Sánchez-Prieto, J. C., Gamazo, A., Cruz-Benito, J., Therón, R., & García-Peñalvo, F. J. (2020). AI-Driven Assessment of Students: Current Uses and Research Trends. In P. Zaphiris & A. Ioannou (Eds.), Learning and Collaboration Technologies. Design, Experiences. 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I (pp. 292-302). Springer Nature. https://doi.org/10.1007/978-3-030-50513-4

AI-Driven Assessment of Students: Current Uses and Research Trends

José Carlos Sánchez-Prieto^{1,4(⋈)}, Adriana Gamazo^{1,4}, Juan Cruz-Benito², Roberto Therón^{1,3,5}, and Francisco J. García-Peñalvo^{1,3,4}

¹ GRIAL Research Group, University of Salamanca, Salamanca, Spain {josecarlos.sp,adrianagamazo,theron,fgarcia}@usal.es
² IBM Quantum, IBM Research, IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598, USA

Juan.cruz@ibm.com

³ Computer Science Department, University of Salamanca, Salamanca, Spain
⁴ Research Institute for Educational Sciences (IUCE),

University of Salamanca, Salamanca, Spain

Abstract. During the last decade, the use of AIs is being incorporated into the educational field whether to support the analysis of human behavior in teaching-learning contexts, as didactic resource combined with other technologies or as a tool for the assessment of the students.

This proposal presents a Systematic Literature Review and mapping study on the use of AIs for the assessment of students that aims to provide a general overview of the state of the art and identify the current areas of research by answering 6 research questions related with the evolution of the field, and the geographic and thematic distribution of the studies.

As a result of the selection process this study identified 20 papers focused on the research topic in the repositories SCOPUS and Web of Science from an initial amount of 129.

The analysis of the papers allowed the identification of three main thematic categories: assessment of student behaviors, assessment of student sentiments and assessment of student achievement as well as several gaps in the literature and future research lines addressed in the discussion.

Keywords: Artificial intelligence · Education · Assessment · Algorithmic evaluation

1 Introduction and Background

The word artificial intelligence is every day in everyone's mouth. It is one of the most over-hyped research areas in media, and society is permeable to such publicity. It is true that artificial intelligence and its advancement are revolutionizing many aspects of our daily lives [1–4] and can improve the efficiency of different processes, but are those advancements adequately communicated?

⁵ VisUSAL Research Group, University of Salamanca, Salamanca, Spain

For years, artificial intelligence has been interacting with users in digital environments in more or less obvious ways [5]. In many cases, users are not aware that their activity is continuously evaluated or intervened by software agents with a minimum of intelligence. However, this non-obvious interaction affects these users and conditions the resources, actions, or states that users reach in the digital environments they use [5, 6]. Today, one of the natural trends in different areas is to introduce intelligent software agents to improve specific areas of computer systems. This has introduced an intense discussion [7–9] about the effects of these agents, giving rise to reflections about the privacy of the users, the ethics in such systems, or the need to create common working frameworks that respect the users and their human characteristics. Too often, disrespectful behavior towards users has been observed, sometimes creating a negative image of the algorithms and how they are used in conjunction with people [7, 10].

These issues are relevant in general, but even more so when they affect such fundamental social and human issues like behavior, freedom, education, or social good. In the case of this article, we will focus on how artificial intelligence is involved in a specific case related to education and the social sciences, such as educational evaluation.

For years, progress has been made in applying computer solutions to the education process, with more or less success [11-14]. These computational approaches are increasingly common in STEM (Science, Technology, Engineering, and Mathematics) areas, while in other areas, they present some difficulties that limit the emergence of such solutions. In the STEM areas, it is possible to assess the results of many operations or learning outcomes using calculation or mathematical processes. Therein lies the growth of computational approaches related to education in such areas and the appearance of algorithmic evaluations or algorithm-based educational approaches [15]. Something as simple for a computer as checking the result of a set of operations, calculations, or processes regulated by mathematics can facilitate the work of teachers and educators when correcting exercises, exams, or assignments within a regulated context. This benefit is even more evident in contexts with many students or people involved in similar processes (like in MOOCs). As a related context, it is possible to cite several works in the field of research in Learning Analytics or Educational Data Mining[16–18] that deal with this type of questions or the use of tools such as Jupyter notebooks [19] in STEM areas to support the educational process and evaluation (through Jupyter-related packages such as nbgrader [20, 21]).

However, to the extent of our knowledge, due to the novelty of the field, there is still a lack of secondary studies that provide a systematic overview of the state of the art that allows the clear identification of lines of research.

This paper deals with these issues and related aspects aiming to fill a research gap through a systematic literature review (SLR) and mapping study [22] focused on how artificial intelligence or similar algorithmic methods have been used for the educational assessment of students.

To do so, the article is composed of the following sections: Sect. 2 describes the methodology employed to perform the SLR, including the research questions, the search string, the repositories and the selection process; Sect. 3 exposes the results of the SLR answering the research questions, and, finally, Sect. 4 provides a discussion on the most relevant findings.

2 Methodology

As we have seen, this study combines two different techniques for literature review, namely a systematic literature review and a mapping study. These two techniques have in common the structured and planned examination of large amounts of information in a replicable way [23]. However, while the main objective of mapping studies is to provide a general overview focused on characteristics external to the content of publications such as the geographical distribution or the evolution of the number of studies over time, SLRs aim to answer more in-depth questions related to the areas of study, the variables used or the methodology among others[22, 24].

In consequence, this investigation employs a combination of this two techniques aiming to answer the following six research questions.

- 1. What has been the evolution of the number of documents covering the use of AIs in student assessment?
- 2. Who are the most relevant authors in the field?
- 3. What is the most common source for the dissemination of results in the field?
- 4. Where do the studies take place?
- 5. Which is the population of the study?
- 6. Which are the most frequent uses of AIs in student assessment?

Following the proposal by Kitchenham and Charters [22], the following inclusion criteria were designed to guide the selection process and ensure the adequacy of the papers selected:

- 1. The research papers are related to the use of IAs for student assessment.
- 2. The research papers include empirical research.
- 3. Research papers have been published after being subjected to a peer review process.
- 4. Research papers are written in English.

If a study did not meet all the inclusion criteria it was excluded from the research during the selection process.

In order to identify the papers in the databases we employed the following search string considering both the research questions and the possible alternative spellings and synonyms: (AI OR "artificial intelligence") AND ("students assessment" OR "students evaluation" OR "assessment of the students" OR "evaluation of the students").

This search string was introduced in the two databases: SCOPUS and Web of Science (WOS). This two databases were selected considering that both of them allow the use of logical expressions, searches in specific fields or full-length searches and are relevant to the field. Additionally the use of this two databases is considered enough to perform valid SLR and mapping studies [25–27]. The introduction of the search string on the two databases provided an initial number of 120 studies (109 from SCOPUs and 20 from WOS) on December of 2019.

These initial results were recorded on a spreadsheet and, after the removal of the duplicates, were subjected to the inclusion criteria screening the titles abstracts and keywords of the papers. When such elements did not include enough information the researchers also screened the full text of the publications (Fig. 1).

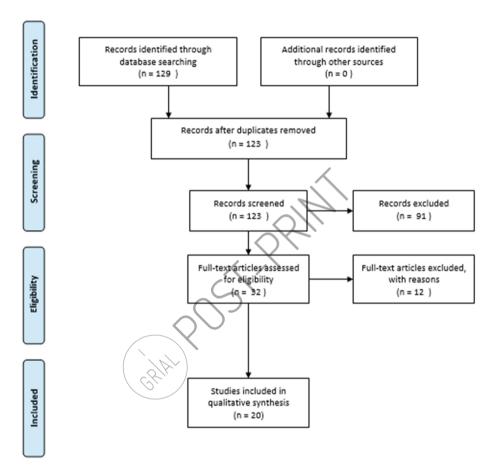


Fig. 1. Selection process. Reported in line with the PRISMA statement [28]

3 Results

The application of the inclusion and exclusion criteria yielded a total of 20 papers on the topic of the use of AI for student assessment. This section presents the analysis of the characteristics of the selected documents in order to answer the research questions posed in the method section, both for the mapping and the SLR.

Q1: What has been the evolution of the number of documents covering the use of AIs in student assessment?

Regarding the evolution of the research field, we can see in figure X that, while there is not a clear upward trend, there does seem to be an evolution towards more research in the past years, especially since 2017, with 14 out of the 20 papers (70%) published within the last three years (Fig. 2).

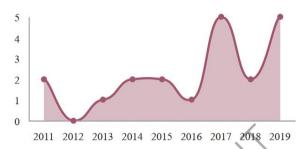


Fig. 2. Year of publication of the studies

Q2: Who are the most relevant authors in the field?

There were 66 different authors in the 20 selected documents, but none of them had authored more than one paper, therefore no author stood out as a prominent researcher in the field.

Q3: What is the most common source for the dissemination of results in the field?

Most of the authors chose to disseminate the results of their research through the elaboration of conference papers (70%, 14 documents), published in the form of proceedings from computer science and AI-based conferences. The rest of the documents (60%, 6 documents) were papers published in scientific journals. While all the papers come from different journals, all of them are from the science and engineering field.

Q4: Where do the studies take place?

Although there is a large variety of countries where the selected studies take place, there are two countries that stand out: the United States of America, with 5 papers [29–33], China with 3 papers [34–36], and Pakistan with 2 papers [37, 38].

Figure 2 shows the dispersion of studies conducted in the European region, which are located in Greece [39], Macedonia [40], Norway [41], Poland[42], Serbia [43] and the United Kingdom [44].

Other countries represented in the selection are Mexico [45] in North America, Turkey [46] in Asia, Mauritius [47] in Africa, and Fiji [48] in the Pacific region¹.

¹ The last two countries are not represented in the map in Fig. 3 due to their small size.

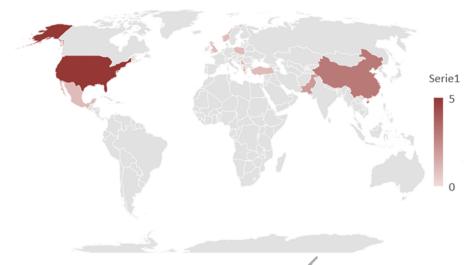


Fig. 3. Countries where the studies took place

Q5: Which is the population of the study?

Out of all 20 papers, 14 were focused on Higher Education students and 1 focused on Secondary Education Students [39]. The other 5 were not aimed at a particular group of students, since they were theoretical in nature [32, 40, 41, 44, 47].

Q6: Which are the most frequent uses of AIs in student assessment?

Aiming to characterize the main aim of the papers analyzed for this study, three thematic categories emerged: assessment of student behaviors, assessment of student sentiments and assessment of student achievement.

There were four papers focused on the study of student behaviors with different aims. Two studies analyzed the social interactions performed by students within the framework of an online course in order to predict their final outcome or achievement in the course [32, 48]. The other two used information on the behavior of students during the performance of a given task in order to personalize their user experience. In one of them, the aim was to provide the student with useful assistance during the task [45] and the other aimed to tailor an enemy bot during a gaming experience [43].

Six studies were focused on student feelings or sentiments, aiming to assess and determine the emotional state of the students when performing a given task or participating in a given course. The aim of four of them was to analyze the students' opinions on the teaching and learning processes and models developed in the course [31, 33, 34, 37, 42], and the other one intended to use the data on student emotions to predict their chances of finishing a MOOC and subsequently provide personalized messages to encourage the students who are less likely to finish the course based on their emotions [36].

There was another block of papers (10) whose main focus was the assessment of student achievement through AI-based methods. More than half of these papers were

focused on the automatization of the correction of different types of student productions in order to assign a grade to their work. There were two studies that applied AIs to multiple-choice tests aiming to improve the fairness of the correction process [41] or to implement an adaptive testing system, where each student gets questions tailored to their answers to previous ones [35]. Another study intended to automatize the correction of short answers to structured questions in order to reduce the instructors' time spent in grading student work [47]. Lastly, there were three studies aimed at grading productions other than tests. Two of them focused on grading the results of lab exercises [29, 46] and another one developed an AI in order to assess concept maps elaborated by the students.

Other researchers used AIs to analyze secondary achievement data from the students (previous grades obtained in tests and/or assignments) in order to either predict the students' final score in the course [38–40] or to categorize the students' work throughout the course according to a pre-established set of grade categories (A, B, C, etc.) [44].

Figure 4 shows a summary of the categories and subcategories detected in the study.

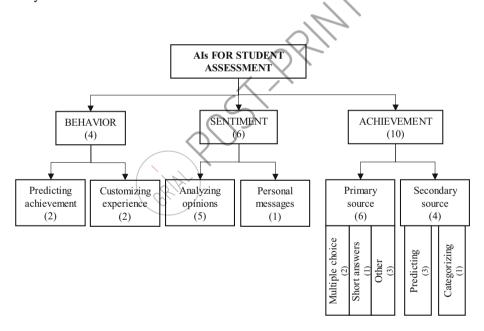


Fig. 4. Summary of AI uses in student assessment

4 Discussion and Conclusions

This literature review has served to address some issues regarding the status of research on the use of AIs for student assessment.

Firstly, the answer to the first research question points out that the research in this field is still at a very early stage of development given that all the selected works have been published in the last decade and more than half of them in the last five years.

This short period of development may have resulted in a wide dispersion both in the number of publications per author and in the geographical distribution, although USA and China concentrate most of the research.

In the same vein, it is striking that all publications have been made in conferences and journals in the field of computer engineering existing a lack of studies focused on the use of AIs for the assessment of the students from a pedagogical perspective. This is a usual phenomenon in the early stages of integration of any technological innovation in the educational field [49] and constitutes a very interesting line of research for future studies.

Additionally, all of the studies except one are focused on the higher education level, therefore extending the research to other formal education levels may constitute another line for the future development of investigations.

In the answer to research question 6, three major thematic areas have been identified, focusing on the study of student behavior, feelings and achievement. Future studies should extend this lines including the use of IAs in more ambitious didactic designs, paying attention to the assessment of different competences or the detection of educational needs.

Finally, this study presents some limitations. In the first place, although SCOPUS and WoS are the more relevant databases of scientific publications they may not include some works of interest published in less relevant journals or conference proceedings, therefore future studies could consider the use additional databases such as Google Scholar in order to diminish the effect of the publication bias.

Secondly, this review has only considered the direct results obtained in the two databases as a first approach to the field, future studies may find of interest to include additional records identified through other sources to increase the number of publications analyzed.

Acknowledgement. This work has been partially funded by the Spanish Government Ministry of Economy and Competitiveness through the DEFINES project (Ref. TIN2016-80172-R).

References

- Makridakis, S.: The forthcoming Artificial Intelligence (AI) revolution: its impact on society and firms. Futures 90, 46–60 (2017)
- Roll, I., Wylie, R.: Evolution and revolution in artificial intelligence in education. Int. J. Artif. Intell. Educ. 26, 582–599 (2016). https://doi.org/10.1007/s40593-016-0110-3
- Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). http://www.darpa.mil/program/explainable-artificial-intelligence. Accessed 24 Feb 2020

- Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–18. ACM, April 2018
- 5. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: the role of humans in interactive machine learning. Ai Mag. 35, 105–120 (2014)
- 6. Cruz-Benito, J.: On data-driven systems analyzing, supporting and enhancing users' interaction and experience. Doctoral dissertation, Universidad de Salamanca (2018)
- 7. O'neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books. New York (2014)
- 8. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J. Ethics **21**, 403–418
- (2017). https://doi.org/10.1007/s10892-017-9252-2
 9. Russell, S., et al.: Letter to the editor: Research priorities for robust and beneficial artificial intelligence: an open letter. AI Mag. 36, 3–4 (2015)
- 10. Peirano, M.: El enemigo conoce el sistema: Manipulación de ideas, personas e influencias
- después de la economía de la atención. Debate (2019) 11. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif.
- Intell. **267**, 1–38 (2019)

 12. Aoun, J.E.: Robot-Proof: Higher Education in the Age of Artificial Intelligence. MIT Press,
- Cambridge (2017)

 13. Henrie, C.R., Halverson, L.R., Graham, C.R.: Measuring student engagement in technology-
- mediated learning: a review. Comput. Educ. **90**, 36–53 (2015)

 14. Jonassen, D., Davidson, M., Collins, M., Campbell, J., Haag, B.B.: Constructivism and computer-mediated communication in distance education. Am. J. Distance Educ. **9**, 7–26
- (1995)
 15. Perrotta, C., Williamson, B.: The social life of learning analytics: cluster analysis and the 'performance' of algorithmic education. Learn. Media Technol. 43, 3–16 (2018)
- 16. Papamitsiou, Z., Economides, A.A.: Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. J. Educ. Technol. Soc. 17, 49–
- 64 (2014)
 17. Roll, I., Winne, P.H.: Understanding, evaluating, and supporting self-regulated learning using learning analytics. J. Learn. Anal. 2, 7–12 (2015)
- 18. Rienties, B., Cross, S., Zdrahal, Z.: Implementing a learning analytics intervention and evaluation framework: what works? In: Kei Daniel, B. (ed.) Big Data and Learning Analytics in Higher Education, pp. 147–166. Springer, Cham (2017). https://doi.org/10.
- 1007/978-3-319-06520-5_10
 19. Kluyver, T., et al.: Jupyter notebooks-a publishing format for reproducible computational workflows. In: Loizides, F., Schmidt, B. (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90. IOS Press, Canada (2016)
- Hamrick, J.B.: Creating and grading IPython/Jupyter notebook assignments with NbGrader.
 In: Alphonce, C., Tims, J. (eds.) Proceedings of the 47th ACM Technical Symposium on Computing Science Education, p. 242. ACM Press, New York (2016)
- 21. Blank, D. S., Bourgin, D., Brown, A., Bussonnier, M., Frederic, J., Granger, B.,... Page, L. nbgrader: A tool for creating and grading assignments in the Jupyter Notebook. The Journal of Open Source Education 2, 32–34 (2019)
- Kitchenham, B., Charters, S.: Guidelines for performing Systematic Literature Reviews in Software Engineering. Version 2.3 (EBSE-2007–01) (2007). https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf. Accessed 24 Feb 2020

- 23. Cruz-Benito, J., García-Peñalvo, F.J., Therón, R.: Analyzing the software architectures supporting HCI/HMI processes through a systematic review of the literature. Telematics and
- Inform. 38, 118–132 (2019) 24. Kitchenham, B.A., Budgen, D., Brereton, P.O.: Using mapping studies as the basis for further research – a participant-observer case study. Inf. Softw. Technol. 53, 638–651 (2011) 25. Kitchenham, B.: What's up with software metrics? – a preliminary mapping study. J. Syst.
- Softw. 83, 37-51 (2010) 26. Neiva, F.W., David, J.M.N., Braga, R., Campos, F.: Towards pragmatic interoperability to
- support collaboration: a systematic review and mapping of the literature. Inf. Softw. Technol. **72**. 137–150 (2016)
- 27. García Sánchez, F., Therón, R., Gómez-Isla, J.: Alfabetización visual en nuevos medios: revisión y mapeo sistemático de la literature. Educ. Knowl. Soc. 20, 1–35 (2019)
- 28. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6, 1-6 (2009)
- 29. Floryan, M., Dragon, T., Basit, N., Dragon, S., Woolf, B.: Who needs help? Automating student assessment within exploratory learning environments. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 125-134. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-1/9773-9_13
- 30. Gurupur, V.P., Pankaj Jain, G., Rudraraju, R.: Evaluating student learning using concept maps and Markov chains. Expert Syst. Appl. 42, 3306-3314 (2015) 31. Newman, H., Joyner, D.: Sentiment analysis of student evaluations of teaching. In: Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-
- Pomsta, K., McLaren, B., du Boulay, B. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 246–250. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2 45 32. Ma, J., Kang, J.-H., Shaw, E., Kim, J.: Workflow-based assessment of student online activities with topic and dialogue role classification. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 187–195. Springer,
- Heidelberg (2011). https://doi.org/10.1007/978-3-642-21869-9_26 33. Tzacheva, A., Ranganathan, J., Jadi, R.: Multi-label emotion mining from student comments. In: Proceedings of the 2019 4th International Conference on Information and Education
- Innovations, pp. 120–124. ACM, New York (2019) 34. Lin, Q., Zhu, Y., Zhang, S., Shi, P., Guo, Q., Niu, Z.: Lexical based automated teaching
- evaluation via students' short reviews. Comput. Appl. Eng. Educ. 27, 194–205 (2019) 35. Wang, M., Wang, C., Lee, C., Lin, S., Hung, P.: Type-2 fuzzy set construction and application for adaptive student assessment system. In: Proceedings of the 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 888–894. IEEE (2014)
- 36. Wang, L., Hu, G., Zhou, T.: Semantic analysis of learners' emotional tendencies on online MOOC education. Sustainability **10**, 1–19 (2018)
- 37. Akhtar, J.: An interactive multi-agent reasoning model for sentiment analysis: a case for computational semiotics. Artif. Intell. Rev., 1–18 (2019). https://link.springer.com/article/10. 1007/s10462-019-09785-6#citeas
- 38. Mahboob, T., Irfan, S., Karamat, A.: A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms. In: Proceedings of the 19th International Multi-Topic Conference (INMIC), pp. 1-8. IEEE (2017)
- 39. Livieris, I.E., Drakopoulou, K., Kotsilieris, T., Tampakas, V., Pintelas, P.: DSS-PSP a decision support software for evaluating students' performance. In: Boracchi, G., Iliadis, L., Jayne, C., Likas, A. (eds.) EANN 2017. CCIS, vol. 744, pp. 63–74. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65172-9_6

- Simjanoska, M., Gusev, M., Bogdanova, A.M.: Intelligent modelling for predicting students' final grades. In: Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1216–1221. IEEE (2014)
- 41. Hameed, I.: A fuzzy system to automatically evaluate and improve fairness of multiple-choice questions (MCQs) based exams. In: Proceedings of the 8th International Conference on Computer Supported Education Volume 1: CSEDU, pp. 476–481. SciTePress (2016)
- 42. Dudek, D.: Survey analyser: effective processing of academic questionnaire data. In: Borzemski, L., Świątek, J., Wilimowska, Z. (eds.) ISAT 2018. AISC, vol. 852, pp. 245–257. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-99981-4_23
- 43. Kuk, K., Milentijević, I.Z., Ranđelović, D., Popović, B.M., Čisar, P.: The design of the personal enemy MIMLebot as an intelligent agent in a game-based learning environment. Acta Polytechnica Hungarica 14, 121–139 (2017)
- 44. Boongoen, T., Shen, Q., Price, C.: Fuzzy qualitative link analysis for academic performance evaluation. Int. J. Uncertainty Fuzziness and Knowl.-Based Syst. 19, 559–585 (2011)
- Zatarain-Cabada, R., Barrón-Estrada, M.L., Ríos-Félix, J.M.: Affective learning system for algorithmic logic applying gamification. In: Pichardo-Lagunas, O., Miranda-Jiménez, S. (eds.) MICAI 2016. LNCS (LNAI), vol. 10062, pp. 536–547. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62428-0_44
- 46. Caliskan, E., Tatar, U., Bahsi, H., Ottis, R., Vaarandi, R.: Capability detection and evaluation metrics for cyber security lab exercises. In: Bryant, A.R., Mills, R.F., Lopez, J. (eds.) Proceedings of the 2017 International Conference on Cyber Warfare and Security, pp. 407–414. Academic Conferences and Publishing International Ltd., UK (2017)
- 47. Luchoomun, T., Chumroo, M., Ramnarain-Seetohul, V.: A knowledge based system for automated assessment of short structured questions. In: Proceedings of the 2019 IEEE Global Engineering Education Conference (EDUCON), pp. 1349–1352. IEEE (2019)
- 48. Singh, S., Lal, S.P.: Educational courseware evaluation using Machine Learning techniques. In: Proceedings of the 2013 IEEE Conference on e-Learning, e-Management and e-Services, pp. 73–78. IEEE (2013)
- 49. Petrova, K., Li, C.: Focus and setting in mobile learning research: a review of the literature. Commun. IBIMA 10 (219-226 (2009)