Advances in Educational Technology and Psychology (2021)
Clausius Scientific Press, Canada

DOI: 10.23977/aetp.2021.59011
ISSN 2371-9400 Vol. 5 Num. 9

A Systematic Literature Review on How to Evaluate Educational Games in K-12

Xinran Wang[a], Jinbo Tan[b]

Faculty of Education, Shandong Normal University, Jinan, China

[a]wxr980630@126.com, [b]yttjb@163.com

Keywords: Educational game, k-12, evaluate

Abstract: Educational games as an emerging strategy for evaluation and learning, have received widespread attention. How to evaluate educational games has become a hot topic in the field of educational technology in recent years. Therefore, the purpose of this article is to introduce the evaluation of the development of educational games in primary and secondary schools in the past five years. We selected three commonly used English databases, searched the papers for the past five years according to keywords, and identified 14 papers from them, describing the different methods (including emerging technology and traditional evaluation) used in the evaluation of educational games. And introduced the dimensions of different evaluation methods. We also describe the tools and methods used in collecting and analyzing data. Most of the papers will use traditional evaluation methods, even if they use emerging technology evaluation methods, they will still use traditional evaluation methods as supplements. Nevertheless, most of the evaluation methods used in the paper are highly targeted, and there are various choices according to different evaluation dimensions. The author finally suggests that the evaluation dimension of educational games can be classified by Bloom's teaching objective classification, and can refer to the educational game evaluation methods proposed in other higher education.

1. Introduction

In recent years，due to the rapid development of technology，our learning method should not be limited to textbooks, and the learning content should not be limited to textbook knowledge. The advent of educational game has broken this deadlock. Educational games can not only be used as a way to learn textbook knowledge but also can learn or improve a certain ability; therefore, educational games have attracted much attention in the educational technology field. How to use games for learning, that is, educational games, has caused great concern in the field of education.

More and more scholars use games as a tool to evaluate student learning, apply educational games to different disciplines, design different educational games according to different research purposes, and take different measures to collect and analyze data to evaluate games what impact has it had on students.

Regarding what is an educational game, Dempsey puts forward his own opinions on topics, concepts, learning skills and attitudes[1]. Connolly et al. believe that games are a instructional strategy to achieve high-level learning[2]. The author believes that there are many definitions of educational games, but its main purpose is to have a positive impact on education.

The quality of educational game is directly related to the effectiveness of students' learning and students' interest. Therefore，how to evaluate educational games with high quality has become a huge challenge now.

Wang et al. regard digital educational game evaluation as a process, and believe that the main body of digital educational game evaluation uses specific evaluation standards, methods, techniques, tools and processes to clarify the advantages or value of digital educational games from a theoretical perspective[3].Regarding the traditional evaluation of educational games, the author believes that the main form is to conduct certain tests according to the evaluation dimension before and after the game to check the effects of educational games.

Thus, in order to elicit the state of the art of how to systematically evaluate educational games, we conducted a systematic literature review. The main contribution of this paper is to explore the main methods used in the evaluation of educational games (the methods of emerging technologies and traditional evaluation methods), as well as the data collection and evaluation methods in the evaluation of educational games.

There are a lot of reviews about the evaluation of educational game. Petri et al. described the method of systematically evaluating educational games and gave a more detailed description[4]. They believe that most of the methods of evaluating educational games are temporarily developed and believe that it is necessary to further study the definition and operability of educational game evaluation. Tahir et al. studied the dimension of educational game evaluation[5]. Petri et al. analyzed the definition of the methods they used, and the way of evaluation and analysis[3]. Wang et al. believe that the current understanding of digital educational game evaluation in foreign digital educational game evaluation is mainly based on objectivist epistemology, without paying enough attention to educational evaluation theory, and the evaluation tools proposed also need to be strengthen the validity test[2].

In previous studies, the researchers did not clearly distinguish between different school segments, and did not explain how the different evaluation methods are connected and what problems exist. So, in this study, we summarized the methods used to evaluate educational games used in k-12 education, without considering higher education even the educational games for the elderly and social people. And we have further explained the traditional evaluation methods and the evaluation methods of emerging technologies.

In order to systematically examine the methods used in evaluating educational games, the review is guided by the following questions:

Q1: Which technology (approaches) exist to systematically evaluate educational games? (We think the technology includes models, scales, or frameworks etc.)

Q2: What methods are used to collect and analyze data? (We think that data includes data generated in the education games as well as experimental data.)

2. Method

In order to select useful papers and ensure the maximum relevance of the papers,we used keyword search as search approaches .We have selected the three most commonly used English database by the author: Springer, Web of Science, and EBSCO.

Table 1: Search statement

Database	Boolean expression
Springer	'(assessment OR evaluation) AND ("educational game " OR "learning game " OR "gaming for education" OR "integrating the game") AND (education OR " game based learning")'
Web of Science	TS= ((assessment* OR evaluation*) AND (“educational game*” OR “learning game*” OR “gaming for education*” OR “integrating the game*”) AND (education* OR “game based learning*”))
EBSCO	(assessment OR evaluation) AND ("educational game " OR "learning game " OR "gaming for education" OR "integrating the game") AND (education OR " game based learning") published between 2015-2020

The following inclusion criteria were set to screen the papers：

Inclusion Criteria 1: Papers must include educational games and evaluation.

Inclusion Criteria 2: Papers belonging to empirical study.

Inclusion Criteria 3: Papers only for primary and secondary school students.

Inclusion Criteria 4: Papers don’t include a review of evaluation of educational games.

Keyword search is a commonly used method to search the target literature. After logging into the appropriate database, enter the correct Boolean expression.

After using keyword search，there are 43 papers in EBSCO, there are 113 papers in Web of Science, there are 545 papers in Springer, in order to ensure maximum relevance we have selected only the first 100 articles. And after screening according to the above criteria, no selected literature was included in EBSCO, 9 papers in the Web of science meet the criteria, 6 papers in Springer meet the criteria. There is an identical paper in Web of Science Therefore，a total of 14 papers were qualified.

3. Results

3.1 What Methods Can Evaluate Educational Games?

In this section, we analyzed the 14 papers from the perspective of the way of educational games evaluation.

Among the final selected papers，11（78.6%）papers used traditional evaluation methods, and 8 of these 11 papers used emerging technology evaluations. The remaining 6 (42.9%) papers only use traditional evaluation methods, as shown in Figure 1. Of the 8 papers, 3 (37.5%) used the model, 1 (12.5%) used the framework, and the remaining 4 (50.0%) used other methods, namely game log files and development systems or software and engine.

We divide the evaluation of emerging technologies into three categories: models, frameworks, and other categories.

Yeni and Cagiltay tries to demonstrate the importance of the academic content-fantasy integration. Therefore, they analyzed the academic and fantasy aspects of the game through the RETAIN model, and analyzed the gaming experience and entertainment aspects through the GameFlow model standard[6]. Both models can provide guidance for game development and teachers who use the game in teaching. Su and Hsaio proposed a novel multiple-criteria decision making model based on Flow theory，and developed an evaluation mobile learning game system (EMGS) based on the proposed model[7]. The results of the study indicate that using the developed system is faster than traditional evaluation methods.Two papers used Technical Acceptance Model in the evaluation process，it mainly used to measure the player’s acceptance of information

technology. Researchers usually use this model as a basis to design questionnaires.Research has proved that TAM is a useful theoretical model that can understand and explain user behavior in information systems[8].With the advent of the era of education informatization, our country also has requirements for the programming ability of primary and secondary school students. And the results of preliminary students indicate that educational games support students to understand basic programming concepts. Duh et al. scholars used the educational game Azbuka and studied the game's acceptance and its impact on children's learning motivation. For the study of acceptance, the three scholars chose the TAM model[9]. When studying the intrinsic motivation of learning, they chose the Intrinsic Motivation Inventory model as the basis for evaluation, and designed four subscales for interest/enjoyment, effort/importance, pressure/tension, value/usefulness. And they found that children have a high internal motivation for learning new technologies, and they will soon be able to accept and adopt new technologies.

One of these 8 papers used framework: TPACK framework. It analyzed effects on practice.Evans, Nino, Deater-Deckard and Chang used the TPACK framework to analyze the impact of the use of mathematics learning games in practice in the middle school setting, and analyzes the three main research areas of TPACK in detail: Technology, pedagogy and content knowledge[10]. The implementation of new learning techniques and teaching methods has brought about significant changes to teachers' teaching, so it is beneficial to use technology integration theoretical frameworks such as TPACK as analytical lenses.

In recent years, the game log files has attracted people's attention as a new way of evaluating educational games. Ke and Hulse, T., et al used the game log files flexibly in their research[11][12]. The log files in the game provides us with a unique innovative evaluation function, which can measure the entire learning process of solving problems in real time. Using the game log files to capture the various behaviors of the player is a more accurate method.

Some scholars use Bayesian networks when evaluating educational games. Bayesian network is probabilistic graphical model that can be used for statistical modeling, inferred statistics, and have good interpretability. Hooshyar, et al. advanced Online Game-based Bayesian Intelligent Tutoring System (OGITS) to enhance programming acquisition and online information searching skills, thus improving students' ability in web-based problem solving through board games[13].

Jakos et al. used Sloodle software, it was integrated into Learning Management System Moodle supports assessment of pupil's progress, assessing student's learning progress[14].

Chaudy et al. presents an assessment engine designed to separate a game and its assessment[15].

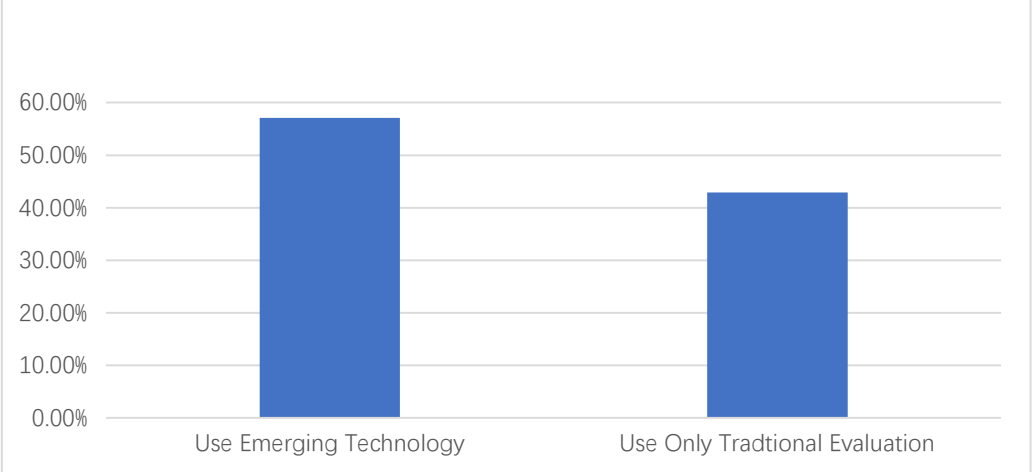

Fig.1: The methods of evaluating educational games

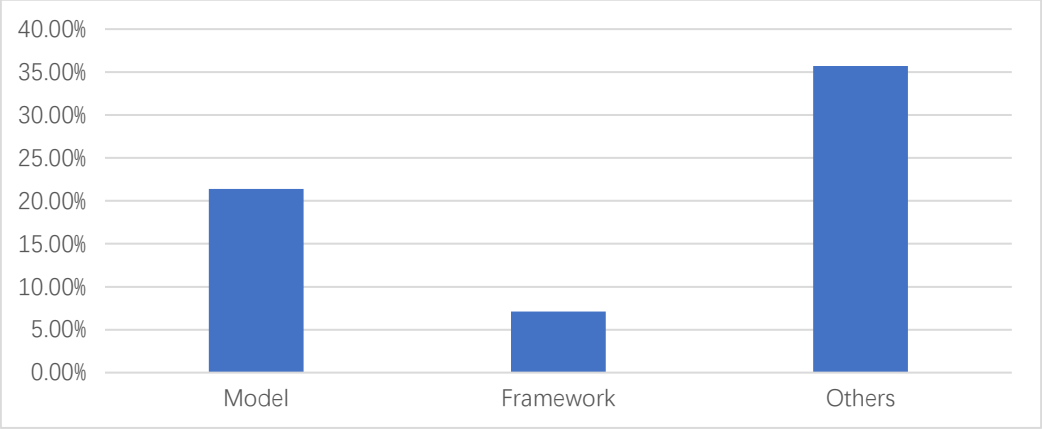

Fig.2: Types of technical evaluation methods

It can be seen from the above that there are few papers that use the framework as the basis for evaluation in the past five years; the choice of models is diverse, and the evaluation dimensions are also different. Most papers have adopted a more novel approach.

Almost all papers use traditional evaluation methods. Even if some papers mainly user emerging technologies to evaluate educational games, they will use traditional evaluation methods as supplementary or auxiliary evaluation methods.

From this we can fine that the methods used for evaluation are selected according to the dimensions of the literature to be evaluated. Even if it is the same evaluation dimension, the evaluation methods that can be used are various.

3.2 What Methods are Used to Collect and Analyze Data?

Most papers use questionnaires based on the model proposed in 3.1, for example: Duh et al. used a questionnaire with adopted statements according to intrinsic motivation model（IMM）method, and another questionnaire with adopted according to TAM methods[9]. Mavridis et al. used used the Attitudes Toward Mathematics Inventory (ATMI) questionnaire that was in the form of a five-point Likert scale[16].

There are also some questionnaires designed by researchers that are not model-based. Kiili and Ketamo used the flow experience and the test anxiety questionnaire, after completing game-based testing, the players were instructed to fill the playability questionnaire[17]. Cheng, Lin, She and Kuo used Game Immersive Questionnaire that was developed on the basis of the immersion theory proposed by Brown and Cairns[18]. Hooshyar, D., et al. used a Likert scale consisting of 15 items was used. It was developed based on scale literature used to evaluate adaptive and intelligent online education environments[13].

There are some more novel methods. Giannakoulas and Xinogalos used a game-activities worksheet[19]. Hulse, T., et al. used the pre- and post- study worksheets to collect student scores and student data log files that interact with the game[12]. Chaudy, Y. and T. Connolly used a Learning Analysis dashboard in EngAGe to collect student's data[15].

Some researchers use multiple data collection methods in their research. Ke used used math knowledge and mental rotation tests, video-and screen-capture of game play behaviors, infield observation, as well as the game log file[11]. Yeni and Cagiltay used three instruments to collect data: RETAIN rubric, GameFlow criteria and an interview schedule[6].

Of the 14 selected papers, 8(71.4%) papers used questionnaires and the questionnaire was in the form of Likert scale, 2(14.3%) papers used tests, 5(35.7%) papers used interviews, 2(14.3%) papers used game log files and 1(7.1%) used observation methods, as shown in Figure 3.

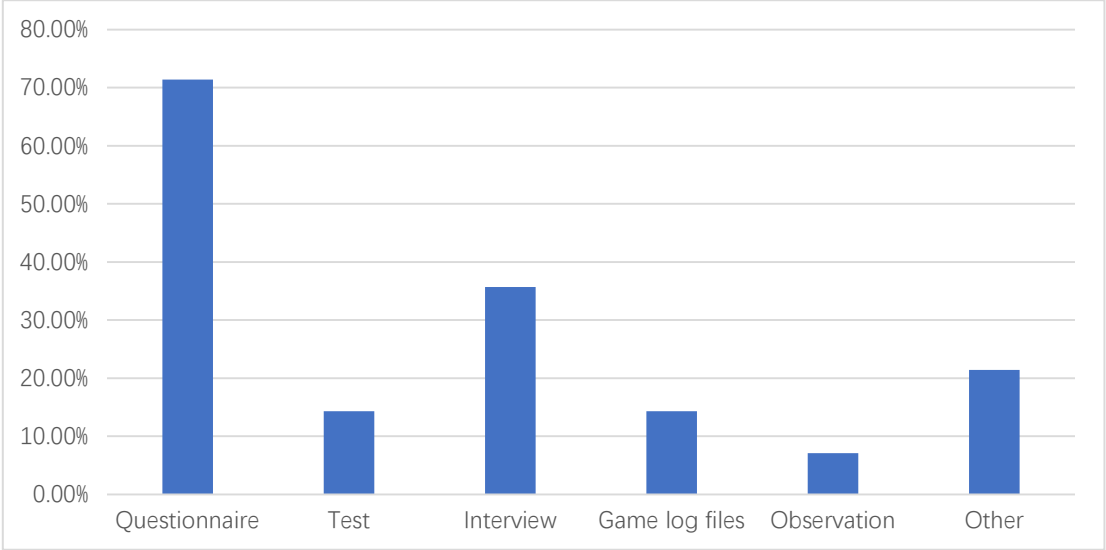

Fig.3: The methods of collect data

It can be seen that in the past five years, the questionnaire is still the most used data collection method. The usage rate of interviews as qualitative data is also very high. With the development of technology, some new methods of data collection have emerged.

When collecting data, most papers uses quantitative and qualitative methods.

In the 14 papers, a total of 12 papers used quantitative methods, 5 papers used qualitative methods, and 5 papers used both quantitative and qualitative methods. There are three papers using the scale as a data collection tool: Seven-point Likert scale, six-point Likert scale and temporal attentive observation scale. And others use questionnaires as a way of quantitative evaluation.

It is worth mentioning that, as mentioned in 3.1, the educational game log files can be used as a new way of evaluation, so the data collection about the educational game log files should also be taken seriously.

Regarding the method of data analysis, 6(42.9%) out of papers used descriptive statistical analysis, 3(21.4%) papers used independent sample t-test, 1(7.1%) paper used exploratory factor analysis, 1(7.1%) paper used regression analysis , 3(21.4%) papers used covariance analysis, 1 (7.1%) paper used multivariate covariance analysis, 2(14.3%) papers used variance analysis, 2(14.3%) papers used correlation analysis, 1(7.1%) paper used comparative analysis, and one paper used cluster analysis, as shown in Figure 4.

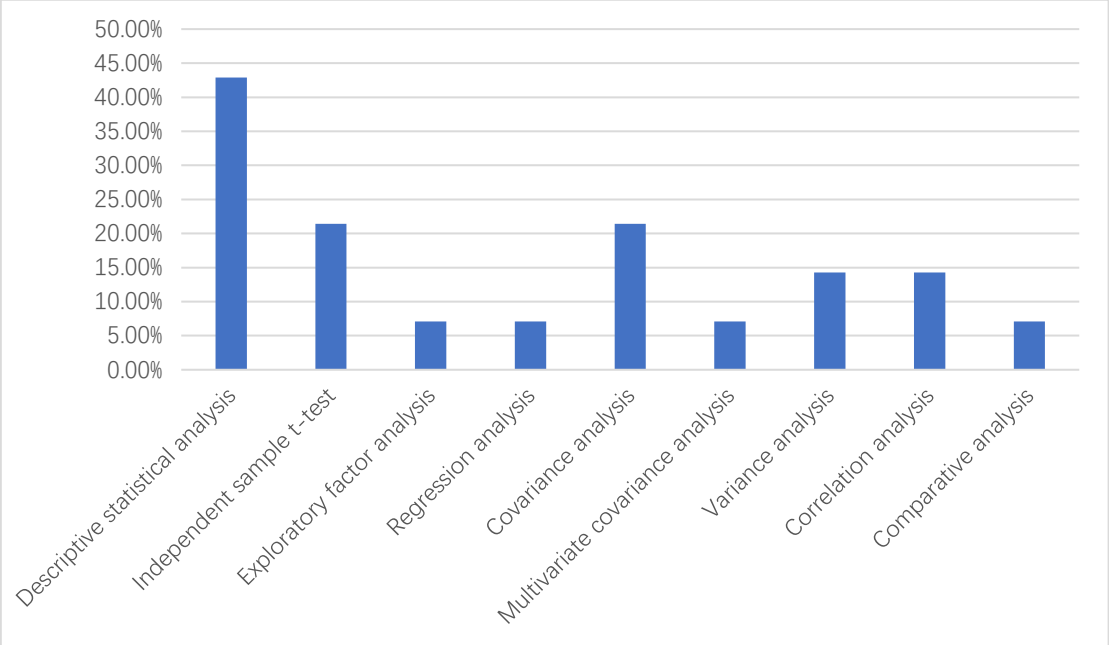

Fig.4: The methods of analyzing data

Regarding the tools used for data analysis, three papers used spss. One paper used Microsoft Excel (2016) and the "R" software environment. One paper used R packages Ez and corrplot. Other papers do not specify what tools are used to analyze the data. Regarding the methods of data analysis, all papers use the method of parameter statistics without exception.

4. Discussion

4.1 Discussion on Emerging Technologies and Traditional Evaluation Methods

We think the ultimate goal of the evaluation of educational games should still be how to improve students' academic performance and their personalized development better. It will be a great progress to make better use of technology to evaluate educational games.

In the selected 14 papers, a total of 4 papers use four models: TAM model, RETAIN model, GameFlow model, MCDM model[6][19]; one of paper uses the framework: TPACK framework; two papers use the game log files, one paper uses Bayesian Intelligent Tutoring System, one uses an evaluation engine: EngAGe[15]. And we found that papers that evaluate emotions, attitudes, and skills use more emerging technologies, but if it is relatively simple to want to explore the impact on performance, there are very few technical requirements.

Regarding the method of game evaluation, we divide it into two categories in this review, namely traditional evaluation methods and emerging technology evaluation methods.

We can see that although the technology is so advanced in today's era, the technology we can use in educational games is still limited. Even if there are many excellent technologies, such as Bayesian network, but most primary and secondary schools still use the traditional evaluation method in application, because the traditional evaluation method has been used for many years and has been developed for many years, and it is more mature. For new technologies, stealth evaluation has also received much attention in recent years. Opening the black box of the game is undoubtedly a major breakthrough in game evaluation in recent years. The evidence-based evaluation design

principles have also attracted more and more scholars' attention. However, some emerging technologies are still in the experimental stage, and because of the higher technical requirements, fewer teachers can achieve this. From this point of view, the traditional evaluation method has limited depth of evaluation, and some emerging technologies can supplement it. Therefore, we can see that the evaluation methods of emerging technologies are not completely separated from the traditional evaluation methods. The papers that use emerging technologies as evaluation methods also use the traditional evaluation methods.

The traditional evaluation method is mainly based on quantitative, qualitative supplemented. Mainly using questionnaire or semi-structured questionnaires, but there is a huge risk that in the process of filling out the questionnaire, how do we ensure the accuracy and objectivity of the students' answer?Attfield believes that when the questionnaire is inaccurate, it is shallow and prone to a halo effect. But the corresponding technology can ensure the accuracy and objectivity of the data we collect to a certain extent. For example, collecting data from student log files through the data set recording system is relatively objective and accurate. Because the traditional evaluation method only collects the data that before and after the game, and using the student log files can record the student's operation behavior during the game very accurately, and can lay a solid foundation for the subsequent data analysis.

4.2 Discussion on Data Collection Methods and Data Analysis

It is crucial that collecting and analyzing in the process of evaluating educational games, In the 14 selected papers, most of the data were collected using quantitative and qualitative methods. Qualitative data is mainly to supplement quantitative data. Quantitative methods focus on questionnaires, while qualitative methods use interviews. The questionnaires are divided into structured questionnaires, semi-structured questionnaires and open questionnaires. The questionnaires are almost all in the form of Likert scale, which is convenient for statistics and data calculation. Although the forms of these questionnaires are similar, due to the different models used in the design of the questionnaires, these questionnaires are highly targeted and difficult to generalize to other papers. Questionnaires are well accepted for measuring diverse factors, such as motivation or user experience. However, Ross believes that although their use may affect the measurement of learning effectiveness, sometimes the questionnaire is sometimes biased and untrustworthy.The design of a non-intrusive evaluation method using self-assessment may lead to results with low validity, if data is collected via ad-hoc questionnaires or interviews. A compromise may be the development of standardized questionnaires increasing the validity and reliability of the data being collected.

Testing is also one of the methods for collecting quantitative data, which is usually evaluated based on the student's pre- and post-test results(scores). However, there are some studies that do not have strict pre- and post-test, but only issue questionnaires to investigate after the game.

In addition, some papers have made greater innovations in collecting data. We mentioned in Section 4.1 that there are some papers that game log files can provide innovative ways for game evaluation. Therefore, in terms of data collection, the method adopted by some papers is mainly based on game log files.

The methods of data analysis used in the selected papers are very extensive. The methods of data analysis also need to be determined according to research, but according to the results in 3.2, we can find that the most used is descriptive statistics. This means that descriptive statistics is a basic data analysis method. Most of the papers mainly calculates the means and standard deviation of the

collected data. According to Lord paradox, the results of statistical analysis may depend on the specific methods used, so the results of data analysis may not be unique.

Regarding the tools used for data analysis, only three papers emphasize the use of spss as a tool for data analysis. Some papers do not describe which data analysis tool is used, but only give specific methods for data analysis, such as independent sample t test, analysis of variance. In addition, as mentioned in 4.1, in addition to the traditional questionnaire survey and structured interviews, the collected data can also be collected for the data in the game, that is, the game log files, which is used as the collection of aluminum data in the game. The ordinary questionnaire is just data collection before and after the game.

5. Conclusion

In this article, we briefly describe how to systematically evaluate the state of educational games over the past five years based on three English databases. The purpose of this study:(1) explore the evaluation methods used in the selected papers, (2) the methods of data collection and data analysis, (3) summarize and analyze the selected paper, and explore the prospect of educational game evaluation. We identified 14 papers describing different methods of evaluating educational games. Most of them are models rather than comprehensive evaluation methods, which indicates a lack of support in how to conduct such evaluations. The methods encountered are also very different in terms of the dimensions of assessment. In addition to evaluating learning effects, they also consider emotions, motivations, etc., which indicates that there is no pattern of factor to be evaluated. In addition, in the preliminary work of this paper review, we found that there are very few game evaluation related methodologies for guidance, most of which are directly developed or designed for application. Since the evaluation methods is determined according to different evaluation dimensions, and the purpose of the evaluation is mainly to test whether students achieve the learning goal or master a certain skill, we can discuss the different evaluation methods of games used for different purpose based on the classification of Bloom's learning goal.

Actually, using games to educate is not a new concept. Learning through play got completely new dimension. There are many researchers who believe that when use intentionally and appropriately, new technologies can provide an extensive scaffolding of learning.

In the process of summarizing and analyzing the selected papers, we found that most of the papers developed for different evaluation factors and the measurement tools are more targeted, so they have weak scalability. In addition, there are many papers that also provide some other technical evaluation methods, such as: An EEG-based methodology for assessing user's confusion in an educational game[20]. Some use scales developed, but because the papers requirements of this review are for K-12 students, some papers are not taken into account, but can we use them for reference? This is also a problem of future research [21,22].

References

[1]Dempsey, J. V., and Johnson, R. B. (1998).The development of an ARCS gaming scale. Journal of

Instructional Psychology, 25,215–221.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., and Boyle, J. M. (2012).A systematic literature review of empirical evidence on computer games and serious games. Computers & Education,59(2),661–686.

[3]WANG Wenjing, ZHAO Xiaochen, XIE Huixin and XIE Qiukui.(2019). A Review on the

Evaluation of Digital Educational Games Abroad. International and Comparative Education,41(03),101-108.

[4] Petri, G., von Wangenheim, C. G. (2017). How games for computing education are evaluated? A systematic literature review. Computers & Education, 107, 68-90.

[5] Tahir, R., Wang, A. I. (2017). State of the art in game based learning: Dimensions for evaluating educational games. Paper presented at the European Conference on Games Based Learning.

[6] Yeni, S., Cagiltay, K. (2017). A heuristic evaluation to support the instructional and enjoyment aspects of a math game. Program-Electronic Library and Information Systems, 51(4), 406-423.

[7] Su, C., Hsaio, K. (2015). Developing and Evaluating Gamifying Learning System by Using Flow-Based Model. Eurasia Journal of Mathematics Science and Technology Education, 11(6), 1283-1306.

[8] Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. Information Management, 40, 191–204.

[9] Duh, E. S., Koceska, N., Koceski, S. (2017). Game-based learning: educational game Azbuka to help young children learn writing Cyrillic letters. Multimedia Tools and Applications, 76(12), 14091-14105.

[10] Evans, M. A., Nino, M., Deater-Deckard, K., Chang, M. (2015). School-Wide Adoption of a Mathematics Learning Game in a Middle School Setting: Using the TPACK Framework to Analyze Effects on Practice. The Asia-Pacific Education Researcher, 24(3), 495-504.

[11] Ke, F. (2019). Mathematical problem solving and learning in an architecture-themed epistemic game. Educational Technology Research and Development, 67(5), 1085-1104.

[12] Hulse, T., Daigle, M., Manzo, D., Braith, L., Harrison, A., Ottmar, E. (2019). From here to there! Elementary: a game-based approach to developing number sense and early algebraic understanding. Educational Technology Research and Development, 67(2), 423-441.

[13] Hooshyar, D., Ahmad, R. B., Wang, M., Yousefi, M., Fathi, M., Lim, H. (2018). Development and Evaluation of a Game-Based Bayesian Intelligent Tutoring System for Teaching Programming. Journal of Educational Computer Research, 56(6), 775-801.

[14] Jakos, F., Verber, D. (2017). Learning Basic Programing Skills With Educational Games: A Case of Primary Schools in Slovenia. Journal of Educational Computer Research, 55(5), 673-698.

[15] Chaudy, Y., Connolly, T. (2019). Specification and evaluation of an assessment engine for educational games: Integrating learning analytics and providing an assessment authoring tool. ENTERTAINMENT COMPUTING, 30(UNSP 100294).

[16] Mavridis, A., Katmada, A., Tsiatsos, T. (2017). Impact of online flexible games on students' attitude towards mathematics. Educational Technology Research and Development, 65(6), 1451-1470.

[17] Kiili, K., Ketamo, H. (2018). Evaluating Cognitive and Affective Outcomes of a Digital Game-Based Math Test. IEEE Transactions on Learning Technologies, 11(2), 255-263.

[18] Cheng, M.-T., Lin, Y.-W., She, H.-C., & Kuo, P.-C. (2016). Is immersion of any value? Whether, and to what extent, game immersion experience during serious gaming affects science learning. British Journal of Educational Technology, 48(2), 246–263.

[19] Giannakoulas, A., Xinogalos, S. (2018). A pilot study on the effectiveness and acceptance of an educational game for teaching programming concepts to primary school students. Education and Information Technologies, 23(5), 2029-2052.

[20] Zhou, Y., Xu, T., Li, S., Shi, R. (2019). Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game. Universal Access In The Information

Society, 18(3SI), 551-563.
[21] Petri, G., von Wangenheim, C. G. (2016). How to Evaluate Educational Games: a Systematic Literature Review. Journal of Universal Computer Science, 22(7), 992-1021.
[22] Ninaus, M., Kiili, K., McMullen, J., Moeller, K. (2017). Assessing fraction knowledge by a digital game. Computer In Human Behavior, 70, 197-206.