Analyzing the educational goals, problems and techniques used in educational big data research from 2010 to 2018

Benazir Quadira,
Department of Information Management School of Business, Shandong University of Technology, Zibo, People's Republic of China;

Nian-Shing Chen
Department of Applied Foreign Languages, National Yunlin University of Science and Technology, Douliu, Taiwan;

Pedro Isaias
Institute for Teaching and Learning Innovation, The University of Queensland, Queensland, Australia

Abstract: The purpose of this study is to review journal papers on educational big data research published from 2010 to 2018. A total of 143 papers were selected. The papers were characterized based on three dimensions: (a) educational goals; (b) educational problems addressed; and (c) big data analytical techniques used. A qualitative content analysis approach was conducted to develop a coding scheme for analyzing the selected papers. The results identified four types of educational goals, with a clear predominance of quality assurance. The identification of the most mentioned educational problems resulted in four main concerns: the lack of detecting student behavior modeling and waste of resources; inappropriate curricula and teaching strategies; oversights of quality assurance; and privacy and ethical issues. With the exception of ethical and privacy concerns, which were solely mentioned by a few publications, all other problems had a similar importance in the reviewed papers. Concerning the most mentioned big data analytical techniques, the coding scheme revealed that the majority of the papers focused on the educational data mining technique followed by the learning analytics technique. The visual analytics technique was mentioned only in a few papers. The results also indicated that the educational data mining technique is the most suitable technique to use for quality assurance and to provide potential solutions for the lack of detecting student behavior modeling and the waste of resources in institutions.

Keywords: Educational goals, educational problems, educational big data, educational data mining, learning analytics meta-analysis.

1. **Introduction**

The notion of big data was first introduced around 1970 with the idea of a "database machine" for storing and analyzing data (Chen et al., 2014). Since then, an increasingly large body of database systems has been introduced around the globe enhancing data volume, including storage and processing capacity. The development of a parallel database system for increasing data volume in 1980 is an example of this growth (DeWitt & Gray, 1992). Since then, big data has become capable of sorting large quantities of data longitudinally and for specific transactions (Picciano, 2012). Most research has involved the collection and organization of the 5Vs (i.e. high volume, velocity, variety, veracity, and value) of the current big data production, which represents new opportunities and challenges in both scale and complexity (Margolis et al., 2014). Moreover, big data has become a useful approach for multidisciplinary problem solving (Brodie, Greaves, & Hendler, 2011).

Thus, there have been numerous studies which have attempted to develop several applications by deploying big data in different sectors, including digital media research (Mahrt & Scharkow, 2013) and education (Vaitsis, Nilsson, & Zary, 2014). While a variety of big data applications exists, they must be used effectively in the context of new knowledge (Margolis et al., 2014). Moreover, big data application in education is an emerging and important research field (Vaitsis et al., 2014). In academia, the advent of big data occurred in 2008 with the publication of a big data special issue in Nature (Chen et al., 2014). In 2012, another special issue was published regarding big data in the European Research Consortium

for Informatics and Mathematics (ERCIM) news (Chen et al., 2014). Since then, there has been increasing interest in big data in the field of education (Vaitsis et al., 2014).

Recently, educational institutions have been trying to apply big data in different departments such as engineering (Xian & Madhavan, 2014), business studies (Gupta, Goul &Dinter, 2015), public affairs (Mergel, 2016), geospatial research (Wang, Liu & Padmanabhan, 2016), statistics (Ridgway, 2016), chemistry (Tetko, Engkvist, Koch, Reymond & Chen), data science (Song&Zhu, 2016). There are many studies also attempting to understand what motivates stakeholders to apply big data in education. For example, Vaitsis et al. (2014) found that the medical curriculum plays an important role in improving medical education as it is used by teachers and directors to plan, design, and deliver teaching and assessment activities and student evaluations. Therefore, they implemented visual analytics, which revealed some unique ways of representing big data for improving medical education systems. In another study, Ping (2013) found a method to use educational big data to make scientific educational decisions and to optimize instruction by educators and learners using learning analytics to analyze educational big data (Siemens & Long, 2011) and to, consequently, uncover the value of the learning process.

Some systematic reviews of journal papers on big data can provide researchers with a clear depiction of the recent trends in this research field. Chen et al. (2014) examined the research in terms of three categories and their corresponding criteria, such as background and the state-of-the-art situation of big data, four phases of the value chain of big data, and finally several representative applications of big data. In another study, Chen and Zhang (2014) offered a deeper view of big data including big data application

opportunities and challenges, and state-of-the-art techniques and technologies for dealing with big data problems. They also discussed several types of methodologies. Moreover, Fan and Bifet (2014) conducted a review study with four papers to present a broad view of the topic, its status, controversies, and forecasts for the future. Such a review provided an initial context in which to analyze the literature of big data from a general perspective. Nonetheless, an up-to-date and targeted investigation of research papers in the field of education with big data is still lacking.

In recent years, a growing number of studies have focused on big data in education. Despite this rising interest, there is a lack of sound analysis and comparison within the topics of educational goals, educational problems and the techniques applied in big data from the year 2010 to 2018. Moreover, the identification of a suitable method to achieve specific educational goals remains unclear. A scrutiny of the aforementioned issues on educational big data research can provide explanatory information to understand this field's status. Therefore, this study aims to identify the educational goals, educational problems addressed, and the techniques used for handling educational big data, by analyzing relevant papers published in journals from 2010 to 2018. This study addresses the following three research questions.

1. What are the main goals for using educational big data?
2. What problems does educational big data address more often?
3. What big data techniques are used more often in educational big data?

2. **Literature review**

2.1 *Big data in the achievement of educational goals*

John Mashey seems to have been the first to use the term “Big Data” in 1998 in a Silicon Graphics slide deck with the title, “Big Data and the Next Wave of InfraStress’’ (Diebold, 2012). Since then, there have been various different explanations of the term “Big data” from volume, variety, velocity (3V) (Doug Laney, 2001) to volume, variety, velocity, veracity (4V) (Gantz & Reinsel, 2011) and 4V to 5V (Fan & Bifet, 2014). These 5Vs are volume, variety, velocity, veracity and value. Various definitions of big data are found in different fields. Mayer-Schönberger and Cukier (2013) defined the term big data as “by collecting, aggregating and analyzing very large amounts of data there are things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value’’. From the perspective of academia, according to Daniel (2015), “big data analytics could be applied to examine student entry on a course assessment, discussion board entries, blog entries or wiki activity, which could generate thousands of transactions per student per course”.

Several attempts have been made to exploit big data’s benefits in education (Xie et al. 2014; MacNeill, Campbell, & Hawksey, 2014; Vaitsis et al., 2014; Williamson, 2015). Xie et al. (2014) proposed a novel computational approach based on time series analysis to assess engineering design processes using a computer-aided design (CAD) tool for improving learning, while MacNeill et al. (2014) focused on the development of education content analytics and considered the legal and ethical aspects

of collecting and analyzing educational data for future development of massive open online course (MOOCs). Williamson (2015) examined the emergence of digital governance in public education, and found that digital governance facilitated by network-based communication and database-driven information processing software. Which is being broadly promoted in education by organizations that are seeking typical educational decision-making to socio-algorithmic forms of power. The aforementioned features have the capability to predict, govern and activate learners' capacities and subjectivities. Thus, one of the aims of this study is to identify the educational goals of deploying big data applications in the published journal articles from 2010 to 2018.

2.2 *Educational problems addressed by big data*

There are numerous educational problems in an institution for learners, teachers, and administrators to deal with. In relation to the learners, problems include chaotic learner behavior such as a lack of responses in discussion forums (Gomez-Aguilar et al., 2015), a lack of engagement (Xie et al., 2014), dropouts (Eynon, 2013), absence from class (Picciano, 2012), missing quizzes (Picciano, 2012), failure to meet course standards, not promptly taking courses, overlooking reading materials (Picciano, 2012) as well as early warnings received from teachers in the study of educational big data perspectives.

With concern to teachers, problems include inappropriate curricula and teaching strategies (Vaitsis et al., 2014), negative interventions in students' learning (Dyckhoff et al., 2012), as well as a lack of quality assurance (Greller & Drachsler, 2012) and feedback service (Tempelaar, Rienties, & Giesbers,

2015). The problems related to administration include lack of funding or financial planning (Selwyn, 2014), such as a reduction of government funding, support from business and the private sector, and growing regulatory demands for accountability and transparency (Hazelkorn, 2007). Nowadays, administrators are more concerned with codes of conduct such as privacy and ethical issues (MacNeil et al., 2014). Moreover, administrators are also facing a lack of online learning activities (Romero-Zaldivar, Parbo, Burgos, & Kloos, 2012), record-based methodological triangulation (Gorissen, Bruggen & Jochems, 2013), and identification of meaningful pedagogical variables (Chiang, Goes, & Stohr, 2012).

Such problems might be very common in an institutional organization and might concern the teachers, learners and administrators. However, these problems can be easily addressed by big data (Brodie et al., 2011). Brodie et al. (2011) mentioned that big data can identify such problems and it can become a useful case for multidisciplinary problem solving. Although big data can handle many problems, Picciano (2012) argued that the use of big data for instructional applications is still in its infancy as it may not be able to solve all of the problems faced by educational institutions. Therefore, another aim of this study is to identify the educational problems in the field of education whereas big data can cope with those educational problems.

2.3 *Different techniques used in educational big data applications*

Numerous studies have attempted to apply big data analytics using different techniques such as statistics, data mining, machine learning, neural networks, social network analysis, signal processing,

pattern recognition, optimization methods and visualization (Chen & Zhang, 2014). In addition, Chen et al. (2014) examined the most important techniques used in big data including structure data analysis, text analysis, web site analysis, multimedia analysis, network analysis and mobile analysis. At present, the main processing methods of big data include bloom filters, hashing, indices and parallel computing (Chen et al., 2014). Although these big data techniques have been applied in different fields including education, there is a concern about how to rapidly extract essential information from massive data to bring value for enterprises, institutions and individuals. As Tulasi (2013) suggested, big data could lead to a new wave of technological advances, which could help improve academic effectiveness.

Moreover, researchers from the educational community have begun to realize the potential application of big data techniques for enhancing education systems and outcomes (Manyika et al., 2011). Siemens and Long (2011) identified two areas, educational data mining (EDM) and learning analytics (LA), which cover the inclusion and exploration of big data capabilities in education. Where EDM is concerned with "developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist" (Romero & Ventura, 2013, p.12). Thus, Berland et al. (2014) described EDM as a powerful method to present actionable data to learners and teachers. For example, due to the real-time assessment of student/teacher processes and progress, formative feedback is radically increased, which can enhance learning/teaching effectiveness with their own data (Berland et al., 2014).

Another emerging field is LA, by which sophisticated analytics tools can improve learning and education (Elias, 2011). For example, LA can help institutions to make decisions for educators, learners and administrators (Pardo & Siemens, 2014). Thus, big data encompasses the promising research field of learning analytics (Siemens & Long, 2011). Daniel and Buston (2013) mentioned that big data in higher education consists of four components: institutional analytics (i.e., assessment policy analytics, instructional analytics, etc.); information technology analytics (i.e., student information, learning management, alumni systems, etc.); academic analytics (i.e., strategic decision-making processes); and learning analytics (i.e., measurement, collection, and analysis and reporting of data about learners and their learning contexts). Daniel (2015) mentioned that the key contribution of big data depends on the application of three models, that is, the descriptive, relational and predictive models. He explained descriptive analytics to identify patterns on recent trends such as student enrollment, graduation rates, etc. Predictive analytics may offer better decisions and actionable insights based on data, and finally, prescriptive analytics helps higher education institutions assess the current situation and make consistent predictions.

While various techniques for big data analytics have been adopted in different fields, the present study aims to identify the most commonly used among those techniques (i.e., LA, EDM, VA) in education to achieve educational goals.

3. **Methods**

3.1 Paper selection

The current study searched Web of Science, Scopus, Google Scholar and the Social Science Citation Index (SSCI) database for papers related to educational big data published between the years of 2010 and 2018. The searching keyword consisted only of "educational big data". The original search returned 235 papers, which were reviewed in terms of the abstract, method and full text to confirm whether they did in fact concern big data in education. The studies that did not pertain to education were excluded. Moreover, conference papers, books, unpublished papers, review papers, position papers, complementary notes, editorial papers and dissertations were equally excluded. As a result, 143 studies remained for further review as shown in Figure 1.

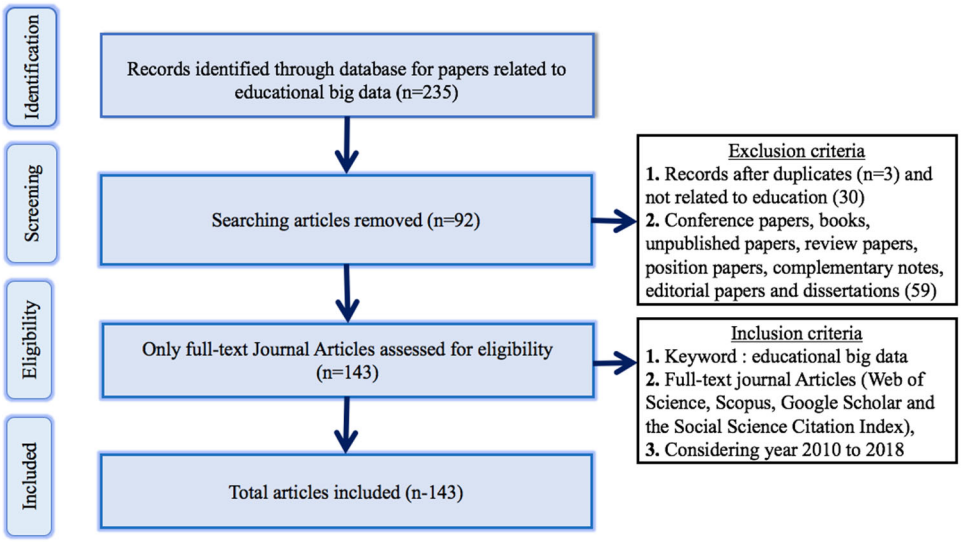

Figure1. Flowchart of the paper selection procedure

3.2 Development of the coding scheme and definitions

The selected papers were analyzed using a qualitative content analysis. The key information corresponding to the research questions were first identified from each study and then synthesized. The synthesis procedure had two main phases. Firstly, the studies were examined in terms of research purposes, research gap, and the techniques, which were applied in the field of educational big data research. Secondly, an inductive method was used to develop the coding categories, and identify patterns in different studies including their educational goals, educational problems, and different big data techniques across the different studies in the field of education. The two authors who contributed to this analysis developed consensus through face-to-face discussion for the items with different coding results. The authors held several meetings to explain and evaluate the relevance of the criteria for selecting the articles and the categorizing coding scheme. Topics relating to the categorization of the articles were clarified through discussions between the authors. Given all decisions were made by consensus during the course of the process, a formal reliability check, such as Krippendorff's alpha (Krippendorff, 2013) did not seem to be viable in this case.

In line with its goals, this review developed a coding scheme as aforementioned for educational goals, educational problems addressed, and the techniques of big data applied in the selected educational big data research. As mentioned in the literature review section, big data in education not only identifies institutional problems, but also has an impact on the development of the learning process, problem-solving approaches, etc. These research subjects for each aspect of institutional demand were then

included as subtopics discussed in the study. Thus, a coding scheme for analyzing the selected paper was developed as shown respectively in Tables 1, 2 and 3.

Sub factors were clustered by their similarity to each other. Some sub factors from different studies, such as to plan future courses, student course scheduling, planning resources allocation, admission and counseling process (Romero & Ventura, 2013) and developing curricula (Vaitsis et. al., 2014) were clustered as factors delineating as appropriate planning and scheduling as one of coding scheme. Numerous studies have been conducted to discover students' learning pattern and guide courses improvement (Song, Zhang, Duan, Hossain, & Rahman, 2018), and enhancing the development of students' learning ability (Shen, 2018). In addition, educational (learner) dashboard systematically delivers timely and continuous feedback on performance in support of improved learning outcomes (Boscardin, Fergus, Hellevig, & Hauer, 2018). Thus, determining students' performance is one of the coding scheme we rectified. Quality assurance is another coding scheme of the study, which need to be make sure while implementing learning systems. Therefore, how the new methods, technologies, and tools of big data can enhance better quality future of online learning environment (Dahdouh, Dakkak, Oughdir, & Messaoudi, 2018) need to be concerned. There are some other studies developing different organizational models/frameworks. For example, a teaching outcome model (TOM), that can be used to inspire and inspect quality of teaching (Ndukwe, Daniel, & Butson, 2018), Teaching Excellence Framework (TEF) which can be viewed as a multi-purpose evaluation tool such as teaching excellence, quality assurance, a measure to provide market information to consumers and allocate fee increases to

institutions (Gunn, 2018). In addition, Chaos optimization cognitive learning model (COCLM) that takes into account the learners' learning motivation, learning task demands, and the change rate of cognitive rules, and transforms the learning process of distance learning into a multi-objective optimization problem (Wen, Zhang, & Shu, 2018). Therefore, such models/frameworks improving educational institutions presenting as the organizational model/framework as one of coding scheme in this study. There were four coding schemes developed for aiming educational goals in an institution as shown in Table 1.

Table 1. Coding scheme for educational goals aimed and their definitions

Coding scheme for educational goals	**Definition of coding scheme of educational goal**
Appropriate planning and scheduling	To plan future courses, student course scheduling, planning resources allocation, admission and counseling process and developing curricula (Romero & Ventura, 2013).
Determine students' performance	To estimate unknown values of student performance, knowledge, activities, test scores and marks (Papamitsiou & Economides, 2014).
Verify quality assurance	The objective is to support the (self-) assessment of improved efficiency and effectiveness of the learning process (Chatti et al., 2012).
The organizational model/framework	To infer parameters of probabilistic models from given data to predict the probability of events of interest (Papamitsiou & Economides, 2014).

One of the coding schemes for educational problem is lack of detecting student behavior modeling and

waste of resources, which consists of different studies with many sub factors such as lack of responses in discussion forums, dropouts and absence from class and quizzes, isolated and warned students. Another important coding scheme is inappropriate curricula and teaching strategy, which can create challenges in the face of rapid practice change. Due to curriculum development nurse educators are encouraged to consider the key areas such as the role of electronic health records (EHRs), wearable technologies, big data and data analytics, and increased patient engagement (Risling, 2017). Santoso, (2017) explored Hadoop as big data analytic tools for data ingestion/staging. They concluded by outlining future directions relating to the development and implementation of an institutional project on Big Data. Thus, oversights on quality assurance cannot be ignored due to improvement and implementation of the institutional projects. Researchers are trying to apply privacy and ethical issues, which seems to be consider another very important coding scheme in the context of educational big data research. In addition, due to lack of security and other concerns relating to campus infrastructure and operations it cannot be used to create customized automated dashboards to support critical decisions (Chaurasia & Frieda Rosin, 2017). There were four schemes developed for identifying educational problems in an institution faced by administrators/educators as shown in Table 2.

Table 2. Coding scheme for educational problems addressed and their definitions

Coding scheme for educational problems	**Definition of coding scheme of educational problems addressed**
Lack of detecting student behavior	To identify students' prerequisite matters (Papamitsiou & Economides, 2014) such as identifying learners' behavior/interaction and network

modeling and waste of resources	patterns, lack of responses in discussion forums, dropouts and absence from class and quizzes, isolated and warned students. Moreover, learners' attitudes (experience level indicators, learning interest, learning styles, learning goals and competences, and background information) and their affective traits should be taken into consideration in the recommendation process
Inappropriate curricula and teaching strategies	To design the curricula and materials that may facilitate the learning process, as well as identifying and developing effective instructional techniques (Vaitsis et al., 2014).
Oversights of quality assurance	Adaptive selection of the most appropriate next task to improve testing outcomes, mostly for below average students (Papamitsiou & Economides, 2014).
Concerns of privacy and ethical issues	Resolving that incompatibility will require new approaches that better balance the protection of privacy and the advancement of science in educational research (Daries et al., 2014).

As Huda et al., (2018) found that big data analytic process provides some advantages to transform the pattern of information fitted into the new environment of online learning resources (OLR) to enhance in developing the learning resources. They also mentioned that big data analytics technique integrated into online learning in the way to addressing the learning behavior is supposed to give insights in contributing the reference model of big data emerging technology for OLR initiative basis. Moreover, Big data analytics is a set of techniques that requires new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of a massive scale (Hashem et al., 2015). Thus, the two important techniques such as learning analytics and educational data mining are becoming the lingua franca for those institutions who seek to improve their strategic and operational decision-making abilities

(Liebowitz, 2017).

In higher education, learning analytics are beginning to be used for a number of applications that address student performance, outcomes and persistence (Picciano, 2012). In addition, learning analytics technique is also used to analyze interaction between the students with online educational resources of distance education (Acevedo, & Marín, 2015). The learning analytics technique includes web analytics (Huda, et al., 2018), activity analytics (Kim, Jo & Park, 2016), academic analytics (Campbell and Oblinger, 2007), and action analytics (Elias, 2011). The input of web analytics data derived from various activities such as conversation, electronic messages, and photos or videos etc. from different social sites. Thus, data variety in finding the corresponding and connection among them, which lead to using the analytic programs (Huda, et al., 2018). Campbell and Oblinger (2007) defined the term academic analytics in that they opted to study issues directly related to "one of higher education's most important challenges: student success." Elias, (2011) mentioned "Action analytics included deploying academic analytics to produce actionable intelligence, service-oriented architectures, mash-ups of information/content and services, proven models of course/curriculum reinvention, and changes in faculty practice that improve performance and reduce costs". Some factors from different studies but similarly to each other, such as academic analytics, web analytics, and action analytics were clustered as factors outlining as learning analytics techniques as one of the coding scheme in the current study. Another important coding scheme named "EDM technique", it includes a set of methods that apply data mining and machine learning techniques such as prediction classification and discovery of latent

structural regularities to the huge volume and idiosyncratic educational data. Those data are generated from many constructionist learning environments which allow students to explore and build their own artifacts, computer programs and media pieces (Berland et al., 2014). Finally, visual analytics (VA) is categorized as the third coding scheme of techniques which showing significant beneficial for users to gain insight into complex data. Keim, Andrienko, Fekete, Görg, Kohlhammer, & Melançon, (2008) mentioned VA integrates data analysis, visual representations, and interactive visual interface. Thus, Vieira, Parsons & Byrd (2018) suggested using VA as a technique to enabling sense making of complex educational data for instructors, students, and administrators. Therefore, three schemes (i.e., learning analytics technique, big data mining technique and visual analytics technique) are identified among the different big data techniques from big data analytics adopted in educational research as listed in Table 3.

Table 3. Coding scheme for big data techniques used and their definitions

Coding scheme for big data techniques	**Definition of coding scheme of big data techniques**
Learning analytics technique	LA has been defined as "an emerging area which explores the measurement, collection, analysis, and reporting of data associated with students' learning and environment" (Kim, et al., 2016; Chatti et al. 2012).
Educational data mining technique	EDM has been defined as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (IEDMS 2009).

Visual analytics technique	The techniques used to create tables, images, diagrams and other intuitive display methods to help understand the data (Chen & Zhang, 2014).

4. Results and discussion

4.1 The profile of the analyzed papers

In this study, a total of 143 papers published from 2010 to 2018 were analyzed. There was 1 paper from 2010, 7 from 2012, 4 from 2013, 17 from 2014, 15 from 2015, 17 from 2016, 42 from 2017, and 40 from 2018. This progressive growth in terms of publications seems to reveal a rising interest in the area of educational big data. With regards to the journals in which the papers were published, there wasn't a significant difference between them in terms of the number of papers: Computers and Human Behavior (6 papers), Computers and Education (5 papers), International Journal of Learning Technology (4 papers), Journal of Asynchronous (2 papers), British Journal of Educational Technology (BJET) (2 papers), and Learning Media and Technology (2 papers). The remaining papers were each published in different journals.

4.2 RQ1: The educational goals that big data applications aimed to address

The coding scheme that was created enabled the depiction of the current educational goals of big data research (Figure 2.).

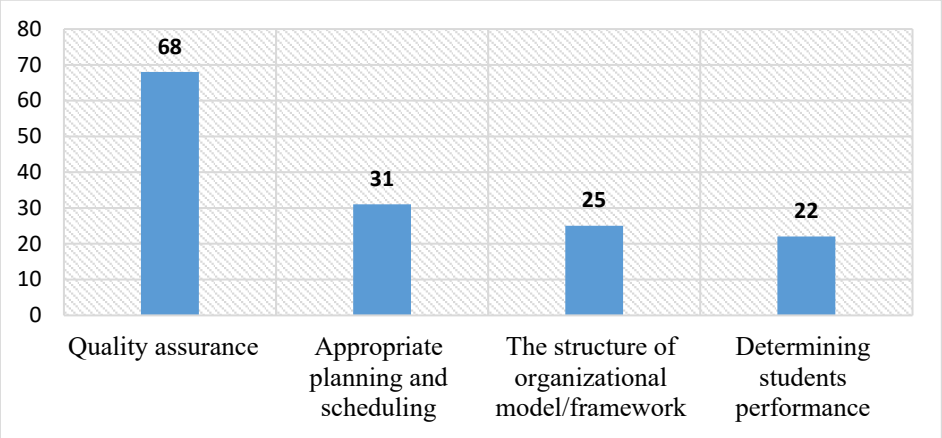

Figure 2. The distribution of educational goals in the educational big data research

There were four main educational goals identified among the selected 143 published papers: verifying quality assurance, appropriate planning and scheduling, organizational model/framework and determining student performance. It should be noted that one study might have addressed more than one goal. The analysis showed that 68 papers discussed quality assurance such as supporting the (self)-assessment of improved efficiency, the effectiveness of the learning process, appropriate teaching strategies, and proper decision making. This corresponds to 47.55% of the papers and demonstrates the predominance of quality related educational goals. These results are coherent with previous studies about the application of big data in other contexts, such as its use in enterprises for the enhancement of production efficiency and competitiveness in many aspects (Chen, et al. 2014) such as marketing, sales planning, operations and finance in enterprise and e-commerce (Chen, et al. 2014). Moreover, Eynon (2013) found that using big data as an administrator can identify failing schools and teachers. Given these results and existing literature, administrators, teachers and other decision makers should examine more

closely the opportunities offered by big data applications to improve the quality of education, by using data to identify and address possible shortcomings.

In total 31 papers stated that big data was used for planning and scheduling, including planning future courses, student course scheduling, planning resources allocation, admission and counseling processes, and developing curricula. As, it is mentioned in the literature, in higher education, academic data such as that, relating to administration and curricula, is of such size and nature that special techniques must be applied to discover new knowledge (Romero & Ventura, 2007). Big data has unlimited potential for effectively scheduling, processing and allocating resources planning in recent educational research. This brings to light a key purpose of the deployment of big data in education, in terms of planning and organization, which can be harnessed by both researchers and practitioners.

There were 25 papers, which focused on the organizational model/framework including predicting the probability of events of interest. The reason might be that big data application in the educational field is very new and will take some more years to mature (Picciano, 2012). Moreover, some researchers focused on developing a model/framework. This study found that 17.50% of papers focused on developing a organizational model/framework as their educational goal, which is the least mentioned educational goal in educational big data research. Thus, the results of this study suggest to stakeholders that to meet institutional demands alone with big data application might be required.

Finally, 22 papers stated that big data was used for student performance, knowledge, activities, test scores, isolated students, and early warnings received. According to previous research, the use of big

data for student performance included estimating the unknown value of student performance, knowledge, activities, test scores, isolated students, early warnings received, and overlooking taking courses (Papamitsiou & Economides, 2014). As Gunnarsson and Alterman (2012) mentioned, students' performance prediction models focus on statistical modeling and data mining techniques. Despite a considerable support by previous studies, and the fact that it is one of educations' main concerns, only five papers mentioned, as their goal, the determination of students' performance. This can indicate that in the field of education, big data is moving on to more complex goals and it is being considered as a strategy to address other, more comprehensive and diverse concerns, rather than to limit its use to more obvious goals.

4.3 RQ2: The educational problems addressed by applying educational big data

In educational big data research, four categories of educational problems were identified, which allow for the examination of the current problems in institutions as it is shown in Figure 3.

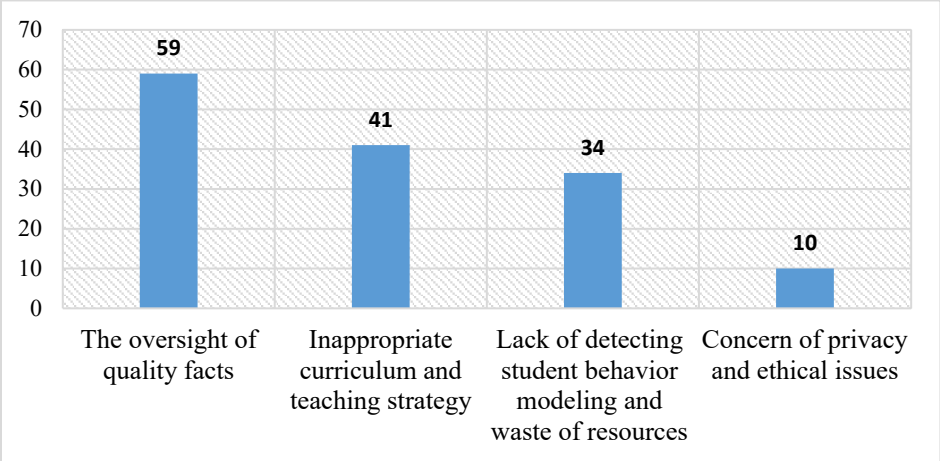

Figure 3. The distribution of problems addressed in the educational big data research

These four categories of educational problems are the oversight of quality facts, inappropriate curricula and teaching strategies, lack of modeling for detecting student behavior and the waste of resources, and concerns of privacy and ethical issues. The oversight of quality facts refers to the adaptive selection of the most appropriate next task to improve testing outcomes, mostly for below average students (Papamitsiou & Economides, 2014). There were 59 papers related to this issue, which similarly to the conclusions in section 4.2, shows an awareness for the importance of quality in education. There are several metrics that can be used for quality assurance using big data, which can be further explored in future research ventures, to assist both institutions and teachers to improve the delivery of education and ensure that the learning process is effective. This study found that 41% of papers focused on the oversight of quality facts in institutions, which are the highest ranking problems in educational big data research. In addition, as it was brought to light by previous studies, big data can potentially track almost all students' activities while in school in order to allow for appropriate scheduling and planning, including designing curricula and materials (Eynon, 2013) which may facilitate learning processes as well as identifying and developing effective instructional techniques (Vaitsis et al., 2014). There were 41 papers which focused on inappropriate curricula and teaching strategies, which places an emphasis on existing concerns pertaining to teaching methodology, which can, with big data, be more closely assessed for effectiveness, and on the curriculum itself, which can also be improved using big data. Addressing these concerns is fundamental for the proficiency of educational settings and it can result in

more teacher accountability, it can be used to test different teaching styles and to experiment with various techniques to teaching and developing curricula.

In total 34 papers addressed the lack of modeling for detecting student behavior and the waste of resources. The subtopics included identifying students' prerequisite matters such as their behavior, interactions and network patterns, lack of response in discussion forums, dropouts, absence from class and quizzes, isolated students, learners' attitudes (experience level indicators, learning interest, styles and goals), expected performance on tasks and recent navigation. Resource allocation and student performance and learning experience are determinant for the success of any educational institution and as such, it is not surprising that they are given significant importance in the papers that were reviewed. The lack of monitoring the student's behavior and activity on a real time basis can result in students missing the real time alerts, which might create failure of performance. Additionally, it can also lead teachers and institutions to be missing key knowledge that is determinant to improve education delivery. Therefore, it is important to focus on these problems and on big data's potential to address them. Finally, ten studies were conducted in the area of privacy and ethical issues, including focusing on better protection of privacy (Daries et al., 2014), and ethical considerations such as privacy, informed consent, and protection from harm (Eynon, 2013). In recent computing applications, in institutions, privacy and ethical issues play an important role. As it was highlighted in the literature, real time interactions of students with specific tasks in a computational platform can produce highly sensitive data, which prompts the need for improvements regarding privacy principles with respect to analytics and their

application in educational settings (Pardo & Siemens, 2014). Such analytics and applications can help institutions to make anchored decisions to protect their data and information. Recently, data scientists have noticed this issue and can further concentrate on overcoming it.

4.4 RQ3: The big data techniques used in the educational big data research

The coding scheme identified three categories of the techniques to the application of big data analytics to address the problems in institutions identified from the 143 studies (Figure 4): educational data mining technique (77 studies), learning analytics technique (71 studies), and visual analytics technique (13 studies). It should be noted that one study might have adopted more than one technique.

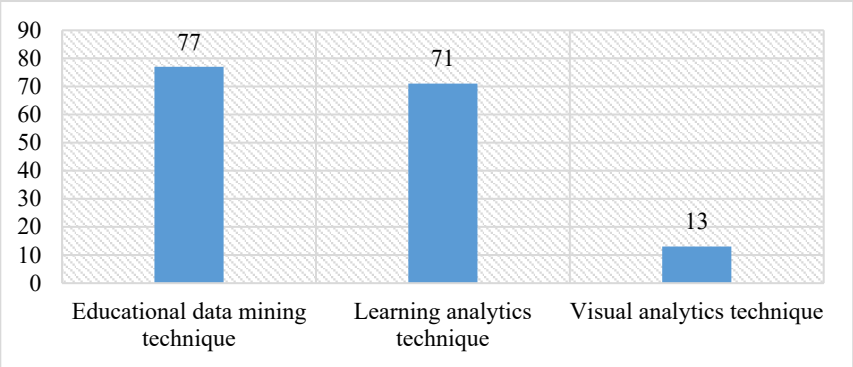

Figure 4. The distribution of big data techniques used in the educational big data research

As was mentioned in the literature, big data applications involves a set of techniques and technologies that requires new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of a massive scale (Hashem et al., 2015). The current study found that 53% of papers focused on EDM techniques in institutions in educational research. More specifically, in EDM, 25% of papers focused on predictive technique and 18% of papers focused on data driven technique, and 10% of papers focused on other techniques of EDM in institutions in the educational research. In the

future, researchers may develop a deeper focus on big data's success in the competitive era. Gomez-Aguilar et al. (2015) mentioned that EDM has techniques, which can be adopted to resolve the problems of educational research. The current study found that 49% of papers focused on the LA technique in institutions, making it as the predominant technique in educational big data research. The reason could be that increasing initiatives of analytics technique will continue transferring to educational sectors.

In Talis Aspire (MacNeill et al., 2014) study, the authors have completed reading list management solution based on usage data to provide educators/learners effective use which enhance opportunity for personalized learning (MacNeill et al., 2014). In the visual analytics technique, tables, figures, diagrams and other intuitive methods of display can be used to understand the data (Chen & Zhang, 2014). This study found that 9 % of papers focused on visual analytics technique in the educational big data research. The visualization aspect is a core to the understanding of the data that was collected. Visualization tools are the first contact that users making senses of the data and the manner in which it is represented determines its comprehensibility and the actions that can be taken.

A subsequent analysis shows how big data techniques have been applied to achieve different educational goals. As mentioned earlier that one study may have addressed more than one goal and technique. Thus, a bar graph is presented according to the three techniques and four goals as shown in figure 5.

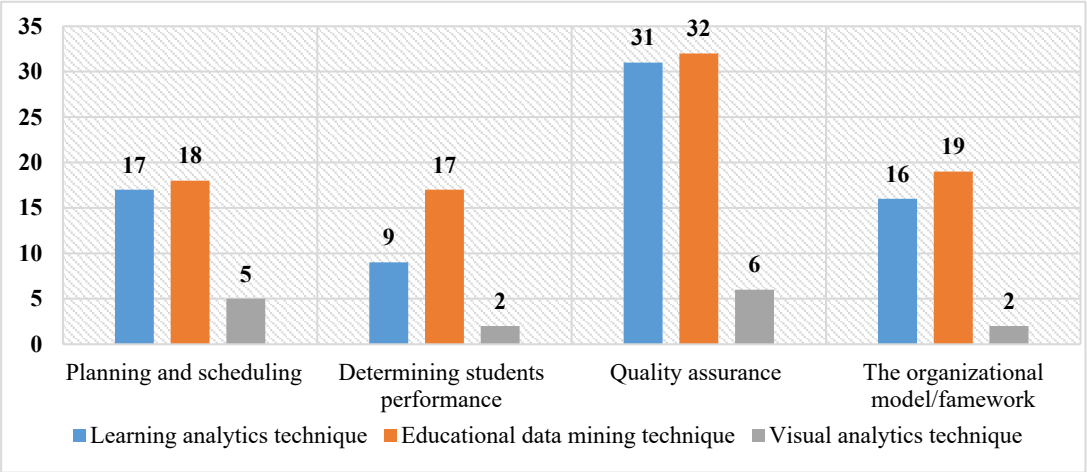

Figure 5. The distribution of big data techniques being applied to achieve different educational goals

Table 4. Big data techniques and the achievement of different educational goals

Educational goals	Learning analytics technique	Educational data mining technique	Visual analytics technique
Appropriate planning and scheduling	23%	21%	34%
Determine students' performance	12%	20%	13%
Verify quality assurance	43%	37%	40%
The organizational model/framework	22%	22%	13%

The educational data mining techniques are mostly used to achieve four educational goals (49%) as shown in Figure5. It is found that educational data mining (EDM) is widely used to ensure quality (37%), followed by the organizational model/framework (22%). EDM techniques can also assist in the achievement of appropriate planning (21%) and scheduling and in determining students' performance (20%) as shown in Table 4. Learning analytics techniques are used to achieve four educational goals (42%) as shown in figure 5. The learning analytics technique is used to achieve 43% of quality assurance

and 23% of appropriate planning and scheduling as shown in Table 4. As Agudo-Peregrina et al. (2014) mentioned, in a first and broad approximation, that learning analytics focuses on the analysis of automatically capturing data of student behavior (Chen & Zhang, 2014). In addition, visual analytics was used for quality assurance (40%) as well as it was for appropriate scheduling and planning (34%) as shown in Table 4. As Vaitsis et al. (2014) suggested, visual analytics could provide novel ways to represent curriculum and educational data. In another study, Gomez-Auiar et al. (2015) stated that "VA facilitates almost instant interaction, identification of patterns and discovery of new information not readily available in learning platforms" (p. 66). Therefore, the results suggest that researchers and data scientists should have a clear view when they are aiming to set any types of educational goals, so that they could choose appropriate techniques to justify those educational goals while applying big data in education.

A subsequent analysis also shows how big data analytics can solve different educational problems. As mentioned earlier that one study may have addressed more than one problems and techniques. The educational data mining techniques are mostly used to solve different educational problems (53%) as shown in Figure 6.

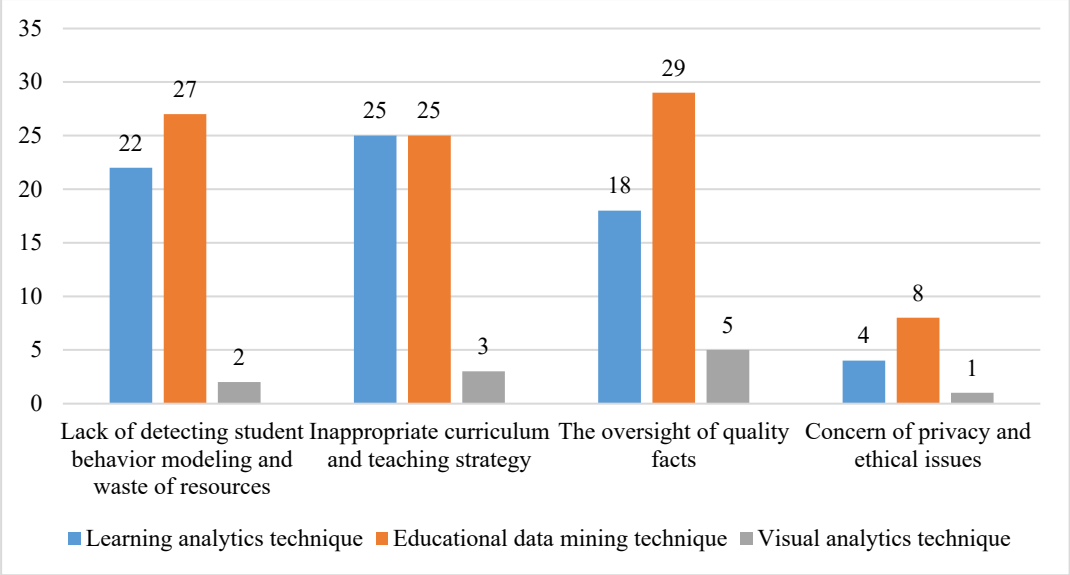

Figure 6. Big data techniques used to problems addressed in the educational big data research

Table5 Big data techniques used to problems addressed in the educational big data research

Educational Problems	Learning analytics technique	Educational data mining technique	Visual analytics technique
Lack of detecting student behavior modeling and waste of resources	32%	30%	18%
Inappropriate curricula and teaching strategies	36%	28%	27%
Oversights of quality assurance	26%	33%	46%
Concerns of privacy and ethical issues	6%	9%	9%

More specifically, the educational data mining techniques are mostly used to detect students' behavior modeling and waste of resources (30%), to fix oversights of quality assurance (33%), to solve inappropriate curricula and teaching strategy (28%) and to identify the concerns of privacy and ethical

issues (9%) as shown in Table 5. The EDM method enables precise formative assessment of complex constructs, and can present comprehensible actionable data to learners and teachers (Berland et al., 2014). The learning analytics techniques are used to solve different educational problems (41% papers) as shown in figure 6. This research's findings highlight that learning analytics technique is to solve the lack of detecting appropriate curriculum and teaching strategy (36%), to detect student's behavior modeling and waste of resources (32%), used to overlooking quality facts (26%) and concerning of privacy and ethical issues (6%) as shown in Table5. As Greller and Drachsler (2012) and Pardo and Siemens (2014) mentioned, for institutional entities, learning analytics play an important role of detecting and addressing issues regarding the retention of students, monitoring of graduation rates, and evaluating and improving courses. Visual analytics was used to fix for oversights quality assurance (46%) as it was to solve inappropriate curricula and teaching strategies (27%) as shown in Table 5. Therefore, administrators have an awareness of applying suitable big data techniques, which are in demand for addressing emergent educational problems.

5. **Conclusion**

This study reviewed the state-of-the-art on big data research in education including its application in addressing educational goals and educational problems, as well as the current techniques of big data in education. Firstly, this paper presented the evolution of big data in education, and identified specific educational goals such as appropriate planning and scheduling, determining student performance, verifying quality assurance, and developing the organizational models/frameworks. The educational

goals mentioned in the reviewed papers placed an emphasis on quality assurance, revealing a rising concern with the quality of education delivery and the examination of how big data can serve this purpose. As educational institutions become increasingly competitive, the quality of the education that they offer their students is central. This prominence of quality, enabled by big data, and the possibilities it affords teachers and institutions to assess how their courses are being delivered and completed by the students, becomes a fundamental part of evaluating learning.

The analysis then focused on the educational problems from the teachers', learners' and administrators' perspectives: the lack of detecting student behavior modeling and waste of resources, inappropriate curricula and teaching strategies, oversight of quality assurance, and concerns of privacy and ethical issues. With respect to the problems that were mainly mentioned in the papers, there wasn't a significant discrepancy between the four categories, with the exception of privacy and ethical concerns, which was mentioned only by a few papers. Despite their importance for learning analytics, for the papers that were sampled for this study, there were not as relevant as the other categories. These findings are important to inform developers of the teachers' needs and to adjust big data techniques to suit them. At the same time, they disclose the concerns that are affecting the educational context in an era of accountability. As it becomes easier to evaluate the shortcomings of education delivery, institutions must use all means available, including technology and big data, to take on their responsibility to ensure that students have access to the best quality education.

Finally, the analysis reviewed several potential techniques to the use of big data in education including the learning analytics technique, the educational data mining technique, and the visual analytics technique to solve the problems that arise in institutions. The subsequent analysis that identified which techniques of big data are more suitable to achieve specific types of educational goals, found that the educational data mining technique is mostly used for and might be very suitable to ensure quality assurance in educational settings. This type of analysis is particularly useful to determine which technique is more relevant in the context of specific educational goals and concerns. This alignment can represent a significant difference in the results that are obtained. Different techniques have different purposes that might be more pertinent to address certain objectives and might be more valuable in addressing particular problems.

In the future, significant challenges and issues need to be addressed by academia. Researchers, practitioners, data scientists and social science scholars should collaborate to ensure the long-term success of applying educational big data in education. Moreover, it is also important to form the development of efficient institutional structures with the various aspects of systematically collecting institutional data and information to enable a big picture for more promising big data applications. Future research ventures, can equally, take the results of this study, and further explored them by complementing these conclusions with a consultation of different stakeholders' opinions, such as students, teachers, data scientists and institutions.

Educational big data is at a point of its evolution where researchers and practitioners are transitioning from a point of mere awareness to action. Whereas the first efforts to promote the use of educational big data focused on disseminating knowledge concerning its benefits and the possibilities it represents, the current research scenario is more determined in examining real application. In addition, the techniques that are used to transform data into usable information assume a more relevant part, as they prove to be a determining factor in educational big data's value.

Acknowledgement

This work was supported by the National Science Council, Taiwan under project numbers MOST-107-2511-H-224-007-MY3, and MOST-106-2511-S-224-005-MY3. This work was also partially supported by Doctoral Foundation Project, Business school, Shandong University of Technology, China.

References

Acevedo, Y. V. N., & Marín, C. E. M. (2015). Towards a decision support system based on learning analytics. *Advances in Information Sciences and Service Sciences*, *7*(1), 1-12.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, *19*(1-2), 205-220.

Boscardin, C., Fergus, K. B., Hellevig, B., & Hauer, K. E. (2018). Twelve tips to promote successful development of a learner performance dashboard within a medical education program. *Medical teacher*, *40*(8), 855-861.

Brodie, M.L., Greaves, M., & Hendler, J.A. (2011) Databases and AI: The Twain Just Met, 2011 STI semantic Summit, Riga, Latvia, July 6-8.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, *42*(4), 40.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Mobile Networks and Applications*, *19*(2), 171-209.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314-347.

Chaurasia, S. S., & Frieda Rosin, A. (2017). From Big Data to Big Impact: analytics for teaching and learning in higher education. *Industrial and Commercial Training*, *49*(7/8), 321-328.

Chiang, R. H., Goes, P., & Stohr, E. A. (2012). Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline. *ACM Transactions on Management Information Systems (TMIS)*, *3*(3).

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, 46(5), 904-920.

Daniel, B. K., & Butson, R. (2013). Technology enhanced analytics (TEA) in higher education, *Proceedings of the International Conference on Educational Technologies*, 29 November –1 December, 2013, KualaLumpur, Malaysia (pp. 89–96).

DeWitt, D., & Gray, J. (1992). Parallel database systems: the future of high performance database systems. *Communications of the ACM*, *35*(6), 85-98.

Dahdouh, K., Dakkak, A., Oughdir, L., & Messaoudi, F. (2018). Big data for online learning systems. *Education and Information Technologies*, *23*(6), 2783-2800.

Dyckhoff, A. L., Zielke, D., Bültmann, M., Chatti, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Educational Technology & Society*, *15*(3), 58-76.

Elias, T. (2011). Learning analytics: Definitions, Processes and Potential. *Learning*, 1-22.

Eynon, R. (2013). The rise of Big Data: what does it mean for education, technology, and media research? *Learning, Media and Technology*, 38(3), 237-240.

Ellaway, R. H., Pusic, M. V., Galbraith, R. M., & Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical teacher*, 36(3), 216-222.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iView*, *1142*, 1-12.

Gunn, A. (2018). Metrics and methodologies for measuring teaching quality in higher education: developing the Teaching Excellence Framework (TEF). *Educational Review*, 70(2), 129-148.

Gómez-Aguilar, D. A., Hernández-García, Á., García-Peñalvo, F. J., & Therón, R. (2015). Tap into visual analysis of customization of grouping of activities in eLearning. *Computers in Human Behavior*, 47, 60-67.

Gorissen, P., van Bruggen, J., & Jochems, W. (2013). Methodological triangulation of the students' use of recorded lectures. *International Journal of Learning Technology*, 8(1), 20-40.

Gunnarsson, B. L., & Alterman, R. (2012). Predicting failure: A case study in co-blogging. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 263–266). ACM.

Gupta, B., Goul, M., & Dinter, B. (2015). Business Intelligency and Big Data in Higher Education: Status of a Multi Year Model Curriculum Development Efforts for business School Undergraduates, MS Graduates, and MBAs,'. *Communication for the Association for Information System, 36*(1), 449-476.

Huda, M., Maseleno, A., Teh, K. S. M., Don, A. G., Basiron, B., Jasmi, K. A., Ismail, M. M., Nasir, B. M.,& Ahmad, R. (2018). Understanding Modern Learning Environment (MLE) in Big Data Era. *International Journal of Emerging Technologies in Learning*, *13*(5), 71-85.

Ju, S. Y., Song, M. H., Ryu, G. A., Kim, M., & Yoo, K. H. (2014). Design and implementation of a dynamic educational content viewer with big data analytics functionality. *International Journal of Multimedia and Ubiquitous Engineering*, *9*(12), 73-84.

Kim, J., Jo, I. H., & Park, Y. (2016). Effects of learning analytics dashboard: analyzing the relations among dashboard utilization, satisfaction, and learning achievement. *Asia Pacific Education Review*, 17(1), 13-24.

Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. *In Information visualization* (pp. 154-175). Springer, Berlin, Heidelberg.

Krippendorff, K. (2013). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. *Annals of the International Communication Association*, *36*(1), 481-499.

Littlejohn, A., & Pegler, C. (Eds.). (2014). Reusing open resources: Learning in open networks for work, life and education. Routledge.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20-33.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. *Robustness of Deep Learning Systems Against Deception*.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., & Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, *21*(6), 957-958.

Mergel, I. (2016). Big data in public affairs education. *Journal of Public Affairs Education*, *22*(2), 231-248.

Ndukwe, I., Daniel, B., & Butson, R. (2018). Data science approach for simulating educational data: towards the development of teaching outcome model (TOM). *Big Data and Cognitive Computing*, *2*(3), 24.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438-450.

Papamitsiou, Z. K., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, *17*(4), 49-64.

Reyes, J. A. (2015). The skinny on big data in education: Learning analytics simplified. *TechTrends*, *59*(2), 75-80.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3), 528-549.

Risling, T. (2017). Educating the nurses of 2025: Technology trends of the next decade. *Nurse education in practice*, *22*, 89-92.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Romero-Zaldivar, V. A., Pardo, A., Burgos, D., & Kloos, C. D. (2012). Monitoring student progress using virtual appliances: A case study. *Computers & Education*, *58*(4), 1058-1067.

Santoso, L. W. (2017). Data warehouse with big data technology for higher education. *Procedia Computer Science*, *124*, 93-99.

Shen, G. R. (2018). Chinese College English Teachers' Ability to Develop Students' Informationized Learning in the Era of Big Data: Status and Suggestions. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(6), 2719-2729.

Shun-ping, W. E. I. (2013). Learning Analytics: Mining the Value of Education Data under the Big Data Era. *Modern Educational Technology*, 2, 003.

Song, I. Y., & Zhu, Y. (2016). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364-373.

Song, J., Zhang, Y., Duan, K., Hossain, M. S., & Rahman, S. M. M. (2018). TOLA: Topic-oriented learning assistance based on cyber-physical system and big data. *Future Generation Computer Systems*, *75*, 200-205.

Tetko, I. V., Engkvist, O., Koch, U., Reymond, J. L., & Chen, H. (2016). BIGCHEM: challenges and opportunities for Big Data analysis in chemistry. *Molecular informatics*, *35*(11-12), 615-621.

Tulasi, B. (2013). Significance of Big Data and analytics in higher education. *International Journal of Computer Applications*, *68*(14).

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6* (70), 1.

Liebowitz, J. (2017). Thoughts on recent trends and future research perspectives in big data and analytics in higher education. In *Big data and learning analytics in higher education* (pp. 7-17). Springer, Cham.

Vaitsis, C., Nilsson, G., & Zary, N. (2014). Big data in medical informatics: improving education through visual analytics. *Studies in health technology and informatics*, 205, 1163-1167.

Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, *122*, 119-135.

Wang, S., Liu, Y., & Padmanabhan, A. (2016). Open cyberGIS software for geospatial research and education in the big data era. *SoftwareX*, *5*, 1-5.

Wen, J., Zhang, W., & Shu, W. (2018). A cognitive learning model in distance education of higher education institutions based on chaos optimization in big data environment. *The Journal of Supercomputing*, *75*(2), 719-731.

Williamson, B. (2015). Governing software: Networks, databases and algorithmic power in the digital governance of public education. *Learning, Media and Technology*, *40*(1), 83-105.

Xian, H., & Madhavan, K. (2014). Anatomy of scholarly collaboration in engineering education: A big-data bibliometric analysis. *Journal of Engineering Education*, *103*(3), 486-514.

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, *47*, 168-181.

Xie, C., Zhang, Z., Nourian, S., Pallant, A., & Hazzard, E. (2014). A time series analysis method for assessing engineering design processes using a CAD tool. *International Journal of Engineering Education*, *30*(1), 218-230.