172. https://doi.org/10.18404/ijemst.408939

Harsh. (2018, May 9). List of top 10 best Android educational apps in India. *The Indian Wire.* https://www.theindianwire.com/education/best-android-educational-apps-india-56205/

Hassen, A. M. (2016). A comprehensive framework for design and evaluation of m-learning applications [Master's dissertation, Qatar University]. https://qspace.qu.edu.qa/handle/10576/5080

Herodotou, C. (2018). Young children and tablets: A systematic review of effects on learning and development. *Journal of Computer Assisted Learning*, *34*(1), 1–9. https://doi.org/10.1111/jcal.12220

Higgins, K., Boone, R., & Williams, D. L. (2000). Evaluating educational software for special education. *Intervention in School and Clinic*, *36*(2), 109–115. https://doi.org/10.1177/105345120003600207

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest*, *16*(1), 3–34. https://doi.org/10.1177/1529100615569721

Hong, Q. N., & Pluye, P. (2018). Systematic reviews: A brief historical overview. *Education for Information*, *34*(4), 261–276. https://doi.org/10.3233/EFI-180219

Hussain, A., Mkpojiogu, E. O. C., & Hassan, F. (2018). Dimensions and sub-dimensions for the evaluation of m-learning apps for children: A review. *International Journal of Engineering and Technology*, *7*(3.20), 291–295. https://doi.org/10.14419/ijet.v7i3.20.19168

InstructionalDesign.org. (2021). *Individualized learning.* https://www.instructionaldesign.org/concepts/individualized

Israelson, M. H. (2015). The app map: A tool for systematic evaluation of apps for early literacy learning. *Reading Teacher*, *69*(3), 339–349. https://doi.org/10.1002/trtr.1414

Kalogiannakis, M., & Papadakis, S. (2017, August). An evaluation of Greek educational android apps for preschoolers. *Proceedings of the 12th Conference of the European Science Education Research Association, Dublin Ireland,* 593–603.

Kay, R. (2018a, March). Creating a framework for selecting and evaluating educational apps. Proceedings of the *12th International Technology, Education and Development Conference, Valencia, Spain,* 374–382. https://doi.org/10.21125/inted.2018.0106

Kay, R. (2018b, October). Developing a framework to help educators select and use mobile apps in the classroom. *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Las Vegas, NV, USA,* 1315–1320.

Kay, R. H., & Knaack, L. (2008). A multi-component model for assessing learning objects: The learning object evaluation metric (LOEM). *Australasian Journal of Educational Technology*, *24*(5). https://doi.org/10.14742/ajet.1192

Kay, R., Lesage, A., & Tepylo, D. (2019, November). Evaluating the learning, design and engagement value of mobile applications: The mobile app evaluation scale. *Proceedings of the 12th International Conference of Education, Research and Innovation, Seville, Spain,* 1103–1107. https://doi.org/10.21125/iceri.2019.0336

Khan, A. I., Al-Khanjari, Z., & Sarrab, M. (2017, April). Crowd sourced evaluation process for mobile learning application quality. *Proceedings of the 2nd International Conference on Information Systems Engineering, Charleston, SC, USA.* https://doi.org/10.1109/ICISE.2017.17

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering: Version 2.3* (EBSE Technical Report EBSE-2007-01), University of Durham, Durham, UK.

Kolak, J., Norgate, S. H., Monaghan, P., & Taylor, G. (2021). Developing evaluation tools for assessing the educational potential of apps for preschool children in the UK. *Journal of Children and Media*, *15*(3), 410–430. https://doi.org/10.1080/17482798.2020.1844776

Kraus, S., Breier, M., & Dasí-Rodríguez, S. (2020). The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, *16*(3), 1023–1042. https://doi.org/10.1007/s11365-020-00635-4

Kucirkova, N. (2019). *Reading to your child? Digital books are as important as print books.* https://sciencenorway.no/books-children-opinion/reading-to-your-child-digital-books-are-as-important-as-print-books/1606950

Leacock, T. L., & Nesbit, J. C. (2007). A framework for evaluating the quality of multimedia learning resources. *Journal of Educational Technology & Society*, *10*(2), 44-59.

Lee, C. Y., & Cherner, T. S. (2015). A comprehensive evaluation rubric for assessing instructional apps. *Journal of Information Technology Education*, *14*(1), 21–53. https://doi.org/10.28945/2097

Lee, J. S., & Kim, S. W. (2015). Validation of a tool evaluating educational apps for smart education. *Journal of Educational Computing Research*, *52*(3), 435–450. https://doi.org/10.1177/0735633115571923

Lisenbee, P. S. (n.d.). Literacy app evaluation tool for teachers: Phonemic awareness and phonics apps rubric. https://conference.iste.org/uploads/ISTE2018/HANDOUTS/KEY_110809656/ISTEresearchdoc06022018_RP.pdf

Lubniewski, K. L., McArthur, C. L., & Harriott, W. A. (2017). Evaluating instructional apps using the app checklist for educators (ACE). *International Electronic Journal of Elementary Education*, *10*(3), 323–329. https://doi.org/10.26822/iejee.2018336190

Maalej, W., Kurtanović, Z., Nabil, H., & Stanik, C. (2016). On the automatic classification of app reviews. *Requirements Engineering*, *21*(3), 311–331. https://doi.org/10.1007/s00766-016-0251-9

Malone, M., & Peterson, M. (2013). Is there an app for that? Developing an evaluation rubric for apps for use with adults with special needs. *Journal of BSN Honors Research*, *6*(1), 39–56. http://hdl.handle.net/2271/1181

Martín-Monje, E., Arús, J., Rodríguez-Arancón, P., & Calle-Martínez, C. (2014). REALL: Rubric for the evaluation of apps in language learning. Proceedings of *Jornadas Internacionales Tecnología Móvil e Innovación en el Aula: Nuevos Retos y Realidades Educativas*. https://www.researchgate.net/publication/255702557_REALL_Rubric_for_the_evaluation_of_apps_in_language_learning

McGrath, J. (2011). *Mobile learning apps > statistics and trends*. Knowledge Direct Learning Management System. https://www.kdplatform.com/mobile-learning-apps-statistics-trends/

McManis, L. D., & Parks, J. (2011). *Evaluating technology for early learners*. Hatch, Inc. https://www.eschoolnews.com/files/2012/01/EvaluatingTechnology_ebook_toolkit.pdf

McQuiggan, S., McQuiggan, J., Sabourin, J., & Kosturko, L. (2015). The business of educational apps. In S. McQuiggan, J. McQuiggan, J. Sabourin, & L. Kosturko, *Mobile Learning: A Handbook for Developers, Educators, and Learners* (pp. 215-235). John Wiley & Sons. https://onlinelibrary.wiley.com/doi/10.1002/9781118938942.ch11

Mengist, W., Soromessa, T., & Legese, G. (2020). Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX*, *7*, 100777. https://doi.org/10.1016/j.mex.2019.100777

Meyer, M., Zosh, J. M., McLaren, C., Robb, M., McCaffery, H., Golinkoff, R. M., Hirsh-Pasek, K., & Radesky, J. (2021). How educational are "educational" apps for young children? App store content analysis using the Four Pillars of Learning framework. *Journal of Children and Media*, *15*(4), 526-548. https://doi.org/10.1080/17482798.2021.1882516

Mishra, P., & Koehler, M. J. (2008, March). *Introducing technological pedagogical content knowledge*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, USA.

Mneumann, M., Wang, Y., Qi, G. Y., & Neumann, D. L. (2019). An evaluation of Mandarin learning apps designed for English speaking preschoolers. *Journal of Interactive Learning Research*, *30*(2), 167–193.

Mobile Action. (n.d.). *Top education apps in India of google play store*. https://www.mobileaction.co/top-apps/education-4/android/in

Mobile App Daily. (2021, May). *Top 10+ best learning apps in 2021 to learn on the go*. https://www.mobileappdaily.com/best-apps-for-learning-anything

More, C. M., & Travers, J. C. (2013). What's app with that? Selecting educational apps for young children with disabilities. *Young Exceptional Children*, *16*(2), 15–32. https://doi.org/10.1177/1096250612464763

Mustaffa, F. Y., Salam, A. R., Muhammad, F., Bunari, G., & Asary, L. H. (2016). Literature review of educational app evaluation rubrics. *Intervention in School and Clinic*, *51*(4), 244–252.

Namukasa, I. K., Gadanidis, G., Sarina, V., Scucuglia, S., & Aryee, K. (2016). Selection of apps for teaching difficult mathematics topics: An instrument to evaluate touch-screen tablet and smartphone mathematics apps. In P. Moyer-Packenham (Ed.), *International perspectives on teaching and learning mathematics with virtual manipulatives* (pp. 275-300). Springer. https://doi.org/10.1007/978-3-319-32718-1_12

Neumann, M. M. (2018). Using tablets and apps to enhance emergent literacy skills in young children. *Early Childhood Research Quarterly*, *42*, 239-246. https://doi.org/10.1016/j.ecresq.2017.10.006

O'Brien, A. M., & McGuckin, C. (2016). *The systematic literature review method: Trials and tribulations of electronic database searching at doctoral level.* Sage. https://doi.org/10.4135/978144627305015595381

Ok, M. W., Kim, M. K., Kang, E. Y., & Bryant, B. R. (2016). How to find good apps: An evaluation rubric for instructional apps for teaching students with learning disabilities. *Intervention in School and Clinic*, *51*(4), 244–252. https://doi.org/10.1177/1053451215589179

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, *37*(1), 879–910. https://doi.org/10.17705/1cais.03743

Papadakis, S. (2021). Tools for evaluating educational apps for young children: A systematic review of the literature. *Interactive Technology and Smart Education*, *18*(1), 18-49. https://doi.org/10.1108/ITSE-08-2020-0127

Papadakis, S., & Kalogiannakis, M. (2017). Mobile educational applications for children. What educators and parents need to know. *International Journal of Mobile Learning and Organisation*, *11*(2), 1. https://doi.org/10.1504/ijmlo.2017.10003925

Papadakis, S., Kalogiannakis, M., & Zaranis, N. (2017). Designing and creating an educational app rubric for preschool teachers. *Education and Information Technologies*, *22*(6), 3147–3165. https://doi.org/10.1007/s10639-017-9579-0

Papadakis, S., Vaiopoulou, J., Kalogiannakis, M., & Stamovlasis, D. (2020). Developing and exploring an evaluation tool for educational apps (E.T.E.A.) targeting kindergarten children. *Sustainability*, *12*(10), 4201. https://doi.org/10.3390/su12104201

Perforce. (2021). *The importance of 5-star app ratings*. https://www.perfecto.io/blog/importance-5-star-app-ratings

Pinkston, G. (2021). *iPad apps for educators*. https://www.garypinkston.com/ipad-web-app.html

Reeves, T. C. (1994). Evaluating what really matters in computer-based education. In M. Wild, & D. Kirkpatrick (Eds.), *Computer education: New perspectives* (pp. 219–246). Mathematics, Science & Technology Education Centre, Edith Cowan University, Australia.

Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO Journal*, 34(2), 243–258. https://doi.org/10.1558/cj.27623

Sanromà-Giménez, M., Lázaro Cantabrana, J. L., Usart Rodríguez, M., & Gisbert-Cervera, M. (2021). Design and validation of an assessment tool for educational mobile applications used with autistic learners. *Journal of New Approaches in Educational Research*, *10*(1), 101-121. https://doi.org/10.7821/naer.2021.1.574

Sarrab, M., Hafedh, A. S., & Bader, A. M. (2015). System quality characteristics. *Turkish Online Journal of Distance Education*, *16*(4), 18–28.

Schrock, K. (2011). *Critical evaluation of a content-based iPad/iPod app*. https://www.kathyschrock.net/uploads/3/9/2/2/392267/evalipad_content.pdf

Sharma, N. (2021) How mobile education apps are improving education system in the world? eLearning Industry. https://elearningindustry.com/mobile-education-apps-improving-education-system-world

Sharma, S. (2016). The what and how of learning apps in pedagogies. *International Journal of Humanities, Arts, Medicine and Sciences*, *4*(2), 269–276.

Sheikh, A., Munro, M., & Budgen, D. (2019). Systematic literature review (SLR) of resource scheduling and security in cloud computing. *International Journal of Advanced Computer Science and Applications*, *10*(4). https://doi.org/10.14569/ijacsa.2019.0100404

Shing, S., & Yuan, B. (2016). Apps developed by academics. *Journal of Education and Practice*, *7*(33).

Shoukry, L., Sturm, C., & Galal-Edeen, G. H. (2015). Pre-MEGa: A proposed framework for the design and evaluation of preschoolers' mobile educational games. In T. Sobh, & K. Elleithy, K. (Eds.), *Innovations and*

Advances in Computing, Informatics, Systems Sciences, Networking and Engineering (pp. 385–390). Springer. https://doi.org/10.1007/978-3-319-06773-5_52

Shuler, C. (2009). *iLearn: A content analysis of the iTunes app store' s education section.* The Joan Ganz Cooney Center. http://www.joanganzcooneycenter.org/publication/ilearn-a-content-analysis-of-the-itunes-app-stores-education-section/

Shuler, C. (2012). *iLearn II: An analysis of the education category on Apple's app store.* Joan Ganz Cooney Center at Sesame Workshop. https://joanganzcooneycenter.org/wp-content/uploads/2012/01/ilearnii.pdf

SiteProNews. (2020, July 14). *Importance of mobile apps in education.* https://www.sitepronews.com/2020/07/14/importance-of-mobile-apps-in-education/

Situated learning. (n.d.). In *EduTech Wiki.* http://edutechwiki.unige.ch/en/Situated_learning

Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile app rating scale: A new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealt*h, *3*(1), e27. https://doi.org/10.2196/mhealth.3422

Swanson, G. (n.d.). *Apps in education.* https://sur.ly/o/appsineducation.blogspot.com/AA000014?pageviewId=desktop-302e30353539343330302031363732313930383534203538313732373038

Tahir, R., & Arif, F. (2014). Framework for evaluating the usability of mobile educational applications for children. *Proceedings of the Third International Conference on Informatics Engineering and Information Science, Lodz, Poland,* 156–170. https://www.researchgate.net/publication/265852684_Framework_for_Evaluating_the_Usability_of_Mobile_Educational_Applications_for_Children

Tahir, R., & Wang, A. I. (2017). State of the art in game based learning: Dimensions for evaluating educational games. In M. Pivec, & J. Gründler (Eds.), *Proceedings of the 11th European Conference on Games Based Learning* (pp. 641–650). Academic Conferences and Publishing International Limited.

Taylor, G., Kolak, J., Bent, E. M., & Monaghan, P. (2022). Selecting educational apps for preschool children: How useful are website app rating systems? *British Journal of Educational Technology*, *53*(5), 1262–1282 https://doi.org/10.1111/bjet.13199

Team Leverage Edu. (2021). *Top 25 free educational apps for students.* https://leverageedu.com/blog/free-educational-apps-for-students/

Tolisano, S. R. (2012). *Ipad app evaluation for the classroom.* Scribd. https://www.scribd.com/doc/94980508/iPad-App-Evaluation-for-the-Classroom

Tu, Y., Zou, D., & Zhang, R. (2020). A comprehensive framework for designing and evaluating vocabulary learning apps from multiple perspectives. *International Journal of Mobile Learning and Organization*, *14*(3), 370–397. https://doi.org/10.1504/IJMLO.2020.108199

Udayan, T. (2022). *10 best free educational apps for students & kids. M*indster. https://mindster.com/free-educational-apps-students/

Vaala, S., Ly, A., & Levine, M. H. (2015). *Getting a read on the app stores: A market scan and analysis of children's literacy apps*. The Joan Ganz Cooney Center at Sesame Workshop.

Vincent, T. (n.d.a). *Educational app evaluation checklist*. Squarespace. https://static.squarespace.com/static/50eca855e4b0939ae8bb12d9/50ecb58ee4b0b16f176a9e7d/50ecb593e4b0b16f176aa976/1330884481041/Vincent_App_Checklist.pdf

Vincent, T. (n.d.b). *Educational app evaluation rubric.* https://www.ocali.org/up_doc/ho_3_pp-rubric3.pdf?1557548866

Vincent, T. (2014, March 4). *Ways to evaluate educational apps*. Learning in Hand with Tony Vincent. https://learninginhand.com/blog/ways-to-evaluate-educational-apps.html

Walker, H. (2010). *Evaluation rubric for iPad apps.* http://learninginhand.squarespace.com/storage/blog/AppRubric.pdf

Walker, H. (2011). Evaluating the effectiveness of apps for mobile devices. *Journal of Special Education Technology, 26*(4), 59-63.

Wang, Y. Y., Wang, Y. S., Lin, H. H., & Tsai, T. H. (2019). Developing and validating a model for assessing paid mobile learning app success. *Interactive Learning Environments*, *27*(4), 458–477. https://doi.org/10.1080/10494820.2018.1484773

Weng, P. L. (2015). Developing an app evaluation rubric for practitioners in special education. *Journal of Special Education Technology*, *30*(1), 43–58. https://doi.org/10.1177/016264341503000104

Zeng, X., Peng, X., & Lu, C. (2017, June). Survey on the quality assessment factors of educational APP. *Proceedings of the International Symposium on Educational Technology, Hong Kong,* 196–200. https://doi.org/10.1109/ISET.2017.52

Appendix. Search Strings

Database	Types of Searches	Search Strings
Google Scholar	**Primary search strings**	**PS1**: (learning apps AND evaluation) **PS2**: (mobile learning apps AND evaluation) **PS3**: (touch screen instructional apps AND evaluation) **PS4**: (tablet learning apps AND evaluation)
	Secondary search strings	**SS1**: (educational software for learning purpose) **SS2**: (rubrics for educational apps) **SS3**: (learning OR literacy apps) **SS4**: (toddlers OR young children OR preschool apps) **SS5**: (iPads OR touch screen OR interactive learning apps)
Academia	**Primary search strings**	**PS1**: (intelligent apps AND their assessment criteria) **PS2**: (preschoolers educational apps AND appraising criteria) **PS3**: (mobile learning apps AND their appraising guidelines)
	Secondary search strings	**SS1**: (e-Learning app) **SS2**: (guidelines for evaluating the educational apps)
j-Gate	**Primary search strings**	**PS1**: (learning apps evaluation frameworks) **PS2**: (rubrics for instructional apps appraising) **PS3**: (checklists for learning apps evaluation)
	Secondary search string	**SS1**: (emerging evaluation OR appraising tools for LAs)
iJET	**Primary search strings**	**PS1**: (educational apps OR instructional apps AND assessment tools) **PS2**: (eLearning apps AND evaluation protocols)
	Secondary search strings	**SS1**: (eLearning apps OR mApps for learning purpose) **SS2**: (online learning through user-friendly apps)
IEEE	**Primary search strings**	**PS1**: (teachers' suggestions mobile apps AND their assessment protocol) **PS2**: (eLearning apps evaluation tools and techniques) **PS3**: (framework OR rubrics on mobile apps assessment)
	Secondary search string	**SS1**: (emerging evaluation tools OR techniques for educational apps)
ERIC	**Primary search strings**	**PS1**: (emerging ratings OR review tools for instructional apps) **PS2**: (appraising models for preschool education) **PS3**: (assessments models for educational apps)
	Secondary search strings	**SS1**: (literacy applications for online learners) **SS2**: (online learning through apps)
Lib-Tech	**Primary search strings**	**PS1**: (iPads apps AND their appraising methods) **PS2**: (interactive mobile apps OR applications AND their evaluation tools) **PS3**: (tablet mobile learning applications AND their appraising models)

Database	Types of Searches	Search Strings
	Secondary search strings	**SS1**: ways to evaluate the mobile learning applications) **SS2**: (educational apps AND their significances) **SS3**: (small apps AND their uses in today's classroom)
Wiley Inter-Science	**Primary search strings**	**PS1**: (instructional apps AND their assessment models) **PS2**: (interactive learning apps AND their evaluation methods) **PS3**: (apps for online learning AND their appraising protocols) **PS4**: (apps for eLearning AND their quality assessment criteria) **PS5**: (handheld software for educational purpose AND their existing appraisal approaches)
	Secondary search string	**SS1**: (reviews methods for learning apps assessment) **SS2**: (teachers' suggestions for selection of desire educational apps) **SS3**: (checklists for appraising existing mobile learning apps)
DOAJ	**Primary search strings**	**PS1**: (the selection guidelines for light-weight educational software) **PS2**: (software applications for learning purpose AND their selection methods) **PS3**: (free OR paid educational apps AND their appraising approaches)
	Secondary search strings	**SS1**: (assessment tools for literacy mobile applications) **SS2**: (scientific evaluation methods for learning apps) **SS3**: (simple reviews OR observations for instructional apps evaluation)

AUTHORS

Shahjad is the corresponding author of this research article. He is a senior research fellow in the Department of Computer Science in Jamia Millia Islamia (a central university), New Delhi, India. He completed his undergraduate and postgraduate studies in computer science. He is currently pursuing a Ph.D. course in the same field. His areas of expertise include eLearning and Learning Apps.

Dr Khurram Mustafa is an IIT Delhi alumnus, who is currently the seniormost professor in the Department of Computer Science at Jamia Millia Islamia (a central university) in New Delhi, India. Despite having completed his PhD on a topic related to eLearning, he continues to supervise students and write/speak on information security, e-learning, and research methods. During his five-year hiatus, he worked as a professor/associate professor at universities in Saudi Arabia, Yemen, and Jordan. In addition to authoring Scientific Research Primer (Ane Books, 2021) and coauthoring two other books, *Software Quality: Concepts and Practices* and *Software Testing: Concepts and Practices* (both published by Narosa, India, and Alpha Science, UK), he has mentored over a dozen PhD candidates. The latter's Chinese edition has also been released. Aside from these, he has coauthored more than a dozen book chapters and over 100 research papers published in international journals/proceedings. He was also the principal investigator for a three-year government-funded information security project and delivered more than 60 invited talks, including several keynote addresses. He is also a member of several professional scientific societies, including ISTE, ICST, CSI, EAI, ACM-CSTA, eLearning Guild, and InfoPier, as well as several academic committees and editorial review boards.