Manuel J. Gomez and Ruipérez-Valiente, J. A. Analyzing the Evolution of Digital Assessment in Education Literature Using Bibliometrics and Natural Language Processing. In Handbook of Research on Digital-Based Assessment and Innovative Practices in Education (pp. 178 - 200). IGI Global. ISBN 9781668424681

DOI: 10.4018/978-1-6684-2468-1.ch009

https://doi.org/10.4018/978-1-6684-2468-1.ch009

Analyzing the Evolution of Digital Assessment in Education Literature Using Bibliometrics and Natural Language Processing

Manuel J. Gomez and José A. Ruipérez-Valiente
University of Murcia, Spain

ABSTRACT

Over the last decade, we have seen a large amount of research being performed in technology-enhanced learning. Within this area, the use of digital assessment has been gaining a lot of popularity. In this work, the researchers aim to identify the main topics in this area within the last 15 years, proposing a new methodology to perform a text analytics and bibliometrics driven approach. Authors collected all the metadata and full text from papers from the last 15 years within this area, trying to find the main topics across all the papers, along with hidden relationships between authors and papers. The analysis in this work has focused on three main objectives: 1) Discover which are the main topics of digital assessment in education based on topic modeling and keywords analysis, using both full text and metadata. 2) Discover the evolution of said topics over the last 15 years of research. 3) Discover the primary authors and papers, along with hidden relationships between existing communities. Authors expect this study to overcome the current limitations in qualitative analysis in research papers, objectively processing large amounts of research and shedding some light on the latest trends within this area.

Keywords: Technology-enhanced Learning, Digital Assessment, Natural Language Processing, Nlp-enhanced Bibliometrics, Topics, Keywords, Network Analysis, Education

INTRODUCTION

Over the last two decades, technology has become an inherent part of education, by generating multiple new applications to improve the learning process (Raja & Nagasubramani, 2018) . These applications have been diverse, including learning management systems (LMSs) to facilitate course development (Sclater, 2008), smart devices and classrooms (Zhu et al., 2016), artificial intelligence applications in education (L. Chen et al., 2020), or even games (Ruipérez-Valiente & Kim, 2020). Multiple research studies have shown the benefits of educational technologies to improve learning. Within the broad spectrum of technology enhanced learning, researchers focus on this chapter specifically on digital assessment, also known as e-assessment (Whitelock, 2009).

Broadly speaking, assessment can be defined as a process of drawing inferences based on evidence. Within the context of education, it has been defined in different ways. Our view of assessment aligns with the following definition provided by Huba and Freed "*Assessment is the process of gathering and discussing information from multiple and diverse sources in order to develop a deep understanding of what students know, understand, and can do with their knowledge as a result of their educational experiences*" (Huba & Freed, 2000). Therefore, when we talk about digital assessment, we mean that this assessment process is supported by digital technologies at some point. In that sense, digital assessment

can take many forms. The simplest form might be the use of digital quizzes and exams in order to facilitate performing assessment over a digital medium (Dellos, 2015). However, that is just the very first step in the process.

Over the last decade there have been multiple digital assessment approaches aiming to improve the educational process. For example, numerous intelligent tutoring systems (ITSs) (Anderson et al., 1985) have emerged that frequently implement adaptive learning algorithms in order to adapt the assessment items to the current knowledge of students automatically. Other examples can include the use of health assessment systems (Saini et al., 2012) to evaluate the recovery of patients that are training to get better or the implementation of digital assessment solutions that can improve the trustworthiness of remote learning against academic dishonesty (Jaramillo-Morillo et al., 2020). Finally, a prominent example is the field of game-based assessment, that aims to perform stealth assessment of competencies and skills through the use of the data generated in games (Gomez et al., 2021). These different applications have reported clear benefits that can improve the assessment process at different levels. Some studies are focused on reporting new tools for digital assessment, others focus on the instructional design to include those digital assessments within the curriculum, and others evaluate the outcomes of using digital assessment.

Therefore, the field of digital assessment is quite broad. In this chapter, we aim to perform a longitudinal study of this literature. However, performing a qualitative review of all the literature is quite consuming, given that this is an ample field with many publications. Consequently, we propose to perform a bibliometrics study enhanced with natural language processing (NLP) techniques (Wolfram, 2016) to automatically extract the topics of the papers based on their full text or abstract. We will investigate the main topics and the evolution of those over the last 15 years. Moreover, we also aim to use network analysis approaches to inspect the research community with a co-author network and a citation paper network. More specifically, we have the following research questions (RQs):

RQ1: What are the main topics of digital assessment in education literature based on keywords and topic modeling?
RQ2: What has been the evolution of such topics over the last 15 years of research?
RQ3: What are the communities of authors and papers in this topic based on a network analysis perspective?

The remainder of the chapter is organized as follows: First, the authors perform a background review of the bibliometrics and network analysis foundations, as well as previous work on digital assessment. Then, the authors extensively describe the methodology pursued to perform the NLP-enhanced bibliometrics study of digital assessment field. Then, the authors present the results of each one of the RQs. We finalize with a discussion about our findings, limitations, conclusion, and future work directions.

BACKGROUND

In this research, authors address a set of different research areas, presenting a novel methodology to provide interesting and useful information through different analyses on digital assessment in education literature. One of these different areas is bibliometrics. The term “statistical bibliography” seems to have been first used by E. Wyndham Hulme in 1922 when he delivered two lectures as the Sandars Reader in Bibliography at the University of Cambridge. However, this term has never been found satisfactory, and this feeling is fairly general by many other researchers in the field (Pritchard & others, 1969). The term is not very descriptive and can be confused with statistics or bibliographies on statistics. Therefore, a better name suggested for this subject is “bibliometrics.”. It was first used, so far as can be ascertained, in the Journal of Documentation in 1969 (Burchfield, 1972). Since then, numerous definitions have emerged: “the application of mathematics and statistical methods to books and other media of communication,” or

“quantitative analyses of the bibliographic features of a body of literature” are only two of the existing definitions of this field (Broadus, 1987). Despite the fact that this statistical analysis of publications has been practiced since the 1920s, bibliometric activity grew significantly with the emergence of new citation mapping tools starting with the ISI’s citation indices in the 1960s (De Bellis, 2009). Moreover, since the turn of the century, there has been a proliferation of bibliometric tools and indicators from the bibliographic database suppliers and academic researchers working in this field (Cox et al., 2019).

Then, bibliometrics could measure, for example, an article's impact. Bibliometric methods can estimate how much influence or impact a selected research article has on future research, and it usually does this by counting the number of times the article is cited after it is published (Cooper, 2015). However, at the level of individuals, bibliometrics usually measures the productivity of research but does not necessarily say anything about quality or the competence of researchers as teachers. In addition, it is important to note that the indicators obtained from bibliometric databases are not indicative in an absolute value terms, but they take taking their full significance only in relative terms when comparing them comparison with those of other groups (Okubo, 1997). The researchers found several studies that aimed to perform bibliometric analysis in research papers: authors in (Kokol et al., 2021) collected 6,557 medical publications from 1970 to 2018 from the Scopus bibliographic database, identifying 33 historical roots and 16 clinical areas, and concluding that the literature production trend was positive. Another example is the work in (Fan et al., 2020), where authors retrieved 864 publications about COVID-19 from both English and Chinese databases, analyzing the different authors and journals, countries and institutions, or the co-occurrence of keywords, among other analyses.

Moreover, bibliometric information can be used to discover more complex information than only quantitative measures, and network analysis is becoming increasingly popular as a general methodology for understanding complex patterns of interaction. It examines actors who are connected directly or indirectly by one or more different relationships, comprising graphical representations of the relationships (edges) between variables (nodes) (Hevey, 2018). Moreover, it assumes that relationships are important, and its benefits include: identifying individuals or teams that play central roles, make out opportunities to accelerate knowledge flows across functional boundaries, or strengthen the efficiency and effectiveness of existing, formal communication channels (Serrat, 2017). Moreover, that use of NLP for mining scientific papers leads us to the research topic on “NLP-enhanced Bibliometrics,”, which aims to promote interdisciplinary research in bibliometrics, NLP, and computational linguistics in order to enhance the ways bibliometrics can benefit from large-scale text analytics and sense mining of papers (Atanassova et al., 2019).

This area of network analysis can also be applied to analyze research papers, including the relationships between authors or between papers. For example, authors in (Ji et al., 2021) collected 83,331 statistical articles published in 36 representative journals, and used network analysis to estimate research interest of authors, but also to discover a multi-layer community tree to visualize the author migrations in different sub-areas. We can see another example in (Zhang et al., 2021), where authors investigated three computer science education conferences (SIGCSE Technical Symposium, ITiCSE and ICER), and analyzed authorship and affiliation details for over 4,500 publications, concluding that the community is open to newcomers, and both the number of authors, and the overall level of collaboration is growing.

While bibliometrics are commonly used to provide “quantitative” indicators about research papers, in this research, Natural Language Processing (NLP) is also used to add a “qualitative” aspect about analyzed publications. NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech in order to do useful things (Chowdhury, 2003). It is considered to be a subfield of Artificial Intelligence and linguistics, and its history started in the 1950s, when Alan Turing published an article called “Machine and Intelligence” (Chopra et al., 2013). Nowadays, with the huge amount of texts available to be analyzed, NLP expertise can be used with

different purposes, such as automatic translation, question answering, indexing and searching large texts, knowledge acquisition, or text generation, among many others (Chowdhary, 2020). However, there are also some limitations in the area. The principal difficulty in processing natural language is the pervasive ambiguity found at all levels, such as lexical, semantic, structural, or pragmatic ambiguity, among others. All these forms of ambiguity may interact and produce an extremely complex interpretation process (Allen, 2003). When examining current literature, authors found many works applying NLP and text mining to discover trends in research papers of many different fields. Those fields include medication usage (E. S. Chen et al., 2007), educational technologies (X. Chen et al., 2020), the COVID-19 pandemic (Boon-Itt & Skunkan, 2020), or even smart cities (Park & Lee, 2019).

Moreover, in this work, researchers aim to use those previous areas to analyze research on digital assessment. On the one hand, assessment is central to the practice of education. For students, good performance gives access to further educational opportunities and employment. For teachers and schools, it provides evidence of success as individuals and organizations (Ridgway et al., 2004). On the other hand, with the increasingly broad use of the Internet and the enormous rise in user-numbers, many businesses realized that the Internet represented a new resource for the search of suitable employees for their recruitment procedures (Laumer et al., 2009). Common advantages of digital assessment tools are the speed, availability, consistency, and objectivity of assessment (Amelung et al., 2010). However, the implementation of e- assessment also faces some challenges: training with inexperienced students, accessibility of computer and internet, difficulties in scoring questions with open responses, or assessing groups of students (Alruwais et al., 2018).

Finally, as far as we are concerned, this is the first study to apply these areas to analyze trends in digital assessment. However, authors found that previous work (Gürcan & Özyurt, 2020) analyzed 27,735 journal articles regarding e-learning, discovering five main dimensions using probabilistic topic modeling. The most representative of those dimensions was assessment, representing a 28.15%. The rest of discovered dimensions were learning environments, teaching models, teaching areas, and teaching tools. Moreover, the topic assessment had sub-topics such as “Feedback assessment,”, “Quality evaluation,”, or “Data analysis.”. In addition, (Gurcan et al., 2021) performed a topic modeling analysis on 41,925 peer-reviews journal articles, revealing that the most important topics were “MOOC,” “learning assessment,” and “e-learning systems.” Since this is the first study to analyze current trends in digital assessment using NLP and bibliometrics, the authors expect to provide an overview of current literature and gather insights about existing communities and how trends have evolved over time.

METHODOLOGY

To conduct the research, the authors divided their work in different stages: a) data extraction, b) data pre-processing, c) final data collection, d) NLP and keyword analysis, and e) network analysis. The entire methodology process is represented in Figure 1.

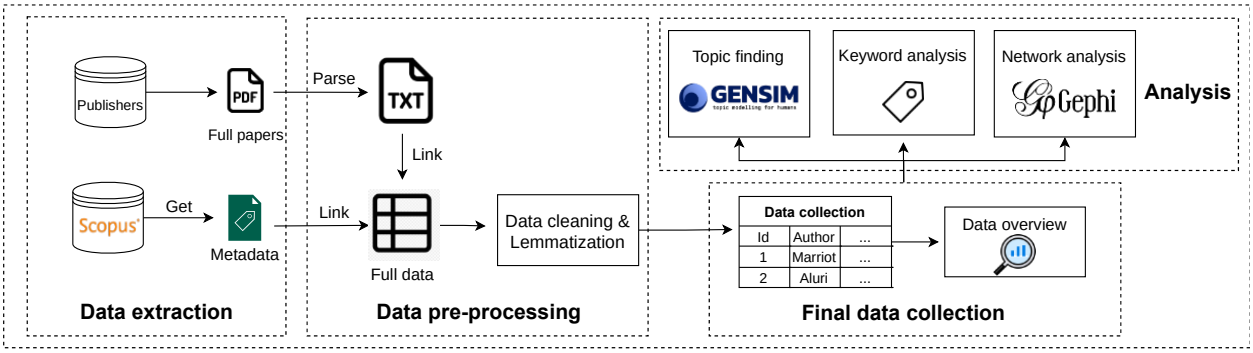

Figure 1. Complete methodology followed to conduct the research

Data Extraction

The first step in the analysis was to get all the metadata necessary to begin the research. To limit the scope of the analysis we have considered the following criteria:

- Papers that contain "education" and ("digital assessment," or "digital-based assessment," or "e-assessment") in the title, abstract or keywords.
- Papers published during the last 15 years.
- Papers published in English.
- Papers that have been published in a book, journal, or conference.

Then, the final query is the following one:

TITLE-ABS-KEY ("education") AND (TITLE-ABS-KEY ("digital assessment") OR TITLE-ABS-KEY ("digital-based assessment") OR TITLE-ABS-KEY("e-assessment")) AND (EXCLUDE (PUBYEAR , 2006) OR EXCLUDE (PUBYEAR, 2005) OR EXCLUDE (PUBYEAR, 2004) OR EXCLUDE (PUB-YEAR, 2003) OR EXCLUDE (PUBYEAR, 2002) OR EXCLUDE (PUBYEAR, 1999) OR EXCLUDE (PUBYEAR, 1997) OR EXCLUDE (PUBYEAR, 1969)) AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar") OR LIMIT TO (DOCTYPE, "ch") OR LIMIT-TO (DOCTYPE, "bk")) AND (LIMIT-TO (LANGUAGE, "English"))

Moreover, the downloaded metadata includes the title, keywords, abstract, source of publication, publication year, authors and so on. Then, the next step was to collect the papers' full text (if available). To achieve this, the authors used different databases from different publishers (e.g., Springer Link database, ACM Digital Library) to download each paper separately. Since not all the paper's full text was available to be downloaded, in these cases the abstract of each paper was used instead.

Data Pre-processing

Once authors had all the necessary data, the next step was to parse every PDF file into a TXT (plain text) file. Researchers considered the use of different libraries to achieve this conversion (namely *pdfMiner*, *pdfPlumbe*r, *pyPdf* and *PdfToText*). To ensure that the library used can parse the papers appropriately, researchers first made a parsing evaluation (testing if the library was able to parse the PDF files), and then a manual text review (comparing some TXT files with the original PDF files to check the quality of the conversion). After this evaluation, results suggested that the best library was *PdfToText* (Palmer, 2021), parsing 100% of the papers with high fidelity, and being able to parse even double-column PDF correctly. Once the authors had the plain text files, the next step was to link each paper's text to its metadata. Python functionalities are used for this purpose, merging the entire manuscript and the metadata in a single data structure. This is done by analyzing each paper's full text's first sentences and comparing it against the paper title originally available in the metadata.

After collecting all the data, authors had to pre-process and clean each paper's full text. First, only the paper's main body is kept (removing title, authors, affiliations, and references from the full text). The researchers performed additional cleaning actions by removing, for example, unnecessary URLs, numbers, or additional space characters. Moreover, to apply NLP techniques afterward, we need to define a set of "stop words," which will not be considered in the text analysis. In addition to the set of stop words defined by default, the authors added some more words after reviewing our data, such as "et," "al,"

“abstract,” “table,” or “figure,” which are common words that appear in every document but do not provide helpful information to the analysis.

From now, we can start treating each paper (or abstract) as a “document,”, since most of the cleaning process has endedfinalized. Once the text is cleaned, the authors lemmatized every document using *pywsd* library. The lemmatization is the process of converting a word to its base form, and its implementation in *pywsd* works as follows: (1) It tokenizes the string, dividing it into a set of tokens (words); (2) It uses a Part-Of-Speech (POS) tagger to map each word to a POS tag (adverb, noun, adjective, etcetera); (3) It calls the lemmatizer with the token and the POS tag to get the base form of the word. The use of the Part-Of-Speech (POS) tagger is crucial to remove language ambiguities. For example, if the lemmatizer finds the sentence “Learning is good,”, the POS tagger will identify “learning” as a name, and the lemmatized word will be “learning.”. However, if the lemmatizer finds the sentence “The student is learning,”, the POS tagger will identify “learning” as a verb, and the lemmatized word will be “learn.”.

Final Data Collection

In this stage, the complete data collection (including full texts and abstracts) is cleaned and ready to be analyzed in depth. After performing further analyses, the authors wanted to make a data collection overview in order to explore the final data and show some descriptive statistics, such as the number of documents, the number of different sources (e.g., conferences, books), or the number of words in the collection.

In this research, the final data collection contains a total of 566 documents from 453 different sources. The most important sources are shown in Table 1. From these 566 documents, 48 of them (8.5%) correspond to paper abstracts (papers whose full text was not available to be downloaded). The corpus contains a total of 2,655,131 words, with 85,344 unique words. In addition, the researchers performed a word cloud model to get an overview of the most representative words, which can be seen in Figure 2. A “word cloud” is a visual representation of word frequency derived from written text, and they can serve as a starting point for a deeper analysis, helping to judge whether a given text is relevant to a specific information need (Heimerl et al., 2014). In the word cloud, the more often the word appears within the corpus being analyzed, the larger it appears in the image generated, providing a synopsis of the main themes contained within the text (Atenstaedt, 2017).

Table 1. Main sources where papers were published in

Source title	Number of publications
IEEE Global Engineering Education Conference	24
Communications in Computer and Information Science	17
ACM International Conference Proceeding Series	14
Lecture Notes in Computer Science	14
Journal of Dental Education	13
British Journal of Education Technology	11
Proceedings of the European Conference on e-Learning	10
Assessment and Evaluation in Higher Education	9
European Journal of Dental Education	7
Advances in Intelligent Systems and Computing	6

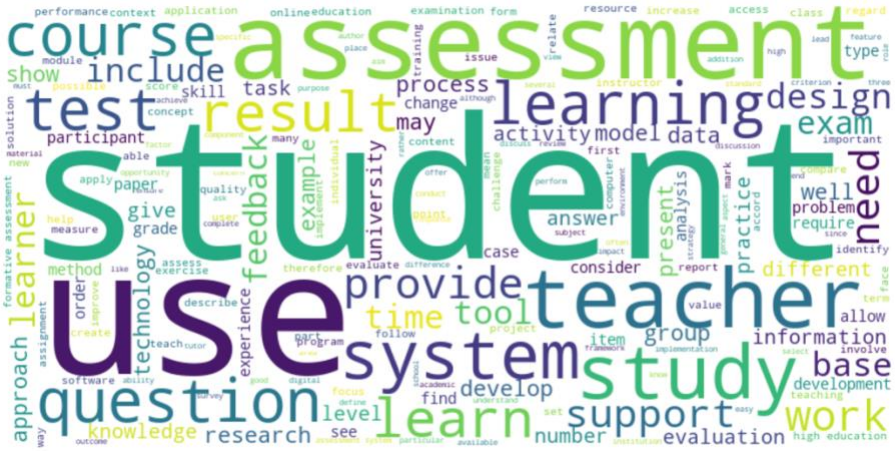

Figure 2. Word cloud model

As expected, the most frequent and important words are related with technology, assessment, and education, such as "student,", "assessment,", "course,", "system," and "test.". As we aimed with our selection of keywords in the initial search, we have gathered a myriad of papers revolving around the desired area, and we further confirm that the studies found are relevant to our search.

NLP and Keyword Analysis

The authors have used Latent Dirichlet Allocation (LDA) topic modeling to the data, using the full text or the abstract of each paper. Specifically, *gensim* library and its *ldaMallet* model have been used. Generally, in LDA each document can be described by a distribution of topics, and each topic can be described by a distribution of words; ldaMallet has almost the same approach as LDA: LDA uses the variational Bayes approach; and *ldaMallet* uses an optimized Gibbs sampling algorithm for LDA (Thooriqoh et al., 2021; Yao et al., 2009).

There are several metrics for evaluating the optimal number of topics. Recent studies have shown that the classic predictive likelihood metric and human judgment are often not correlated, and even sometimes slightly anti-correlated (O'callaghan et al., 2015). This has led to many studies that have focused upon the development of topic coherence measures. In this study, the authors use two of these coherence measures, both of which have been shown to match well with human judgements of topic quality: *C_v* and *C_umass* (Röder et al., 2015; Stevens et al., 2012). Then, to apply topic modelling, we followed the next steps:

Step 1: We created multiple models using different number of topics to decide the optimal number, based on the coherence measures specified previously. Researchers determined that eight was the optimal number of topics, with a *C_v* score of 0.38 and a *C_umass* score of -0.39.

Step 2: We made an initial manual topic labeling based on the first five keywords of each one.

Step 3: We manually reviewed ten random papers having a high proportion (>50%) of each topic (including title, keywords and abstract) to delimit topics more precisely.

Step 4: We assigned the final label to each topic.

Moreover, another approach to discover the main topics is to use the papers' keywords. To analyze them, the authors applied to the keywords the same cleaning method previously applied to the papers and abstracts, and then inspected the data manually to merge similar keywords into a single one. In that way,

when we count the keywords, if we find, for instance, “e-assessment” and “digital assessment,”, both keywords are merged, and their number of appearances are aggregated.

Network analysis

The next last step in this work was to perform network analysis using the metadata from the collected papers. Specifically, researchers built two different networks:

- **Co-authorship network**. Co-authorship is one of the most tangible and well documented forms of scientific collaboration, brining different talents together to give scientific credibility (Glänzel & Schubert, 2004; Kumar, 2015). A co-authorship network is an undirected graph that describes the authors working together within a collection of documents. Each node in the graph represents an author in the collection, and each edge is connected from one author to another that have shared one or more papers. In the network, each author is represented with its full name as identifier.
- **Citation network**. A citation network is a directed graph that describes the citations within a collection of documents. Each node in the graph represents a document in the collection, and each edge is directed from one document toward another that it cites. Since citations of others papers are hand-picked by the authors as being related to their research, the citations can be considered to judge relatedness (Lu et al., 2007). In this network, an identifier is generated for each paper, concatenating the first author’s name with the first work of the paper title and the year of publication (e.g., if the paper title is “Electronic integrity issues in e-assessment security” (Apampa et al., 2008), the first author is “Apampa K.M.,” and the year of publication is 2008, the identifier will be “ApampaElectronic2008”).

In this research study, the researchers used the papers’ full text to obtain the information about the citations of each paper, and the metadata to obtain the information about the different authors. Then, both networks are created using Gephi, an open-source network analysis and visualization software package written in Java on the NetBeans platform (Bastian et al., 2009).

RESULTS

RQ1. Main Topics of Digital Assessment in Education Based on Keywords and Topic Modeling

As a preliminary output of the results, we find the following eight predominant topics based on the NLP topic finding model, where each topic is identified by their most representative words in terms of the Term Frecuency-Inverse Document Frecuency (TF-IDF). We can find a description of each topic, along with its title and most representative terms, in Table 2. Next, we will discuss the importance of each one of these topics across the entire corpus, and also the temporal evolution of said topics over the last 15 years, to see which ones are new emerging trends, which ones are disappearing, and which ones are constant over time.

Table 2. Main topics discovered by the ldaMallet model

Topic	Description	Main terms
Alternative e-assessment & adoption	Includes papers about alternative ways of e-assessment, such as the use of games, tablets, or smartwatches, among others.	Item, game, word, video, test, tablet, diffusion
Formative assessment & feedback	Papers about formative assessment, which is the use of assessment to provide	Feedback, teacher, student, formative, assessment, study

	feedback to teachers and students over the course of instruction (Boston, 2002).	
Professional development	This topic contains studies which aim to apply digital assessment into workplaces.	Learn, teach, staff, college, development, professional, work
Analytics & assessment	Contains papers aiming to use analytics in order to collect information useful for assessment.	Assessment, activity, competence, process, tool, skill, indicator
Medical assessment	Papers using medical assessment in education.	Digital, grade, clinical, dental, patient, visual, faculty
E-assessment technologies	This topic involves the use of different technologies into the assessment, such as the automatic generation of questions, or the automatic assessment based on existing data.	Question, answer, automatic, automate, generate, test, submission
E-assessment systems	These papers aim to present new digital assessment systems, such as the use of cloud solutions, or the use of IoT.	User, service, system, web, network, design, interface
Trustworthy assessment	Includes papers that aim to make sure that the assessments being performed are trustworthy and secure.	Exam, authentication, security, face, plagiarism, cheat, online

Since LDA is a mixed membership algorithm, each document can be assigned to several topics with a certain weight. Based on these topics discovered, the authors evaluated each document to get its topics associated and calculated the proportion of each topic using the following equation:

$$Proportion\_topic_j = \frac{\sum_{i=1}^{N} weight\_topic_{ij}}{N} * 100$$

Then, the proportion of topic *j* would be the summation of each weight assigned to the topic *j* in each document from *i* to *N*, divided by the number of documents in the corpus (*N*). We can see the topics' distribution across the last ten years in Figure 3. As we can observe, the most frequent topics have been " Analytics & assessment" (19.3%), "Formative assessment & feedback" (15.3%), and "E-assessment technologies" (15.1%), as opposed to "Alternative e-assessment & adoption" (7.5%) and "Medical assessment" (6.5%), which have not been so popular.

Moreover, a similar analysis was conducted in order to analyze papers' keywords. In this case, the researchers used the following equation:

$$Proportion\_keyword_j = \frac{n\_occurrences_j}{total\_occurrences} * 100$$

Then, the proportion of keyword *j* would be the number of occurrences of *j*, divided by the total number of occurrences of all keywords (*total_occurrences*). The top 10 keywords proportion across the last 15 years were calculated. As we see in Figure 4, the most frequent keywords appearing are "e-assessment" (13.6%), "e-learning" (4.0%) and "high education" (1.8%).

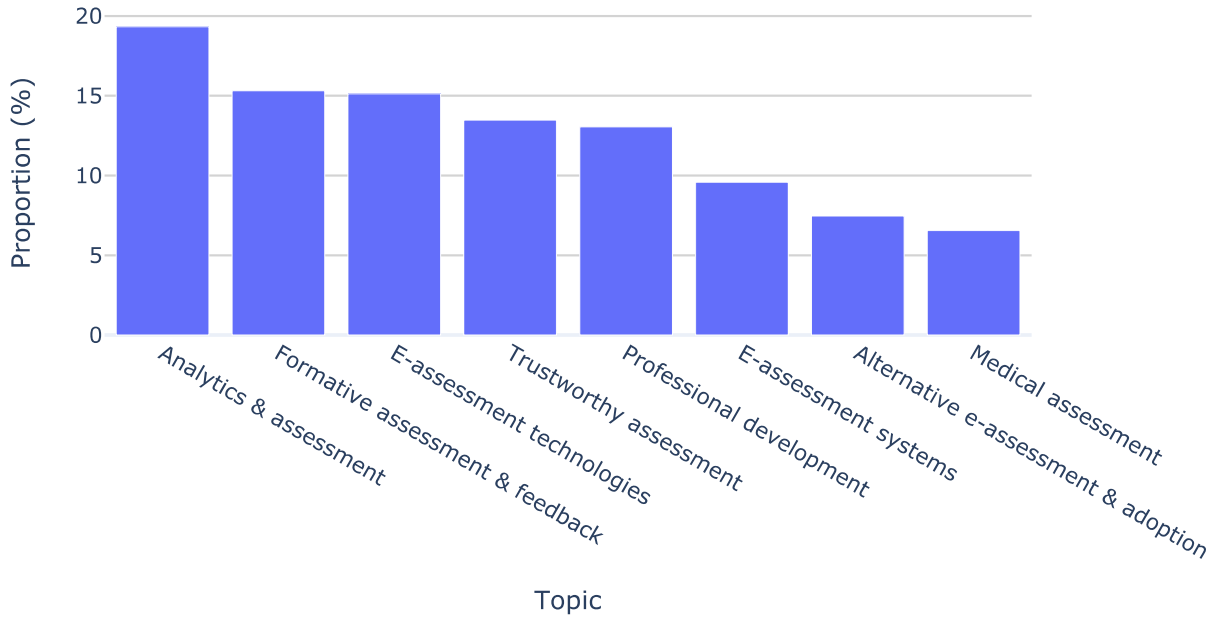

Figure 3. Topic's distribution across all papers

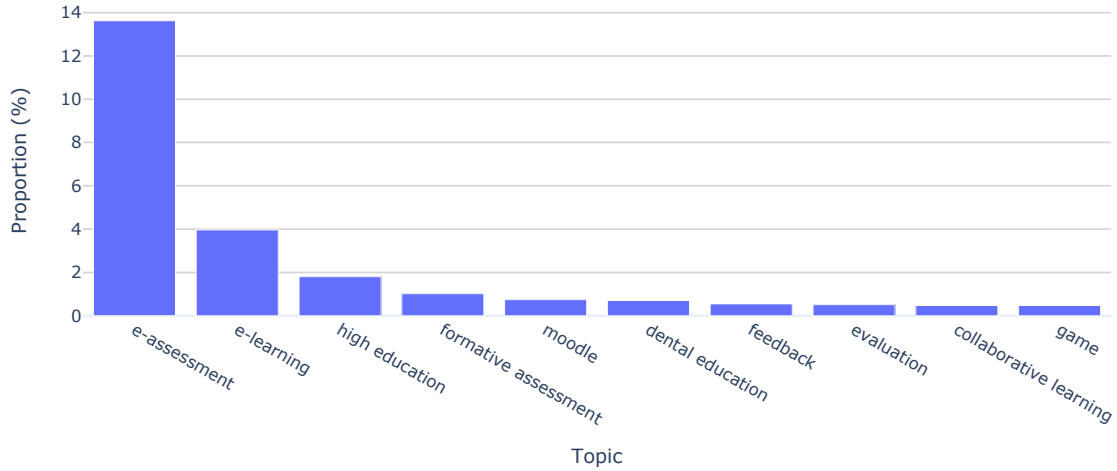

Figure 4. Keyword's distribution across all papers

RQ2. Evolution of Topics Over the Last 15 Years of Research

After presenting an analysis covering the last 15 years of research aggregated, now the authors present a second analysis, taking into account each one of these years individually. The proportion each topic in each year is now calculated using the same formula as above, but this time using only the papers corresponding to each year, instead of all of them.

Figure 5 shows the evolution of each topic's distribution over the years. Some topics, such as "Medical assessment," have gained popularity over the years, going from a frequency of 3.6% in 2007 to a maximum of 12.1% in 2009, and ending with 8.1% in 2021. Other topics like "Formative assessment & feedback" or "E-assessment systems" have maintained their frequency stable over the years, and finally, some other topics, such as "Trustworthy assessment"), have lost some popularity over time: between

years 2009-2015, this topic maintained a proportion above 15%, and then it ended up with a minimum of 5% in 2019.

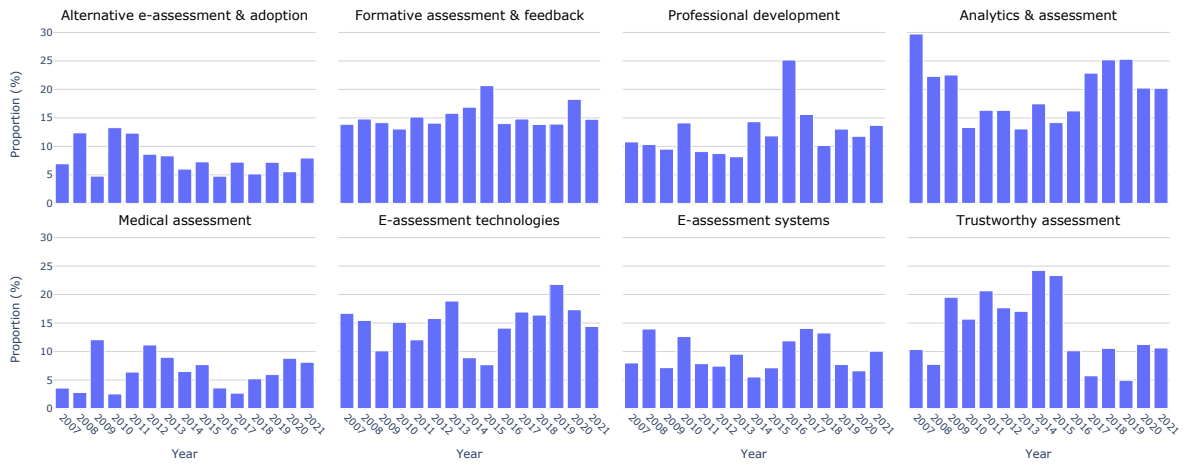

Figure 5. Topic's distribution by year

Analogously, in Figure 6 we have the evolution of each keyword's distribution over time. We see some keywords that have never been as trendy as others, but they keep appearing year after year. This is the case of "high education" which has a similar distribution every year. Moreover, we see that most of the keywords have a tiny proportion compared to "e-assessment" and "e-learning.". In fact, the minimum proportion of "e-assessment" was a 10% in 2015, while the maximum proportion of any of the other keywords (excluding "e-assessment" and "e-learning") is a 3.6%, corresponding to the keyword "formative assessment" in the year 2007. This fact evidences the dominance of this keyword with respect to the others. We also see that not all keywords appear every year: for instance, the keyword "evaluation" did not appear in 2007, 2008, or 2011. Another example is the case of "dental education,", that started appearing in 2012 (did not appear at all before that year) and then it has kept its proportion stable over time.

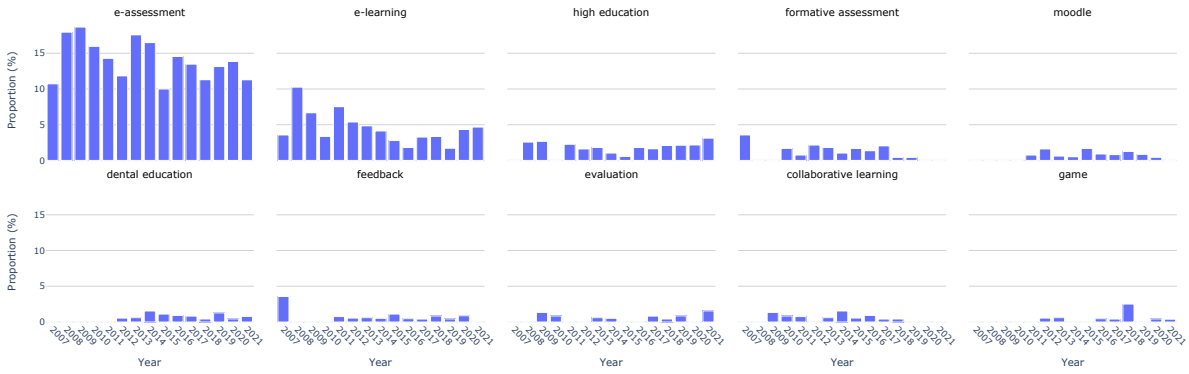

Figure 6. Keyword's distribution by year

RQ3. Communities of Authors and Papers in this Topic Based on a Network Analysis Perspective

As stated in previous sections, the researchers created two different networks: a co-authorship network and a citation network. In Figure 7 we can see the co-authorship network built in Python and then streamed into Gephi, where each node represents an author. Note that, in the plot, only the giant component is shown (a giant component is a connected component of a network that contains a significant proportion of the entire nodes in the network).

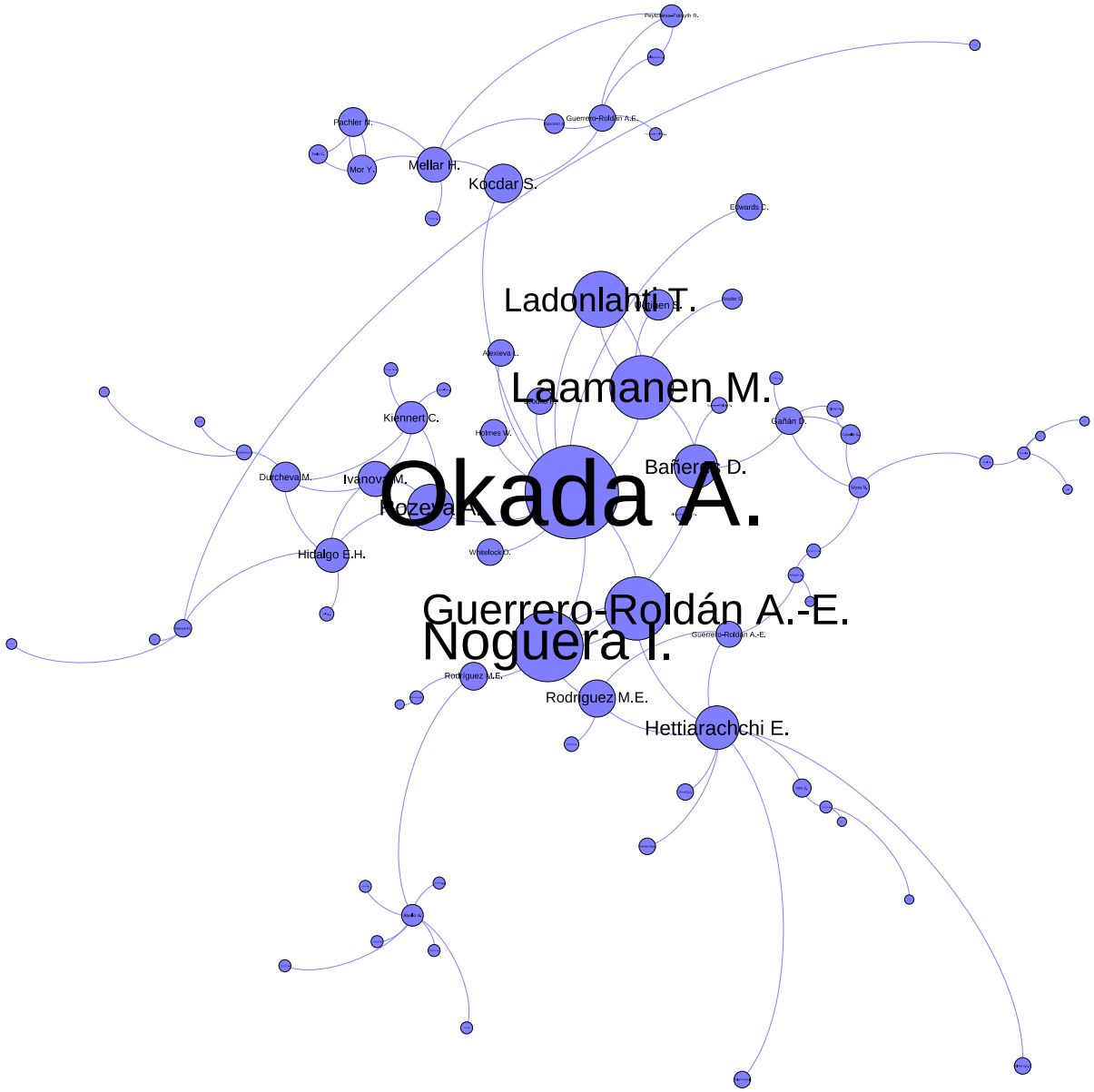

Figure 7. Co-authorship network built on Gephi

There are 1,810 author names across papers (an average of 3.2 authors per paper), and 1,458 of those author names are unique. Furthermore, the top central authors are “Okada A.,” “Noguera I.,” “Laamanen M.,” “Ladonlahti T.,” and “Hettiarachi E.,”, among others. Thus, we can see that these authors are the ones that collaborate with a large proportion of other authors that are also considered as central authors in the co-authorship graph. This study also revealed the most frequent authors in the collection (i.e., the

authors with a larger proportion of published papers): “Gusev M.” (2.06% of the papers), “Ristov S.” (1.86%), “Huertas M.A.” (1.44%), “Hettiarachchi E.” (1.44%), and “Schroeder U.” (1.44%).

This giant component shown in the graph has 76 nodes (5.2% of the authors), meaning that the rest of authors (94.8%) are not collaborating with the authors shown in the giant component, so they represent sub-communities within the digital assessment in education community. In general, we see a fragmented community that is weakly connected (overall density of only 0.034).

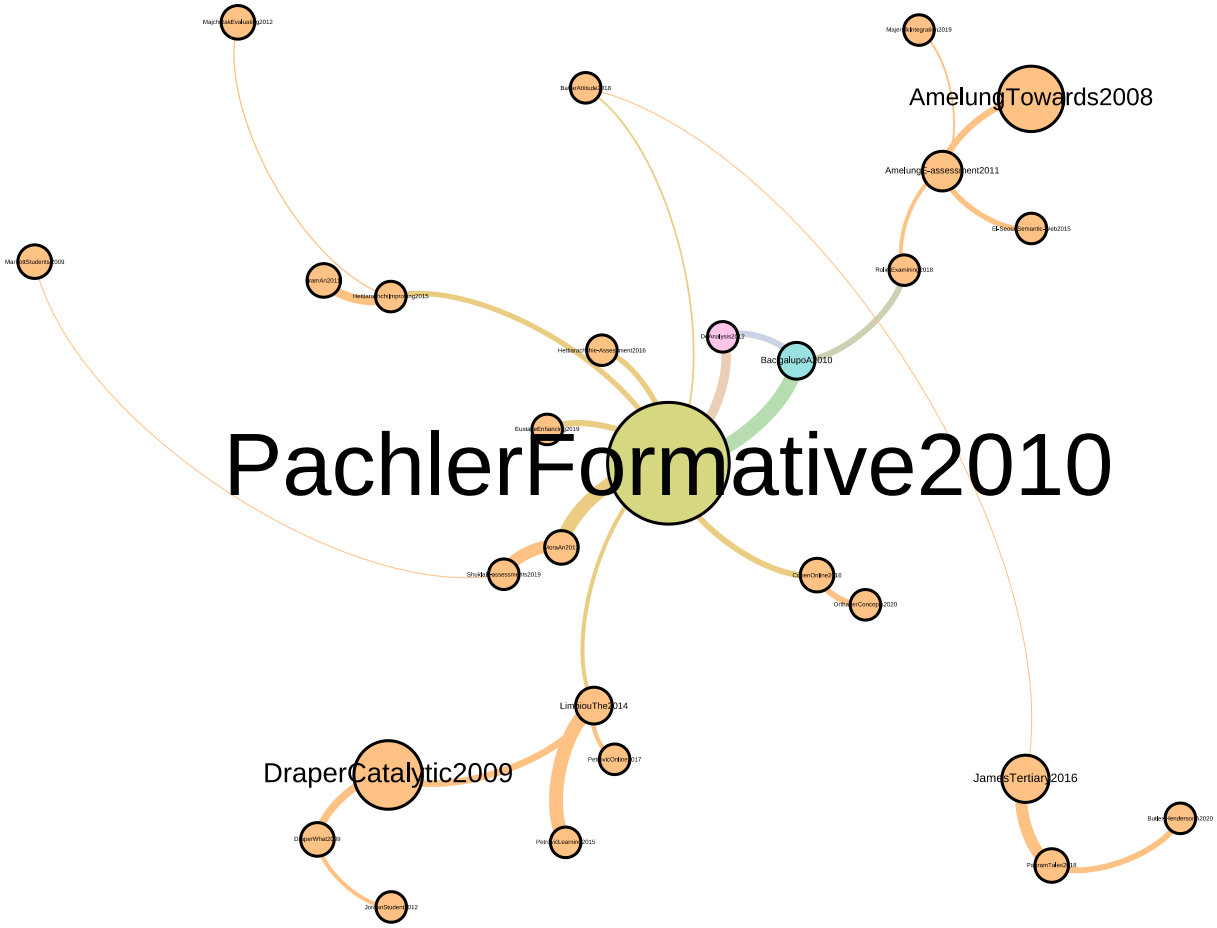

Figure 8. Co-citation network built on Gephi

Moreover, in Figure 8 we can see the citation network built in Python and then streamed into Gephi. Each node represents a paper in the network, and each edge is a citation between the two papers (nodes). Note that, in our plot, only the giant component is shown. This giant component shown in the graph has 28 nodes (5% of the papers), meaning that the rest of papers are not strongly connected (being cited or citating) with the rest of papers in the collection. Again, we see a fragmented community that is weakly connected (overall density of only 0.037).

The algorithm has found 28 references between papers of the collection. The top central papers are “Formative e-assessment” (Pachler et al., 2010), “Towards generic and flexible web services for e-assessment” (Amelung et al., 2008), “Catalytic assessment: understanding how MCQs and EVS can foster deep learning” (Draper, 2009), and “Tertiary students attitudes to invigilated, online summative

examinations" (James, 2016). Looking at the top central papers, we note that one of them was published in 2008, one of them in 2009, one of them in 2010, and another one in 2016. Since these are the papers that have been most cited between them, we note that they are the ones having a major influence between the community being analyzed.

DISCUSSION

This study aimed to answer three different research questions based on two primary sources from papers: the full manuscripts and the metadata (authors and keywords). We can see a summary of the main findings in Table 3. Comparing topics revealed from full manuscripts and keywords, we see that some of the topics revealed are common, such as "dental education," which can be included into "medical assessment," or "game," which can be included into "alternative e-assessment & adoption." However, most of the topics discovered are different, and also their distributions and evolution over the years. This could indicate that topic modeling is revealing hidden topics based on the papers' full text, which may be more realistic that keywords that are arbitrarily chosen by the authors.

Moreover, if we take a look at the network analysis results, we see that being a central author does not imply that your paper will be also a central paper within the collection, as we are measuring separately collaboration between authors and citations between papers. Moreover, we also see that central and most frequent authors are usually different. However, this is not always the case, since "Hettiarachchi E." is a central author but also a top prolific author. In addition, as (Swacha, 2021) previously noted in his work, there are a little small percentage of authors (15%) that have contributed with more than one paper, which also indicates that research in these areas is usually a short-time activity rather than an area of scientific specialization. Further research could explore the transfer of these trends to many other communities, to check the similarities and differences between communities in these aspects.

Previous work has reviewed the state of the e-assessment literature, revealing trends and future challenges in the area. (Stödberg, 2012) conducted a review aiming to summarize some research on e-assessment, providing an overview based on 76 articles from three well-established scientific journals. This study revealed that most of the literature focused on using closed questions, which is normally considered fair and secure for students. However, digital assessment can also be used to measure more complex competencies and skills (such as 21-st century competencies), which are traditionally difficult to measure using conventional forms of assessment (Ruiperez-Valiente et al., 2020). One of the ways to measure those competencies is using new forms of assessment, such as the assessment through games (game-based assessment). In fact, we see that this is a trend in current digital assessment literature, since we see "game" as a central keyword in the collection, and also this same keyword appearing within the topic "alternative e-assessment & adoption" in the LDA model. In this study, the author also revealed some examples of the use of digital assessment, stating that automated e-assessments can be used to save the assessor's time but also to provide immediate feedback to students on their achievements, which has been shown to have a positive effect. In this study, the topic "e-assessment technologies" discovered includes papers covering automated e-assessments and showing that this is one of the most prominent topics in the area. Moreover, a topic that is not usually mentioned but appears in this study is "trustworthy assessment.". The aim of this sub-area is to assure that the assessment being done is secure, and to ensure that students are not cheating while performing the different tasks.

In the literature review, authors identified some previous studies that also tried to combine several techniques to analyze research areas. For example, authors in (Gurcan et al., 2021) performed a bibliometric analysis using a corpus from e-learning field, and then the abstract of each paper to build a LDA model to discover existing trends in the area. Moreover, (Zhang et al., 2021) investigated three computer science education conferences and analyzed authorship and affiliation details from 4,500 publications. In this research, authors combine the use of different techniques in order to enhance the

classic bibliometrics approach and go beyond literature. Combining both full manuscripts and metadata, we can perform quick analysis combining two different sources of information. On the one hand, using authors from metadata and citations from full manuscripts allow us to conduct network analysis and reveal the most central papers and authors. On the other hand, we can perform effective trend identification based on full-text data and keywords.

Table 3. Summary of the main findings

Question	Findings
What are the main topics of digital assessment in education based on keywords?	The main topics found are "e-assessment," "e-learning," "high education," "formative assessment," "moodle," "dental education," "feedback," "evaluation," "collaborative learning," and "game."
What are the main topics of digital assessment in education based on topic modeling?	The main topics found are "analytics & assessment," "formative assessment & feedback," "e-assessment technologies," "trustworthy assessment," "professional development," "e-assessment systems," "alternative e-assessment & adoption," and "medical assessment."
What has been the evolution of such topics over the last 15 years of research based on keywords?	The topic that has gained more popularity has been "e-learning", while other topics such as "e-assessment," "high education," and "formative assessment" have kept a stable popularity over time. The topic "e-assessment" shows a clear dominance with respect to the rest.
What has been the evolution of such topics over the last 15 years of research based on topic modeling?	The topic that has gained more popularity has been "professional development," while other topics such as "e-assessment systems," or "e-assessment technologies" have kept a high popularity during those years. Moreover, other topics like "trustworthy assessment" have been gaining popularity until it reached a peak, and then the interest started decreasing.
What are the communities of authors and papers in this topic based on a network analysis perspective?	The most central authors are "Okada A.," "Noguera I.," "Laamanen M.," "Ladonlahti T.," and "Hettiarachi E.". The most prolific authors are "Gusev M." (2.06% of the papers), "Ristov S." (1.86%), "Huertas M.A." (1.44%), "Hettiarachchi E." (1.44%), and "Schroeder U." (1.44%). The most central papers are "Formative e-assessment" (Pachler et al., 2010), "Towards generic and flexible web services for e-assessment" (Amelung et al., 2008), "Catalytic assessment: understanding how MCQs and EVS can foster deep learning" (Draper, 2009), and "Tertiary students attitudes to invigilated, online summative examinations" (James, 2016).

Although the authors have tried to use a search query to cover a wide range of publications in the area, we might be missing some existing publications that are using other keywords in their papers. However, this study is the first one of its class, and the authors believe that it can be helpful to discover current trends in the area in a very easy way. Although the researchers tried to use the full manuscripts when it was possible, not all the studies were available to be retrieved, and the abstract replaced the full text in these cases. This could introduce some bias, since previous research suggested that using full text data provides better results, especially in a small corpus of documents (Syed & Spruit, 2017).

CONCLUSION, LIMITATIONS AND FUTURE WORK

This study has performed a full bibliometric analysis of the last 15 years within the field of digital assessment, including various techniques coming from bibliometrics, NLP, and network analysis. Our results have indeed shown that this is a lively field with contributions from numerous researchers as well as diverse topics and application purposes, with Table 3 summarizing all these results. Therefore, we find that digital assessment is a promising field for the coming decade now that the systems are more mature.

However, our work is not without limitations. First, we have limitations in terms of the data collection of papers that we have retrieved, which have been based only on Scopus database, that even though it is the one that indexes more papers, we might be missing some. In addition, our search was also limited by the keywords selected. We believe these keywords truly represent the core of the digital assessment topic, but there could be papers missed because authors are using different terms that we did not contemplate.

We encourage future work to further pursue analyzing the main trends of digital assessment in education. Additional research can include bibliometric studies with an extended literature, as well as adding more in-depth scoping reviews that can analyze qualitatively each of the papers. Future researchers performing case studies within this area, should also heavily take into account best practices to implement these systems. Longitudinal meta-review papers of the digital assessment field can help establish guidelines of what works and what does not work and in which context. Therefore, these analyses can help establish the base so that these systems can be adopted in educational systems. Students and teachers can greatly benefit from digital assessment implementation at many different levels.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Allen, J. F. (2003). Natural language processing. In *Encyclopedia of computer science* (pp. 1218–1222).

Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, *8*(1), 34–37.

Amelung, M., Forbrig, P., & Rösner, D. (2008). Towards generic and flexible web services for e-assessment. *Proceedings of the 13th Annual Conference on Innovation and Technology in Computer Science Education*, 219–224.

Amelung, M., Krieger, K., & Rösner, D. (2010). E-Assessment as a Service. *IEEE Transactions on Learning Technologies*, *4*(2), 162–174.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, *228*(4698), 456–462.

Apampa, K. M., Wills, G. B., Argles, D., & Marais, E. (2008). Electronic integrity issues in e-assessment security. *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, 394–395.

Atanassova, I., Bertin, M., & Mayr, P. (2019). Mining scientific papers: NLP-enhanced bibliometrics. *Frontiers in Research Metrics and Analytics*, *4*, 2.

Atenstaedt, R. (2017). Word cloud analysis of the BJGP: 5 years on. *British Journal of General Practice*, *67*(658), 231–232.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Third International AAAI Conference on Weblogs and Social Media*.
Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, *6*(4), e21978.
Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, and Evaluation*, *8*(1), 9.
Broadus, R. N. (1987). Toward a definition of "bibliometrics." *Scientometrics*, *12*(5–6), 373–379.
Burchfield, R. W. (1972). *A supplement to the Oxford English dictionary*.
Chen, E. S., Stetson, P. D., Lussier, Y. A., Markatou, M., Hripcsak, G., & Friedman, C. (2007). Detection of practice pattern trends through Natural Language Processing of clinical narratives and biomedical literature. *AMIA Annual Symposium Proceedings*, *2007*, 120.
Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, *8*, 75264–75278.
Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Computers & Education*, *151*, 103855.
Chopra, A., Prashar, A., & Sain, C. (2013). Natural language processing. *International Journal of Technology Enhancements and Emerging Engineering Research*, *1*(4), 131–134.
Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of Artificial Intelligence*, 603–649.
Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89.
Cooper, I. D. (2015). Bibliometrics basics. *Journal of the Medical Library Association: JMLA*, *103*(4), 217.
Cox, A., Gadd, E., Petersohn, S., & Sbaffi, L. (2019). Competencies for bibliometrics. *Journal of Librarianship and Information Science*, *51*(3), 746–762.
De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. scarecrow press.
Dellos, R. (2015). Kahoot! A digital game resource for learning. *International Journal of Instructional Technology and Distance Learning*, *12*(4), 49–52.
Draper, S. W. (2009). Catalytic assessment: understanding how MCQs and EVS can foster deep learning. *British Journal of Educational Technology*, *40*(2), 285–293.
Fan, J., Gao, Y., Zhao, N., Dai, R., Zhang, H., Feng, X., Shi, G., Tian, J., Chen, C., Hambly, B. D., & others. (2020). Bibliometric analysis on COVID-19: a comparison of research between English and Chinese studies. *Frontiers in Public Health*, *8*, 477.
Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In *Handbook of quantitative science and technology research* (pp. 257–276). Springer.
Gomez, M. J., Ruipérez-Valiente, J. A., Martínez, P. A., & Kim, Y. J. (2021). Applying Learning Analytics to Detect Sequences of Actions and Common Errors in a Geometry Game. *Sensors*, *21*(4), 1025.
GÜRCAN, F., & ÖZYURT, Ö. (2020). Emerging trends and knowledge domains in E-learning researches: Topic modeling analysis with the articles published between 2008-2018. *Journal of Computer and Education Research*, *8*(16), 738–756.
Gurcan, F., Ozyurt, O., & Cagitay, N. E. (2021). Investigation of Emerging Trends in the E-Learning Field Using Latent Dirichlet Allocation. *The International Review of Research in Open and Distributed Learning*, *22*(2), 1–18.
Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *2014 47th Hawaii International Conference on System Sciences*, 1833–1842.
Hevey, D. (2018). Network analysis: a brief overview and tutorial. *Health Psychology and Behavioral Medicine*, *6*(1), 301–328.
Huba, M. E., & Freed, J. E. (2000). *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning*. ERIC.

James, R. (2016). Tertiary student attitudes to invigilated, online summative examinations. *International Journal of Educational Technology in Higher Education*, *13*(1), 1–13.
Jaramillo-Morillo, D., Ruipérez-Valiente, J., Sarasty, M. F., & Ramírez-Gonzalez, G. (2020). Identifying and characterizing students suspected of academic dishonesty in SPOCs for credit through learning analytics. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–18.
Ji, P., Jin, J., Ke, Z. T., & Li, W. (2021). Co-citation and Co-authorship Networks of Statisticians. *Journal of Business \& Economic Statistics*, 1–17.
Kokol, P., Blažun Vošner, H., & Završnik, J. (2021). Application of bibliometrics in medicine: a historical bibliometrics analysis. *Health Information \& Libraries Journal*, *38*(2), 125–138.
Kumar, S. (2015). Co-authorship networks: a review of the literature. *Aslib Journal of Information Management*.
Laumer, S., von Stetten, A., & Eckhardt, A. (2009). E-assessment. *Business \& Information Systems Engineering*, *1*(3), 263–265.
Lu, W., Janssen, J., Milios, E., Japkowicz, N., & Zhang, Y. (2007). Node similarity in the citation graph. *Knowledge and Information Systems*, *11*(1), 105–129.
O'callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645–5657.
Okubo, Y. (1997). *Bibliometric indicators and analysis of research systems: methods and examples*.
Pachler, N., Daly, C., Mor, Y., & Mellar, H. (2010). Formative e-assessment: Practitioner cases. *Computers \& Education*, *54*(3), 715–721.
Palmer, J. A. (2021). *pdftotext*.
Park, K. C., & Lee, C. H. (2019). A study on the research trends for smart city using topic modeling. *Journal of Internet Computing and Services*, *20*(3), 119–128.
Pritchard, A., & others. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, *25*(4), 348–349.
Raja, R., & Nagasubramani, P. C. (2018). Impact of modern technology in education. *Journal of Applied and Advanced Research*, *3*(1), 33–35.
Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment.* Futurelab.
Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408.
Ruiperez-Valiente, J. A., Gaydos, M., Rosenheck, L., Kim, Y. J., & Klopfer, E. (2020). Patterns of Engagement in an Educational Massive Multiplayer Online Game: A Multidimensional View. *IEEE Transactions on Learning Technologies*, *X*(X), 1–14. https://doi.org/10.1109/TLT.2020.2968234
Ruipérez-Valiente, J. A., & Kim, Y. J. (2020). Effects of solo vs. collaborative play in a digital learning game on geometry: Results from a K12 experiment. *Computers & Education*, *159*, 104008.
Saini, S., Rambli, D. R. A., Sulaiman, S., Zakaria, M. N., & Shukri, S. R. M. (2012). A low-cost game framework for a home-based stroke rehabilitation system. *2012 International Conference on Computer & Information Science (ICCIS)*, *1*, 55–60.
Sclater, N. (2008). Web 2.0, personal learning environments, and the future of learning management systems. *Research Bulletin*, *13*(13), 1–13.
Serrat, O. (2017). Social network analysis. In *Knowledge solutions* (pp. 39–43). Springer.
Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.
Stödberg, U. (2012). A research review of e-assessment. *Assessment \& Evaluation in Higher Education*, *37*(5), 591–604.
Swacha, J. (2021). State of research on gamification in education: A bibliometric survey. *Education Sciences*, *11*(2), 69.

Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174.
Thooriqoh, H. A., Fatichah, C., Purwitasari, D., & others. (2021). Topic Detection in Sentiment Analysis of Twitter Texts for Understanding The COVID-19 Effect in Local Economic Activities. *2021 13th International Conference on Information \& Communication Technology and System (ICTS)*, 354–359.
Whitelock, D. (2009). e-assessment: developing new dialogues for the digital age. *British Journal of Educational Technology*, *40*(2), 199–202.
Wolfram, D. (2016). Bibliometrics, information retrieval and natural language processing: natural synergies to support digital library research. *Proceedings of the Joint Workshop on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, 6–13.
Yao, L., Mimno, D., & McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 937–946.
Zhang, J., Luxton-Reilly, A., Denny, P., & Whalley, J. (2021). Scientific Collaboration Network Analysis for Computing Education Conferences. *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, 582–588.
Zhu, Z.-T., Yu, M.-H., & Riezebos, P. (2016). A research framework of smart education. *Smart Learning Environments*, *3*(1), 1–17.

ADDITIONAL READING

Allen, J. F. (2003). Natural language processing. In *Encyclopedia of computer science* (pp. 1218–1222).
Mikropoulos, T. A., Sampson, D. G., Nikopoulos, A., & Pintelas, P. (2014). The evolution of Educational Technology based on a bibliometric study. In *Research on e-Learning and ICT in Education* (pp. 15-24). Springer, New York, NY.
Atanassova, I., Bertin, M., & Mayr, P. (2019). Mining scientific papers: NLP-enhanced bibliometrics. *Frontiers in Research Metrics and Analytics*, *4*, 2.
Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109–127.
Eyal, L. (2012). Digital assessment literacy—The core role of the teacher in a digital environment. *Journal of Educational Technology & Society*, *15*(2), 37-49.
Jordan, S. (2013). E-assessment: Past, present and future. *New Directions*, *9*(1), 87-106.
Alruwais, N., Wills, G., & Wald, M. (2018). Advantages and challenges of using e-assessment. *International Journal of Information and Education Technology*, *8*(1), 34-37.
Amelung, M., Krieger, K., & Rösner, D. (2010). E-Assessment as a Service. *IEEE Transactions on Learning Technologies*, *4*(2), 162-174.

KEY TERMS AND DEFINITIONS

Bibliometrics: Bibliometrics is the use of statistical methods to analyze books, articles and other publications.

Natural Language Processing: branch of computer science which aims to understand and produce language the same way as human beings can.

Lemmatization: the algorithmic process of determining the lemma of a word based on its intended meaning.

Topic modeling: a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents.

Metadata: a set of data that describes and gives information about other data.

Digital assessment: the presentation of evidence, for judging student achievement, obtained through the use of computer technology.

Text analytics: process of drawing meaning out of written communication.

Network analysis: set of techniques which allow to depict relations among actors and to analyze the social structures emerging from those relations.

Corpus: a collection or body of knowledge or evidence.