Samuelsen *et al. Research and Practice in Technology Enhanced Learning*
(2019) 14:11
https://doi.org/10.1186/s41039-019-0105-4

Research and Practice in
Technology Enhanced Learning

RESEARCH Open Access

Integrating multiple data sources for learning analytics—review of literature

Jeanette Samuelsen[1,2*], Weiqin Chen[1,3] and Barbara Wasson[1,2]

* Correspondence: Jeanette.Samuelsen@uib.no
[1]Centre for the Science of Learning & Technology, University of Bergen, Bergen, Norway
[2]Department of Information Science & Media Studies, University of Bergen, Bergen, Norway
Full list of author information is available at the end of the article

Abstract

Learning analytics (LA) promises understanding and optimization of learning and learning environments. To enable richer insights regarding questions related to learning and education, LA solutions should be able to integrate data coming from many different data sources, which may be stored in different formats and have varying levels of structure. Data integration also plays a role for the scalability of LA, an important challenge in itself. The objective of this review is to assess the current state of LA in terms of data integration in the context of higher education. The initial search of six academic databases and common venues for publishing LA research resulted in 115 publications, out of which 20 were included in the final analysis. The results show that a few data sources (e.g., LMS) appear repeatedly in the research studies; the number of data sources used in LA studies in higher education tends to be limited; when data are integrated, similar data formats are often combined (a low-hanging fruit in terms of technical challenges); the research literature tends to lack details about data integration in the implemented systems; and, despite being a good starting point for data integration, educational data specifications (e.g., xAPI) seem to seldom be used. In addition, the results indicate a lack of stakeholder (e.g., teachers/instructors, technology vendors) involvement in the research studies. The review concludes by offering recommendations to address limitations and gaps in the research reported in the literature.

Keywords: Learning analytics, Higher education, Data integration, Multiple data sources, Interoperability, Scalability

Introduction

Learning analytics (LA) includes collecting, computationally analyzing, and reporting data to stakeholders, to gain insights and enable decision-making and interventions related to questions about learning and learning environments (Siemens, 2011). The data can be stored in different formats, have varying levels of structure, and originate from different sources. When data from multiple sources are collected and merged, i.e., the data are integrated, they may reflect the dispersed activities of learners in a more precise way than what is possible for each individual data source (Chatti, Muslim, & Schroeder, 2017). In addition, data integration can also lead to more useful analysis, since many LA techniques require large-scale and possibly diverse data (Cooper & Hoel, 2015). Data integration is one contributing factor to the scalability of LA. Factors important for scaling up LA include the technical solution but also factors such as organizational hierarchy

(Buckingham Shum & McKay, 2018), management structures (Dawson et al., 2018), policy and regulations (European Union, 2016).

Data integration is closely related to interoperability, which involves technical, semantic, legal, and organizational levels. The *semantic level* is about ensuring that data format and meaning is preserved and understood. The *technical* level includes services for data exchange and data integration. The *legal level* addresses aspects such as enabling collaboration despite different legal frameworks and organizational policies. With regard to this interoperability level, an important factor is to protect the privacy of users. Finally, *organizational* interoperability includes aligning processes for common organizational goals and addressing user expectations and requirements (European Commission, 2017).

Data integration and interoperability are general challenges when working with Big Data and are not specific to any one domain (Kadadi, Agrawal, Nyamful, & Atiq, 2014). In an educational setting, collecting and combining data from multiple sources can help provide insights that have implications for areas such as learning, instruction, retention, and curriculum design. The data collected are often activity data (i.e., data from learners' activity in digital learning environments). As stated by Chatti et al. (2017), adopting widely used specifications is important for interoperability. Within education and LA, two well-known educational data specifications are xAPI (2019) and IMS Caliper Analytics (2019). These two specifications enable the exchange of data among different applications, and integration of data from multiple sources in a central data store. In addition, the specifications provide a specific format and syntax to describe learning events that have occurred in learning environments; and, they enable (but do not enforce) the use of controlled vocabularies, which can be used to define concepts and relationships for describing and representing a domain of interest (Allemang & Hendler, 2011). Thus, the two specifications target the technical and, at least to some degree, semantic interoperability levels.

In most cases, only a limited number of data sources are used in combination for data analysis in the context of LA in higher education. Existing LA projects addressing data integration needs tend to place emphasis on the technical level of interoperability (Apereo, 2018; JISC, 2019; OnTask, 2019). The new EU general data protection regulation (European Union, 2016) addresses, to a large degree, legal and organizational concerns. Semantic interoperability, enabling shared data meaning, is typically less emphasized, even though it can enable more effective merging of data through reuse of common data specifications. While all the interoperability levels are important for data integration, the focus in this review is on the semantic and technical levels.

In this article, we report on a systematic literature review that examines publications on LA research studies in higher education that use and/or combine data from multiple sources. The reason for focusing on higher education as the domain of study is twofold: (1) higher education is the working context of the authors and the datasets we are using in our studies come from a higher education setting and (2) previous research has found that the majority of LA research is conducted in higher education (Misiejuk & Wasson, 2017); thus, this restriction of scope would not be expected to significantly affect the number of publications included in the review.

The aim of our research is to assess the current state of LA in terms of data integration in the context of higher education. As a conclusion, we identify shortcomings in existing research and provide recommendations to address such limitations and gaps.

Review method and results

This systematic review follows the guidelines by Kitchenham and Charters (2007). The following review stages will be detailed: planning the review, conducting the review, and data synthesis and reporting the results.

Planning the review

In the planning phase, we clarified the needs for a review, by going through relevant existing reviews and state-of-the-field reports (e.g., Misiejuk & Wasson, 2017; Sclater, Peasgood, & Mullan, 2016; Shahiri, Husain, & Rashid, 2015) and identified gaps in the knowledge base. This process resulted in the following questions that this review seeks to answer:

- *RQ1.* What types of data and data sources are being used for LA in higher education?
- *RQ2.* How and to what extent are different data being used/combined for LA research in higher education?
- *RQ3.* What methods are used, and what issues are being addressed through using/combining multiple data sources?

Conducting the review

Six academic databases were initially selected for search: ACM Digital Library, IEEE Xplore, SpringerLink, Science Direct, Wiley, and AISEL. Later, we added common venues for publishing LA and the related field of educational data mining (EDM) research as additional sources: Journal of Learning Analytics (jLA), Journal of Educational Data Mining, International Conference on Learning Analytics & Knowledge, International Conference on Educational Data Mining, and Learning at Scale (L@S) conference. For journals, issues from 2014 (when jLA was first published) and later were included. For conferences, we included proceedings from 2017, 2018 and 2019 (the most recent conferences, as research in earlier conferences would most likely have been published in journals by this time).

The search string was constructed based on the research questions, as follows:

> ("multiple data sources" OR multimodal OR "multi-modal" OR "multiple data sets" OR "multiple datasets") AND ("learning analytics" OR "educational data mining") AND "higher education"

As can be seen, the search string encompasses relevant fields (LA and EDM), context (higher education), and concepts (multiple data sources).

Before beginning the systematic literature review, we defined inclusion and exclusion criteria. These criteria were based on the research questions and helped with the assessment of the relevance of each publication for the review. The inclusion and exclusion criteria are listed in Tables 1 and 2 below.

Conducting the review included the following steps: identification, screening, eligibility, and inclusion. Figure 1 shows the results of each step using a PRISMA flow diagram (Moher, Liberati, Tetzlaff, Altman, & PRISMA group, 2009).

The initial search of academic databases was conducted on September 5, 2017. Forty-nine results (journal articles) were returned from the academic databases, no duplicates were among the results. Two researchers judged the relevance of the publications for the

Table 1 Inclusion criteria

Inclusion criteria	Explanation
Publication domain	The domain of the study must be higher education or massive open online courses (MOOCs).
Publication type	Studies must be peer-reviewed and published in a journal/conference proceedings.
Data source	The research must make use of multiple data sources.
Implementation	Publications must include frameworks for combining data sources or details about implemented systems using/combining multiple data sources (not just proposed design and/or architecture).
Language	Publications must be written in English.

systematic review in two phases. First, the title, abstract, and keywords were read for each publication, to filter out results that were not relevant for the review based on the inclusion and exclusion criteria. This initial screening resulted in 26 publications being excluded. Second, the remaining 23 results were assessed for eligibility, based on reading the full-text of the publications and checking this against the inclusion and exclusion criteria. This phase removed 17 publications and resulted in six for inclusion in the review. On April 18, 2019, to include publications that had been published since the initial search, an identical search string was applied to search the same academic databases. This search yielded 32 new results (journal articles). During the initial screening, 19 publications were excluded. The remaining 13 publications were assessed for eligibility, leading to the removal of eight publications. Thus, five new publications were included in the review following this newer search.

Additional literature was downloaded from online repositories for common venues for publishing LA and EDM research, between July 24 and July 26, 2018. The downloaded literature was indexed based on full-text PDF files, thereafter the review search string was applied on the indexed literature. Indexing and search were done using Zotero (reference management software). Initially, the search string resulted in 22 publications (journal articles and conference papers). Assessing the relevance of the publications for the systematic review, 11 publications were removed during the initial screening; then, three publications were removed during full-text assessment for eligibility. Thus, eight additional publications were found eligible for inclusion. On April 18, 2019, we added all new research from the LA and EDM venues to Zotero (i.e., new conference proceedings and journal issues), before re-applying the full-text search. This search gave 12 new results. Out of the resulting publications, eight were removed during the initial screening, and three were removed when assessing the publications for eligibility. Thus, this newer search of LA and EDM venues resulted in one new publication being included in the systematic review.

In total, there are 20 publications included in the review. They are listed in Table 3.

Table 2 Exclusion criteria

Exclusion criteria	Explanation
Data analysis	Publications that primarily focus on data analysis methods and algorithms, but have little information regarding data, are excluded.
Publication type	Literature reviews are excluded.

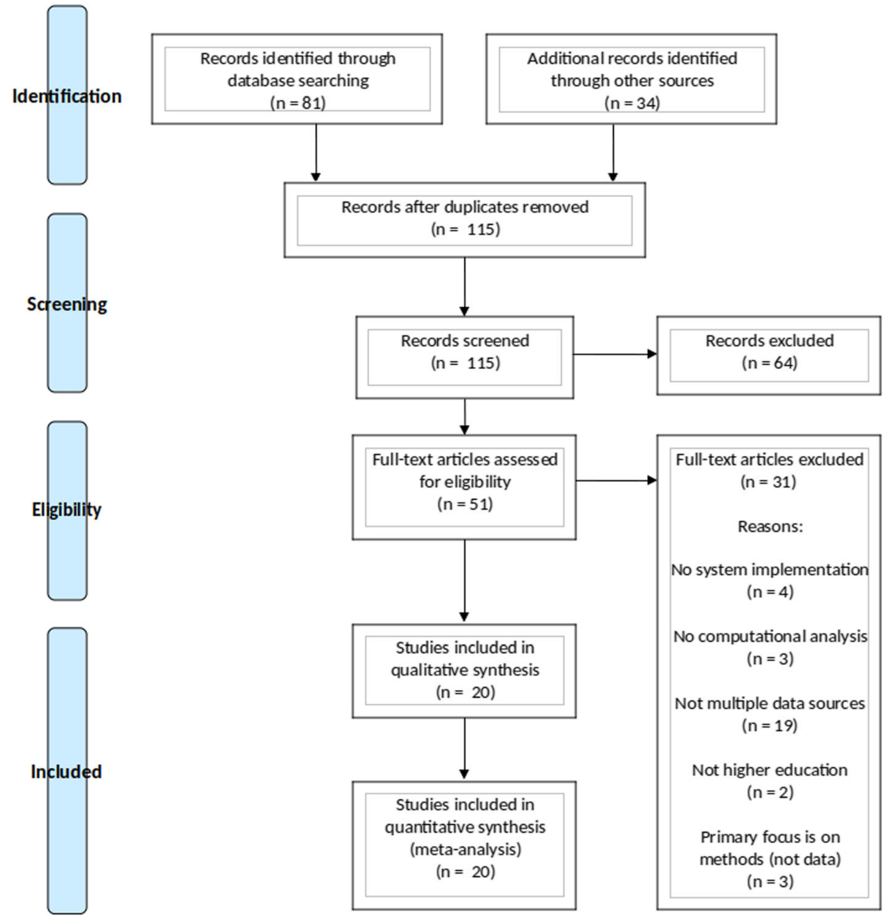
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 81)
Additional records identified through other sources (n = 34)
Records after duplicates removed (n = 115)
Records screened (n = 115)
Records excluded (n = 64)
Full-text articles assessed for eligibility (n = 51)
Full-text articles excluded (n = 31)
Reasons:
No system implementation (n = 4)
No computational analysis (n = 3)
Not multiple data sources (n = 19)
Not higher education (n = 2)
Primary focus is on methods (not data) (n = 3)
Studies included in qualitative synthesis (n = 20)
Studies included in quantitative synthesis (meta-analysis) (n = 20)

Fig. 1 Review process diagram

Data synthesis and reporting the results

The publications included in the review were read, reviewed (using data extraction forms to record details for every relevant criterion) and synthesized. The following criteria were assessed for each publication (see analysis of the articles in Table 3):

- Types of data sources
- Types of data
- Data sources integrated?
- Data integration approach (manual or automatic)
- System supports educational data specifications?
- Issue addressed in publication
- Methods used
- Records for how many participants analyzed
- Data storage technology (see Table 7)

Integration of data sources was judged based on the information in the publications' text about integration, or (if the former was not available) which data were analyzed in combination to address a specific issue. Similarly, we looked for clues

Table 3 Overview of the reviewed publications

Publication	Issue addressed in publication	Types of data sources	Types of data	Data sources integrated?	Data integration approach (manual or automatic)	Methods used	Records for how many participants analyzed
Lopez Guarin, Guzman, and Gonzalez (2015)	Predict the loss of academic status at a certain time	Student information system (× 2)	Student background information (× 2), performance test data, final grades	Yes	Automatic—first join admissions data sets into one table. Then join with academic information	Decision trees, naive Bayes	1532 students
Park, Yu, and Jo (2016)	Classify blended learning courses in a Korean higher education institution	LMS (× 2)	Activity log, course data	Yes	Automatic (most likely—not explicitly stated). Combine course data and log data on course ID (anonymized)	Latent class analysis	N/A (Records regarding 612 courses which were found suitable for analysis)
Thompson, Kennedy-Clark, Wheeler, and Kelly (2014)	Automatic tagging of text part of speech; for the identification of types of micro-events that learners enact, and the determination of whether learners complete functions that are crucial for task success	Corpora (× 2) (both are mini-corpora of collaborative problem-based learning activities)	Text (× 2)	No (data are analyzed separately)	Data are not integrated	Part of speech tagger—trained on Penn Tree Bank corpora. Visualization of timing and speaker for each utterance in one mini corpora.	Corpora 1: total 6 dyads (12 students + teacher) Corpora 2: four postgraduate students
Zheng, Bender, and Nadershahi (2017)	Data were extracted from tools to provide data on faculty's application of digital tools and to assess the impact of the lecture annotation tool on students' learning behavior	LMS, lecture annotation tool	Activity log data (× 2)	No (data are analyzed separately)	Data are not integrated	N/A	N/A

Table 3 Overview of the reviewed publications *(Continued)*

Publication	Issue addressed in publication	Types of data sources	Types of data	Data sources integrated?	Data integration approach (manual or automatic)	Methods used	Records for how many participants analyzed
Pardos and Kao (2015)	Bayesian network analysis to assess student current and prior knowledge for problems in a MOOC (with visualization); and visualization of course structure (not based on preceding analysis)	MOOC (× 2)	Activity log, student background information (possibly more)	No, platform can currently only integrate EdX MOOC data with other EdX MOOC data. Platform also supports Coursera	Automatic for integrating EdX MOOC data with other EdX data (for Coursera MOOC data this is not addressed). Approach: use HarvardX tool to integrate different types of EdX files into one csv file (loosely based on xAPI). For visualizations: read csv file(s) into memory	Bayesian network analysis, visualization	N/A
Liu et al. (2017)	Examine use of an adaptive system through analysis of usage patterns	Student information system, adaptive platform, LMS, performance test	Student background information, activity log, performance test data (× 3)	Yes	N/A (publication explicitly mentions combination of data, yet does not specify how)	Spearman correlation, visualizations, regression analyses	128 first-year students entered into pharmacy program
Raca, Tormey, and Dillenbourg (2016)	Compare student behaviors (levels of movement) and connect with attention (self-reported)	Video, questionnaire	Video-derived data, questionnaire data	Yes	N/A	Descriptive statistics (e.g., mean, percentage), correlations	56 bachelor level students
Di Mitri et al. (2017)	Predict learners performance during self-regulated learning	Physiological signals wristband, software tracking tool, questionnaire, weather information	Physiological arousal data, software category, questionnaire data, location data, weather data	Yes	Automatic. A tool (Learning Pulse Server) imports data from different APIs and stores events in a Learning Record Store (xAPI format)	Linear mixed effects models	9 PhD students (the multimodal data set originally contained approximately 10,000 records)
Ochoa et al. (2018)	Provide automatic feedback on oral presentation skills	Video, audio, presentation slide	Video derived data, audio derived data, presentation slide derived data	No (data are analyzed separately)	Data/data sources are not integrated	Various classification algorithms (e.g., random forest)	83 engineering students

Table 3 Overview of the reviewed publications *(Continued)*

Publication	Issue addressed in publication	Types of data sources	Types of data	Data sources integrated?	Data integration approach (manual or automatic)	Methods used	Records for how many participants analyzed
Hutt et al. (2017)	Detect mind wandering during a lecture using eye tracking	Eye tracker, questionnaire	Eye tracker data, questionnaire data	Yes	N/A	Bayesian network classifier	32 undergraduate students from a Canadian university
Jayaprakash, Moody, Lauría, Regan, and Baron (2014)	Detect students who are in academic difficulty	LMS, student information system	Activity log data, partial course grades, course data, student background information (× 2)	Yes	Automatic. Uses Pentaho Business Intelligence Data Integration (ETL approach)	Logistic regression, support vector machines, J48, naive Bayes	15,150 undergraduate students
Rodríguez-Triana, Prieto, Martínez-Monés, Asensio-Pérez, and Dimitriadis (2018)	Identify deviations between the desired learning state (based on learning design) and the actual state in blended/CSCL scenarios	LMS, wiki, online writing application, attendance list, human observation, instructional design information, questionnaire	Activity log data (× 2), attendance information, teacher comments, instructional design information, questionnaire data	Yes	Automatic (at least in part). Third-party tools were integrated into virtual learning environment (GLUE)	N/A (three binary classifiers were built to identify deviations between desired learning state and actual state)	165 students
Gray, McGuinness, Owende, and Hofmann (2016)	Predict at-risk students	Student information system, questionnaire, exam results	Student background information, questionnaire data, GPA	Yes	N/A	Correlations, *t* test/ANOVA. Classification (e.g., naive Bayes, decision trees)	1207 first-year students (records from 2010 to 2012)
Wang, Paquette, and Baker (2014)	Identify career path for MOOC learners	MOOC, organization member information	Student background information, questionnaire, organization member information	Yes (partly, questionnaire is analyzed separately)	N/A (most likely manual)	Chi-square, descriptive statistics	N/A (536 MOOC participants answered questionnaire)
Mangaroska, Vesin, and Giannakos (2019)	Predict student performance	E-learning portal (× 2), Integrated Development Environment (IDE)	Performance test data, activity log data (× 3)	Yes	Automatic. System collects and aggregates data from different sources. Data are integrated in a Learning Record Store	Descriptive statistics, Spearman correlation, linear regressions, visualization	21 (one teacher and 20 computer science students)

Table 3 Overview of the reviewed publications *(Continued)*

Publication	Issue addressed in publication	Types of data sources	Types of data	Data sources integrated?	Data integration approach (manual or automatic)	Methods used	Records for how many participants analyzed
Villano, Harrison, Lynch, and Chen (2018)	Examine the relationship between student retention and an early alert system (controlling for a number of variables)	Student information system, early alert system	Student background information, final grades, workload, school data (e.g., location, fee), early alert system data	Yes	Automatic. University collects and integrates data from different IT systems in a data warehouse	Survival analysis	N/A (16,142 records captured from 2011 to 2013 were analyzed)
Wong, Kwong, and Pegrum (2018)	Examine if an augmented reality app for integrity and ethics can help change student's perspectives on these subject matters	AR platform, LMS	Activity log data, text (× 2)	No (data are analyzed separately)	Data/data sources are not integrated	Descriptive statistics, text analysis, visualization	N/A (1259 students participated, but not all participants' data were included in the subsequent analyses)
Sandoval, Gonzalez, Alarcon, Pichara, and Montenegro (2018)	Prediction of students who are at risk of failing classes	Student information system, LMS	Student background information, final grades, activity log data	Yes	Automatic. Extract data from data sources and encrypt, then re-codify some of the attributes into similar types before integrating in a relational database	Linear regressions, random forest	21,314 students (over three semesters)
Sun, Xie, and Anderman (2018)	Examine the effect of self-regulation on academic achievement in flipped classrooms	Questionnaire, LMS	Questionnaire data, performance test data, partial course grades	Yes	Manual. Combine grades obtained from instructors with survey data	Structural equation modeling, multi-level regression	151 US undergraduate students
Giannakos, Sharma, Pappas, Kostakos, and Velloso (2019)	Examine if including physiological sensing data provides advantages for predicting skill acquisition (and more generally for the design of learning technologies)	Eye tracker, physiological signals wristband, EEG cap, video, game	Eye tracker data, physiological arousal data, EEG data, video derived data, activity log data, performance test data	Yes	Automatic. The features for each data source were extracted separately, then data were integrated using R	LASSO regression, random forest, ANOVA	17 participants from a major European university

Table 4 Data sources and data integration

Observation	Frequency	Publications
Multiple data sources, same format	14	Lopez Guarin et al. (2015), Park et al. (2016), Di Mitri et al. (2017), Rodríguez-Triana et al. (2018), Liu et al. (2017), Gray et al. (2016), Hutt et al. (2017), Jayaprakash et al. (2014), Raca et al. (2016), Mangaroska et al. (2019), Villano et al. (2018), Sandoval et al. (2018), Sun et al. (2018), Giannakos et al. (2019)
Multiple data sources, different formats	6	Thompson et al. (2014), Zheng et al. (2017), Pardos and Kao (2015), Ochoa et al. (2018), Wang et al. (2014), Wong et al. (2018)
Support educational data specifications	3	Pardos and Kao (2015), Di Mitri et al. (2017), Mangaroska et al. (2019)

of the data integration approach in the text. In the case where this information was not explicitly stated, we would sometimes make cautious assumptions in Table 3 (e.g., marked "N/A - most likely manual") based on the data being integrated (e.g., only integrating two lists of names could be achieved easily and more efficiently in Excel than programmatically). In one instance (Villano et al., 2018), the type of data source was not explicitly stated, in which case we had to infer it based on the types of data the system provided.

For the review, we counted one tool or system as one data source (possibly with multiple types of data). Stand-alone digital sources (e.g., corpora, list, questionnaire, and video) were also registered as data sources. Other data sources included equipment with sensors. For types of data sources and data, we iteratively defined broader categories, to allow for comparisons of data sources and data types among the studies reported in the different publications. This meant that for some of the publications, we changed data source and/or type into a broader category. For instance, the sources "Academic Information System" and "Direction of Admissions" were upon further inspection both changed into the broader category "Student Information System," and the type of data "Teamwork questionnaire" was later changed into the broader category "Questionnaire".

Table 4 presents the main observations from the reviewed literature with regard to data sources and data integration. Fourteen of the publications report on studies that combine data that are already available in the same format but come from different data sources. Six of the 20 reviewed publications report on studies that analyze data of different formats that originate from different sources, without a common format. Thus, these data are not integrated but rather analyzed separately. Only three of the studies (Di Mitri et al., 2017; Mangaroska et al., 2019; Pardos & Kao, 2015) in the reviewed publications support educational data specifications. Di Mitri et al. (2017) and Mangaroska et al. (2019) combined data with the same format, while Pardos and Kao (2015) analyzed data from multiple data sources with different formats.

Information about the data sources and types of data that occur most often in the reviewed literature are presented in Figs. 2 and 3. The most used data source is learning management system (LMS), which is used in eight of the studies reported in the reviewed publications. Student information systems and questionnaires are used as data sources in six of the studies; video is used in three; while MOOC, physiological signals wristband, and eye trackers are used in two. The majority of data sources appear only in one of the studies reported in the reviewed publications, as seen in Fig. 2.

The most used type of data is activity log, which is used in ten of the studies reported in the publications. Student background information (e.g., demographics,

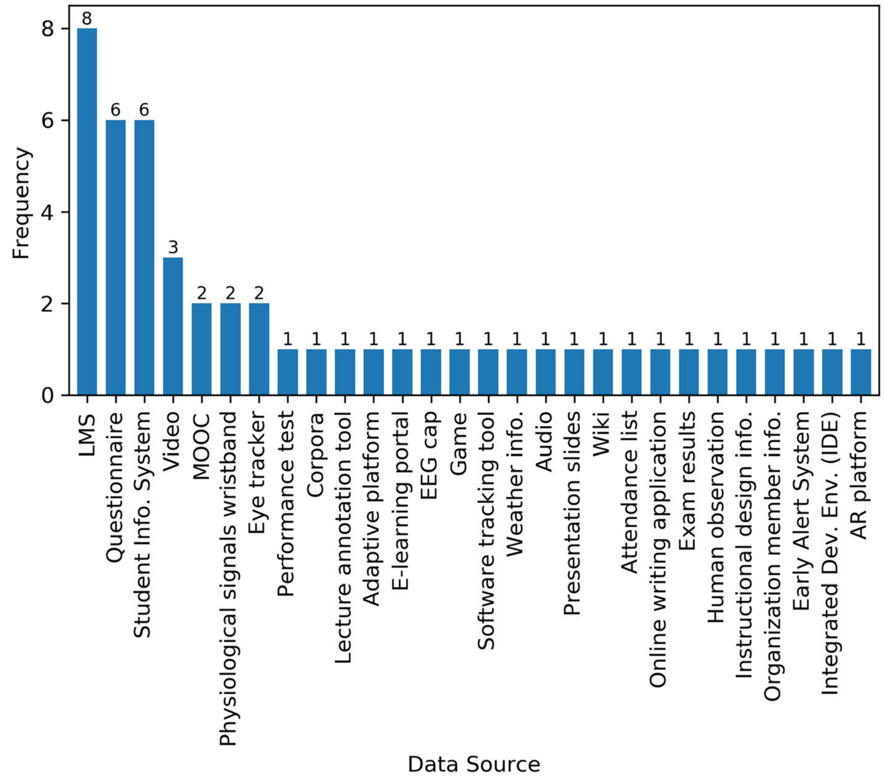

Fig. 2 Data source frequency

socio-economic information, prior academic performance) is used in eight of the studies reported in the publications, questionnaire data is used in seven, performance test data is used in five, while final grades and video derived data are used in three of the studies. Course data, text, physiological arousal data, eye tracker data, and partial course grades are each used in two of the studies reported in the publications. The majority of data types appear only in one of the reviewed publications, as seen in Fig. 3.

Table 5 shows the number of data sources used in the studies reported in each of the publications. As seen, the majority (13 of 20) use only two data sources. The other studies use three, four, five, and seven data sources, respectively.

As seen in Table 3, for data integration, 10 of the reviewed publications report on studies that use automatic integration (using tools). Manual integration requires human effort for the process of combining data, for instance, through copying and pasting data from data sources into excel sheets. Only one publication clarifies that the data are integrated manually (Sun et al., 2018). Five studies do not specify how data integration is achieved. The rest of the reported studies do not integrate data at all. Regarding issues addressed in the reviewed literature, the reported studies address a wide variety of issues. A common theme is identifying struggling students (Gray et al., 2016; Jayaprakash et al., 2014; Lopez Guarin et al., 2015; Sandoval et al., 2018).

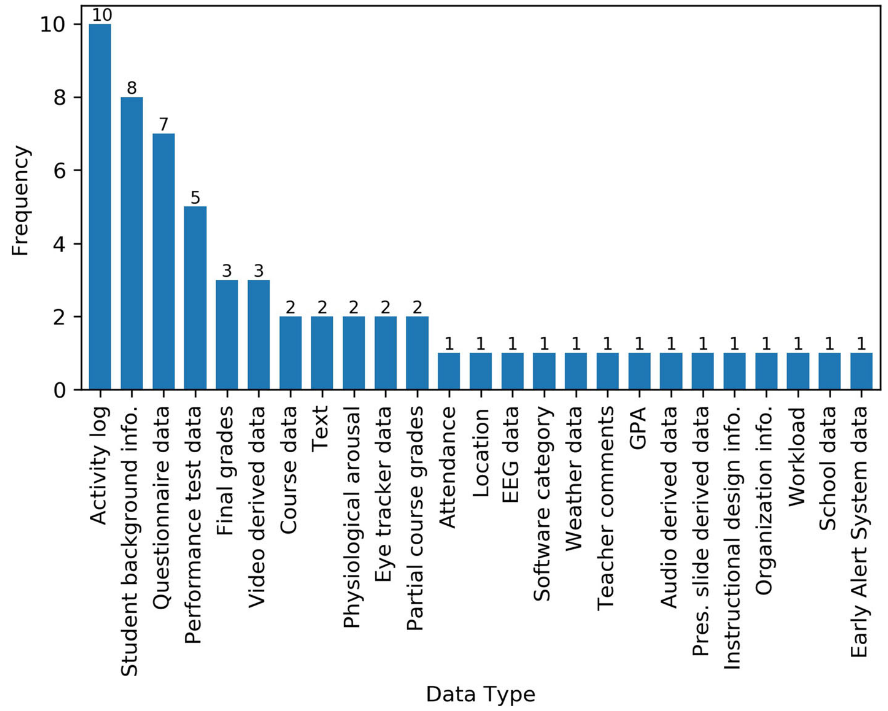

Fig. 3 Data type frequency

As shown in Table 6, different data analysis methods were used for the studies reported in the publications. The most common methods were classification (used in eight publications), regression (used in six publications), visualization (used in five publications), correlation (used in four publications), and descriptive statistics (used in three publications). Two publications (Rodríguez-Triana et al., 2018; Zheng et al., 2017) did not specify methods used in their reported studies. The number of participants in the research studies, whose data were analyzed, vary widely, ranging from 9 to 21,314 (see Table 3). Six of the publications (Pardos & Kao, 2015; Park et al., 2016; Villano et al., 2018; Wang et al., 2014; Wong et al., 2018; Zheng et al., 2017) did not list the number of participants whose data were analyzed in their reported studies.

Table 5 Number of data sources

Number of data sources	References
2	Lopez Guarin et al. (2015), Park et al. (2016), Thompson et al. (2014), Zheng et al. (2017), Pardos and Kao (2015), Raca et al. (2016), Hutt et al. (2017), Jayaprakash et al. (2014), Wang et al. (2014), Villano et al. (2018), Wong et al. (2018), Sandoval et al. (2018), Sun et al. (2018)
3	Ochoa et al. (2018), Gray et al. (2016), Mangaroska et al. (2019)
4	Liu et al. (2017), Di Mitri et al. (2017)
5	Giannakos et al. (2019)
7	Rodríguez-Triana et al. (2018)

Table 6 Methods used

Method used	Publications
Classification methods	Lopez Guarin et al. (2015), Park et al. (2016), Ochoa et al. (2018), Hutt et al. (2017), Jayaprakash et al. (2014), Gray et al. (2016), Sandoval et al. (2018), Giannakos et al. (2019)
Regression methods	Liu et al. (2017), Di Mitri et al. (2017), Mangaroska et al. (2019), Sandoval et al. (2018), Sun et al. (2018), Giannakos et al. (2019)
Visualization	Thompson et al. (2014), Pardos and Kao (2015), Liu et al. (2017), Mangaroska et al. (2019), Wong et al. (2018)
Correlation	Liu et al. (2017), Raca et al. (2016), Gray et al. (2016), Mangaroska et al. (2019)
Descriptive statistics	Raca et al. (2016), Mangaroska et al. (2019), Wong et al. (2018)
ANOVA	Gray et al. (2016), Giannakos et al. (2019)
Text analysis	Thompson et al. (2014), Wong et al. (2018)
t test	Gray et al. (2016)
Network analysis	Pardos and Kao (2015)
Chi Square	Wang et al. (2014)
Structural equation modelling	Sun et al. (2018)
Survival analysis	Villano et al. (2018)

There is also some variation (e.g., relational database, text file, learning record store) in the reported data storage technologies used, as seen in Table 7. However, out of the 20 publications, only seven report the type of technology used in the studies.

Results summary and discussion

Having analyzed the reviewed literature, we can now address the three research questions proposed in this paper. A number of other patterns and trends have emerged in relation to LA in terms of data integration in the context of higher education; these patterns and trends are also discussed in the following.

Re-visiting research questions

What types of data and data sources are being used for LA in higher education?

In the reviewed literature, we see that three data sources appear in many of the research studies, namely LMSs, student information systems, and questionnaires. The use of LMSs and student information systems is most likely because of the wide use of these systems in higher education institutions and the amount of digital information that can be extracted from them. In the case where data are not already collected through systems, questionnaires allow stakeholders to self-report the information. The most common types of data identified are activity logs, student background information, questionnaire data, and

Table 7 Data storage technologies

Data storage technology	Publication
Relational database	Lopez Guarin et al. (2015), Sandoval et al. (2018)
Text file	Thompson et al. (2014)
Comma separated values file	Pardos and Kao (2015)
Learning record store	Di Mitri et al. (2017), Mangaroska et al. (2019)
Data warehouse	Villano et al. (2018)

performance test data. This is perhaps not surprising given that these types of data are often collected in the data sources that have been found to be the most common.

From 2017, there seems to be a shift, where multimodal data are used to a larger extent in the studies (Di Mitri et al., 2017; Giannakos et al., 2019; Ochoa et al., 2018; Rodríguez-Triana et al., 2018). Diverse data are collected from data sources containing sensors, such as physiological signals wristbands, eye trackers, and EEG caps. The use of sensor data in the research, however, is generally still at an early stage, facing challenges such as synchronization in addition to data integration.

How and to what extent are different data being used/combined for LA research in higher education?

The results indicate that the majority of the studies in the reviewed publications that combine data from different sources use tools for automatic integration. The tools used include Business Intelligence software (Jayaprakash et al., 2014), tools developed ad hoc for a specific research project (Di Mitri et al., 2017), SQL (Lopez Guarin et al., 2015), and R scripts (Giannakos et al., 2019). However, out of sixteen studies reported in the publications that integrate data, five of the studies do not specify the integration approach, while one specifies that the approach is manual. Most publications report on studies that combine data of similar formats. A number of the studies that use multiple data sources do not integrate data but rather analyze them separately.

The results also show that the number of data sources used tends to be limited. The majority of studies reported in the publications use two data sources. Only seven out of 20 used more than two data sources. As mentioned earlier, we also see that a small number of data sources appear in many of the research studies. However, higher education often includes a broad variety of tools, such as digital exam systems, library systems, and key card access systems. When only a limited number of data sources are taken into account for LA, this may give only a partial picture of student behavior and learning, thereby also potentially biasing analysis results. On the other hand, when increasing the number of data sources, the complexity of the problem of interoperability is also increased. Challenges such as meaningful data integration, storage, and processing requirements need to be addressed.

What methods are used, and what issues are being addressed through using/combining multiple data sources?

As the results have shown, the most common methods used are classification methods. This is in accordance with findings in previous research (Shahiri et al., 2015). Regression methods, visualization, correlation, and descriptive statistics are also used to analyze data in a number of the publications. While the issues addressed in the reviewed literature vary greatly, a common theme is identifying struggling students. The publications addressing this theme of issues have been published between 2014 and 2018, indicating that addressing this type of theme has been of interests for a few years. It is clear that identifying and helping students who are at risk can have positive consequences, not only for the individual students, but also for the higher education institutions (e.g., in terms of economy via student throughput and reputation).

Additional findings

Lack of details in technical solutions

The reviewed literature tends to lack details about the technical solutions for combining multiple data sources in the implemented systems. There is generally little information about data formats, data storage solutions, and data integration techniques. Only four publications report on studies that integrate data sources and provide information (implicitly or explicitly) on all three of formats, storage solutions, and integration techniques. Lopez Guarin et al. (2015) detail a data integration technique of joining database tables, which implies the use of a relational database for storage, with the relational model as the underlying format. Sandoval et al. (2018) also integrate data into a relational database. Di Mitri et al. (2017) and Mangaroska et al. (2019) format data using the xAPI specification. The collected data are stored and integrated in a Learning Record Store. For the five publications that do not specify the data integration technique used in the studies, we may speculate that they are combining the data manually. However, without any specific information, it is hard to draw any hard conclusions.

Lack of involvement of stakeholders

Most of the reviewed publications have no information regarding who chose, combined, and managed the data used in the studies. This indicates a lack of involvement of stakeholders and adoption of participatory or co-design approaches. Successfully involving and engaging different stakeholders may be a challenge at the institutional level, both in terms of management and organization (Buckingham Shum & McKay, 2018; Dawson et al., 2018). The reviewed publications tend to make clear that it was the researchers themselves who did the feature selection and analysis, while further data management is not addressed. One exception is found in the publication by Rodríguez-Triana et al. (2018). Here, a teacher collaborated with the researchers to customize a multimodal LA solution for blended learning, with emphasis on the data-gathering phase (e.g., what questions are to be answered by the solution and what data are to be used given constraints and affordances).

Lack of use of educational data specifications

Using educational data specifications such as xAPI and IMS Caliper Analytics enables a more uniform representation of activity data, thus supporting the combination of data that follow the same standards. In the reviewed publications, however, we only find evidence of three studies using the xAPI specification (Di Mitri et al., 2017; Mangaroska et al., 2019; Pardos & Kao, 2015), while none use Caliper Analytics. Even though xAPI and Caliper are mainly focused on the technical level of interoperability (with added capabilities for the semantic level), there are gaps in the specifications that can make it challenging to describe parts of learning and learning environments, and inconsistent use of a specification among different communities of practice may make it difficult to efficiently integrate data, these specifications are still a good starting point for data integration in LA. It is not clear why xAPI and Caliper Analytics seem to be used so seldom in LA research. One possible answer may be that those implementing LA solutions lack skills in using such specifications. With regards to some individual data sources, e.g., digital tools such as LMSs, it may also be a factor that EdTech/digital tool vendors do not tend to provide data in xAPI or Caliper Analytics formats, in which case, those using educational data specifications need to transform data from the original format to the relevant standardized format themselves. It is clear that this transformation effort can be time-consuming, even for LA solution implementers that are skilled using the educational data specifications (for some digital tools, there are plugins that may help with the transformation process [JISC, 2019]). In addition, it may also be a factor that some of the usage of xAPI and Caliper has not been illustrated through this review, as there are few publications included on platforms. Further examination is needed in order to understand why educational data specifications do not seem to be widely used for data integration in LA.

Combination of similar data formats as low-hanging fruit

In the reviewed literature, we see that most publications report on studies that combine data of similar formats, as mentioned earlier. For some of these studies, the data did not originally have a common format; for instance, Raca et al. (2016) get data from the diverse sources of video and questionnaires. Other studies combine data that are already in similar formats. For instance, Di Mitri et al. (2017) and Rodríguez-Triana et al. (2018) both get data from application programming interfaces (APIs), which tend to output JSON format. The transformation effort is trivial when data are originally in the same format, but more challenging when the formats are not originally equivalent. Thus, focusing on data already in the same format is a low-hanging fruit in terms of data integration. However, there is still the significant challenge of semantics, ensuring the data have the same meaning. In general, the alignment of concepts is usually done ad hoc when programming an analytics solution. Another approach to the alignment of concepts would be to use an ontology, "a formal, explicit specification of a shared conceptualization" (Studer, Benjamins, & Fensel, 1998, p. 184). Ontologies are related to controlled vocabularies, although in general, they tend to have more complex and formal collections of concepts and relationships (W3C, 2015). When using an ontology, it is possible to add descriptions and meaning to data coming from various sources, and to combine, support, and reuse different specifications (Allemang & Hendler, 2011). Ontologies also enable inference, meaning we can state new and related facts from one stated fact.

Recommendations and conclusion

This review has examined LA research in higher education that uses and/or combines data from multiple sources, aiming to assess the current state of LA in terms of data integration in the context of higher education.

We acknowledge that there are some limitations in this systematic literature review. Limiting the language to English may have excluded relevant research published in other languages. Choosing to focus on LA in the context of higher education has excluded some research that use multiple data sources which take place in other levels of education, such as K-12 (Chang et al., 2017; Mutahi et al., 2015). While the research studies were excluded for not taking place in higher education, both examine how students engage with interactive learning content, whether individually or in groups. The information provided and research focus varies for the publications included in the review. Some publications lack information that we consider important for the review (lacking information is marked N/A in Table 3). While we are aware that there are LA platforms such as OnTask (2019), JISC (2019), and Apereo (2018) that address data integration issues, the review process only resulted in two publications with research where it is clear that they were using platforms (Mangaroska et al., 2019; Pardos & Kao, 2015), leading us to believe that the published research available on such solutions is limited. It is possible, however, that some of the research studies were using platforms without stating it clearly, since many of the publications did not provide much detail about the technical solutions.

While the choice of data sources used in research depends of the purpose of the learning analytics (e.g., identifying dropouts or successful students, giving feedback), we reason that the actual choice of sources might also follow from access to data sources not always being that easy (this is the case from our own experience as well). In addition, integrating data from multiple sources is a challenge. Thus, based on the finding that few studies actually use and combine multiple data sources, we identify a need for more research studies that use and combine more data sources and report the details so others can learn from the experience (including how they managed to access and acquire the data).

Integrating multiple data sources plays an important role in scaling up LA. The problem of scalability in terms of LA in higher education can be viewed as a natural extension of the problem of data integration. In addition to the problem of combining data from multiple sources, it encompasses additional dimensions such as the institutional (the solution should at minimum scale across departments and faculties within a higher education institution) and the temporal (the solution should scale across semesters and years, to allow for a more complete picture of learning). In higher education, there are multiple tools that generate data, and being able to integrate these data can promote the usefulness of LA to a different level than one data source can do by itself. If the studies conducted are to result in realistic analysis results, it is important that they reflect learning and learning environments as realistically as possible; thus, it may be beneficial to combine more than one or a limited set of data sources.

With regard to the problem that the reported studies use limited data sources, we argue that institutional policy should guide data use. To break down data silos and enable more efficient application of diverse data sources, such policy should cover not only the technical levels of data exchange, but also factors related to organization and management.

Privacy, which is guided by national and international laws, is essential for LA in general and for data integration in particular. Securing user data includes prevention of unauthorized access to the systems. In addition, de-identification of user information yields added privacy. When aggregating de-identified data from multiple sources, there is the added chance that users can become indirectly identifiable from the collective information that is stored. On the other hand, there is the risk that ensuring data privacy may limit access to data that could be important for LA research (Flanagan & Ogata, 2017). Such considerations need to be balanced, complying with relevant laws.

Having conducted the review, we found that in general researchers do not describe their process of data integration in enough detail (or at all). Our recommendation is that they make this important part of LA research clearer, as building on existing knowledge can help push the boundaries of the state of the art and help in replication of earlier studies.

As mentioned in the discussion, it is not clear why xAPI and Caliper Analytics seem to be so seldom used, even though these specifications help enable interoperability on the technical level and have capabilities for the semantic level. Here, we would suggest that studies look further into the specifications and their usage, in order to understand the challenges in adoption and the opportunities afforded, from the LA stakeholders' point of view.

In terms of stakeholder participation, we would recommend more inclusion of different stakeholders in LA research. To really understand the pressing issues, such as how to interpret analysis results in a given context, it is not enough to include only researchers and technical staff. Those that have more thorough knowledge of the problem domain (e.g., teachers for classroom studies or instructors for university course studies) also need to be included. There also needs to be more focus on participatory design or co-design, where stakeholders with diverse competences and strengths use their individual knowledge for implementation and scalability of LA in higher education. Finally, there needs to be a dialog between the LA research community and the EdTech/digital tool vendor sector about the need to have standardized data.

Abbreviations

API: Application Programming Interface; EDM: Educational Data Mining; LA: Learning Analytics; LMS: Learning Management System; MOOC: Massive Open Online Course

Acknowledgements

There are no acknowledgements.

Authors' contributions

The systematic review presented in this article is a part of the PhD project conducted by JS. BW is her main supervisor, and WC is her co-supervisor. WC has been working closely with JS on judging the relevance of each identified publication for the review, and assisted in planning the review and article writing. BW has contributed to judging the relevance of problematic publications, planning the review, and article writing. All authors read and approved the final manuscript.

Funding

This research is part of JS's PhD funded by the Centre for the Science of Learning & Technology (SLATE), University of Bergen, Norway.

Availability of data and materials

Data for the review was extracted from publications identified through searching academic databases and a limited number of conference proceedings/journals for publishing LA research.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Centre for the Science of Learning & Technology, University of Bergen, Bergen, Norway. [2]Department of Information Science & Media Studies, University of Bergen, Bergen, Norway. [3]Oslo Metropolitan University, Oslo, Norway.

Received: 25 January 2019 Accepted: 26 July 2019
Published online: 28 August 2019

References

Allemang, D., & Hendler, J. (2011). *Semantic web for the working ontologist: effective modeling in RDFS and OWL* (2nd ed.). Morgan Kaufmann.

Apereo. (2018). Learning Analytics Initiative | Apereo. Retrieved from https://www.apereo.org/communities/learning-analytics-initiative

Buckingham Shum, S., & McKay, T. (2018). Architecting for learning analytics: Innovating for sustainable impact. *EDUCAUSE Review*.

Chang, C.-J., Chang, M.-H., Liu, C.-C., Chiu, B.-C., Fan Chiang, S.-H., Wen, C.-T., et al. (2017). An analysis of collaborative problem-solving activities mediated by individual-based and collaborative computer simulations: Collaborative problem solving. *Journal of Computer Assisted Learning, 33*(6), 649–662 https://doi.org/10.1111/jcal.12208.

Chatti, M. A., Muslim, A., & Schroeder, U. (2017). Toward an open learning analytics ecosystem. In B. K. Daniel (Ed.), *Big Data and learning analytics in higher education* (pp. 195–219) https://doi.org/10.1007/978-3-319-06520-5_12.

Cooper, A., & Hoel, T. (2015). *Data sharing requirements and roadmap*.

Dawson, S., Poquet, O., Colvin, C., Rogers, T., Pardo, A., & Gasevic, D. (2018). *Rethinking learning analytics adoption through complexity leadership theory* (pp. 236–244). ACM Press https://doi.org/10.1145/3170358.3170375.

Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., & Specht, M. (2017). *Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data* (pp. 188–197). ACM Press https://doi.org/10.1145/3027385.3027447.

European Commission. (2017). New european interoperability framework. Retrieved from https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf

European Union. (2016). Regulations. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Flanagan, B., & Ogata, H. (2017). Integration of learning analytics research and production systems while protecting privacy. In *The 25th International Conference on Computers in Education, Christchurch, New Zealand* (pp. 333–338).

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management, 48*, 108–119 https://doi.org/10.1016/j.ijinfomgt.2019.02.003.

Gray, G., McGuinness, C., Owende, P., & Hofmann, M. (2016). Learning factor models of students at risk of failing in the early stage of tertiary education. *Journal of Learning Analytics, 3*(2), 330–372 https://doi.org/10.18608/jla.2016.32.20.

Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., & D'Mello, S. K. (2017). Gaze-based detection of mind wandering during lecture viewing. In *Proceedings of the 10th International Conference on Educational Data Mining* (pp. 226–231).

IMS Caliper Analytics. (2019). Caliper Analytics | IMS Global Learning Consortium. Retrieved from https://www.imsglobal.org/activity/caliper

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: an open source analytics initiative. *Journal of Learning Analytics, 1*(1), 6–47 https://doi.org/10.18608/jla.2014.11.3.

JISC. (2019). Learning records warehouse: technical overview: Integration overview. Retrieved from https://docs.analytics.alpha.jisc.ac.uk/docs/learning-records-warehouse/Technical-Overview:%2D%2DIntegration-Overview

Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of data integration and interoperability in big data. *2014 IEEE International Conference on Big Data (Big Data)*, 38–40 https://doi.org/10.1109/BigData.2014.7004486.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University (UK).

Liu, M., Kang, J., Zou, W., Lee, H., Pan, Z., & Corliss, S. (2017). Using data to understand how to better design adaptive learning. *Technology, Knowledge and Learning, 22*(3), 271–298 https://doi.org/10.1007/s10758-017-9326-z.

Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje, 10*(3), 119–125 https://doi.org/10.1109/RITA.2015.2452632.

Mangaroska, K., Vesin, B., & Giannakos, M. (2019). Cross-platform analytics: A step towards personalization and adaptation in education. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19* (pp. 71–75) https://doi.org/10.1145/3303772.3303825.

Misiejuk, K., & Wasson, B. (2017). *State of the field report on learning analytics*. Bergen: *SLATE Report 2017-2* Retrieved from http://bora.uib.no/bitstream/handle/1956/17740/SoF%20Learning%20Analytics%20Report.pdf.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA group. (2009). *Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement* (p. 7).

Mutahi, J., Bent, O., Kinai, A., Weldemariam, K., Sengupta, B., & Contractor, D. (2015). Seamless blended learning using the cognitive learning companion: A systemic view. *IBM Journal of Research and Development, 59*(6), 8:1–8:13 https://doi.org/10.1147/JRD.2015.2463591.

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). *The RAP system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors* (pp. 360–364). ACM Press https://doi.org/10.1145/3170358.3170406.

OnTask. (2019). Home | OnTask. Retrieved from https://www.ontasklearning.org/

Pardos, Z. A., & Kao, K. (2015). *moocRP: An open-source analytics platform* (pp. 103–110). ACM Press https://doi.org/10.1145/2724660.2724683.

Park, Y., Yu, J. H., & Jo, I.-H. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. *The Internet and Higher Education, 29*, 1–11 https://doi.org/10.1016/j.iheduc.2015.11.001.

Raca, M., Tormey, R., & Dillenbourg, P. (2016). Sleepers' lag: Study on motion and attention. *Journal of Learning Analytics, 3*(2), 239–260 https://doi.org/10.18608/jla.2016.32.12.

Rodríguez-Triana, M. J., Prieto, L. P., Martínez-Monés, A., Asensio-Pérez, J. I., & Dimitriadis, Y. (2018). *The teacher in the loop: Customizing multimodal learning analytics for blended learning* (pp. 417–426). ACM Press https://doi.org/10.1145/3170358.3170364.

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. *The Internet and Higher Education, 37*, 76–89 https://doi.org/10.1016/j.iheduc.2018.02.002.

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education: A review of UK and international practice. Jisc. Retrieved from https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v3.pdf

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414–422 https://doi.org/10.1016/j.procs.2015.12.157.

Siemens, G. (2011). 1st international conference on learning analytics and knowledge. Technology Enhanced Knowledge Research Institute (TEKRI). Retrieved from https://tekri.athabascau.ca/analytics/

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. *Data and knowledge engineering, 25*(1), 161–198.

Sun, Z., Xie, K., & Anderman, L. H. (2018). The role of self-regulated learning in students' success in flipped undergraduate math courses. *The Internet and Higher Education, 36*, 41–53 https://doi.org/10.1016/j.iheduc.2017.09.003.

Thompson, K., Kennedy-Clark, S., Wheeler, P., & Kelly, N. (2014). Discovering indicators of successful collaboration using tense: Automated extraction of patterns in discourse: Discovering indicators of successful collaboration. *British Journal of Educational Technology, 45*(3), 461–470 https://doi.org/10.1111/bjet.12151.

Villano, R., Harrison, S., Lynch, G., & Chen, G. (2018). Linking early alert systems and student retention: a survival analysis approach. *Higher Education, 76*(5), 903–920 https://doi.org/10.1007/s10734-018-0249-y.

W3C. (2015). Ontologies. Retrieved from https://www.w3.org/standards/semanticweb/ontology

Wang, Y., Paquette, L., & Baker, R. (2014). A Longitudinal Study on Learner Career Advancement in MOOCs. *Journal of Learning Analytics, 1*(3), 203–206 https://doi.org/10.18608/jla.2014.13.23.

Wong, E. Y. W., Kwong, T., & Pegrum, M. (2018). Learning on mobile augmented reality trails of integrity and ethics. *Research and Practice in Technology Enhanced Learning, 13*(1) https://doi.org/10.1186/s41039-018-0088-6.

xAPI (2019). xAPI.com Homepage: What is xAPI (the Experience API). Retrieved from https://xapi.com/

Zheng, M., Bender, D., & Nadershahi, N. (2017). Faculty professional development in emergent pedagogies for instructional innovation in dental education. *European Journal of Dental Education, 21*(2), 67–78 https://doi.org/10.1111/eje.12180.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.