International Journal of Innovation in Education

ISSN online: 1755-1528 - ISSN print: 1755-151X
https://www.inderscience.com/ijiie

Assessment analysis: methods and implementation options for multiple-choice exams

Annastiina Rintala

DOI: 10.1504/IJIIE.2023.10051577

Article History:

Received:	11 November 2021
Accepted:	08 June 2022
Published online:	23 January 2023

Assessment analysis: methods and implementation options for multiple-choice exams

Annastiina Rintala

School of Engineering Science,
LUT University,
Lappeenranta, FI-53850, Finland
Email: annastiina.rintala@lut.fi

Abstract: Automatic assessment can reduce teacher workload and offer flexibility for students, but if the teacher does not assess the exams manually, the teacher's view of the students' competence and exam-related behaviours will be meagre. This drawback can be mitigated through appropriate analytics. To support the design of an analytics module for an electronic assessment system, this paper investigates what kinds of analyses are useful for multiple-choice exams and how the analysis can be implemented. Three types of analysis were found useful: 1) descriptive statistics of exam answers; 2) analysis of errors in answers; and 3) analysis of students' exam-taking behaviours. Though these analyses are to some extent generalisable, analysis needs vary, for example, by time, exam type and user. Therefore, it is suggested that to enable user-specific analyses in a resource-efficient manner, assessment software providers should facilitate access to assessment data in a structured format.

Keywords: multiple-choice exams; automatic assessment; electronic assessment; analytics tools; design science; electronic learning; learning analytics; assessment analytics; innovation; education.

Reference to this paper should be made as follows: Rintala, A. (2023) 'Assessment analysis: methods and implementation options for multiple-choice exams', *Int. J. Innovation in Education*, Vol. 8, No. 1, pp.20–39.

Biographical notes: Annastiina Rintala received her Master's and PhD in Lappeenranta University of Technology, Finland. Her expertise area is supply chain and operations management. She currently works as a Lecturer in School of Engineering Science at the Lappeenranta University of Technology, Finland.

1 Introduction

One of the potentially time-consuming aspects of teaching is the assessment of student exams and exercises. Though assessments are an integral part of teaching, the time spent assessing one student's work does not usually benefit the other students. The use of electronic assessments with automatic evaluation is especially relevant for large student groups. Online examination resources can also provide many benefits to students, including self-paced learning, access to resources without time constraints and instant feedback. Previous studies also imply that students find the use of e-assessment more engaging and less stressful than traditional assessment (Dermo, 2009; Holmes, 2015;

McCann, 2010). However, if the teacher does not assess the students' work manually, the natural feedback loop between the teacher and students will be broken, and there is a risk that the teacher's view of the students' competence will be short-sighted. This loss can be mitigated through appropriate analytics.

Learning analytics is a relatively new discipline that addresses many points of interest, ranging from student engagement to predictive modelling. One of the leading themes in previous research objectives is the use of analytics in the redesign of learning activities (Mangaroska and Giannakos, 2019). However, relatively few studies focus on assessment analysis (Ellis, 2013; Nouira et al., 2019; Saqr, 2017). The development of learning analytics practices also lags behind the technological possibilities. Some electronic learning systems do not lend themselves to any type of analytics, as the example discussed in this paper shows. For the electronic assessment system that is widely used in Finnish universities, the development of an analytics module is in the initial phase. Prior to technological implementation, it is important to be aware of the needs of users, such as what kind of information should be obtained and what kind of decisions should be supported with assessment analytics. To support the design of an analytics module of a rather simple electronic assessment system, this study focuses on the following research questions:

- What kind of analytics are suitable for multiple-choice exams?
- How can the analysis be implemented in practice?

Taking a design science approach and focusing on a single case, this paper investigates the need for assessment analytics and options for implementation from the teachers' point of view.

This paper is organised as follows. In Section 2, the related literature of learning analytics and automatic assessment of student work is reviewed. In Section 3, the research design is presented. Section 4 presents and discusses the results, and in Section 5 some concluding remarks are offered.

2 Learning analytics/educational data mining

Learning analytics is an interdisciplinary field embracing methods and approaches from various disciplines. It involves machine learning, artificial intelligence, information retrieval, statistics, and visualisation. The Society for Learning Analytics Research defines learning analytics as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Mangaroska and Giannakos, 2019).

The practice of learning analytics has evolved around the idea of harnessing the power of digital technologies to collect traces that users leave behind in order to understand activities and behaviours associated with users' learning (Siemens, 2013). Recently, learning analytics has highlighted the gradual shift from a technological towards an educational perspective, despite its roots in business intelligence, recommender systems, and educational data mining (Ferguson, 2012).

Although multiple e-learning environments store user data automatically, exploitation of the data for learning and teaching is still very limited. These educational datasets offer unused opportunities for the evaluation of learning theories, learner feedback and

support, early warning systems, learning technology, and the development of future learning applications. Thus, the importance of learning analytics has been increasingly recognised by governments, educators, funding agencies, research institutes, and software providers (Greller and Drachsler, 2012).

Learning analytics may concern several stakeholder groups, primarily students, teachers and institutions. For students, learning analytics can offer different types of learning processes and reflection visualisations. For example, Dipace et al. (2019) present an analytics dashboard for a massive open online course (MOOC). De Barba et al. (2016) examine how motivation and participation predicted performance in a MOOC. Similarly, Pursel et al. (2016) provide early insights into variables such as interaction data that show some relationship to MOOC completion. In principle, these variables can be used in predicting student completion and devising methods to keep students engaged. Kitto et al. (2017) present an approach in which machine learning is used to classify the behaviour patterns of students during learning activities to facilitate students' self-reflection. Chatti et al. (2016) present a collaborative annotation tool for video-based learning that allows users to annotate sections of interest, reply to each other's annotations and locate the most viewed and annotated parts of the video. Charleer et al. (2018) present a dashboard that visualises grade data and provides an overview of students' study progress, thereby supporting the adviser-student dialogue in advising sessions.

Teacher-oriented learning analytics studies are concerned with monitoring student performance and behaviour to identify developmental targets and design interventions into the course and learning activities. For example, Melero et al. (2015) present an application for mobile learning that offers visualisations of group activity enactment (time used to answer questions, scores obtained etc.) that helps in the evaluation of the overall design of the activity route and identification of the most problematic questions. Florian-Gaviria et al. (2013) present a software suite that provides visualisations of study performance data on different social planes: single students, collaborative groups and whole classes. Pardo et al. (2015) present a case study in which digital footprints of students are coupled with questionnaire data concerning students' approaches to learning. Berland et al. (2015) present a tool for supporting teachers' pairing decisions with real-time analyses of students' programming progressions. Martinez-Maldonado et al. (2015) present a conceptual model for capture, analysis and presentation of tabletop interaction data to provide understanding of the collaborative learning process. Rodríguez-Triana et al. (2015) present a model for monitoring students' work in computer-supported collaborative learning settings, and Tervakari et al. (2014) present a similar model in social media–enhanced learning environments. For institutions, learning analytics can offer a tool to monitor the performance of students and identify students at risk of underperforming (e.g., Bainbridge et al., 2015; Choi et al., 2018), though Lawson et al. (2016) note that ethical dilemmas related to the interventions in such cases may explain why there is relatively little research on this topic.

In parallel with the concept of learning analytics is the concept of educational data mining. While having basically the same objectives, the latter emphasises the data accumulated by different educational information systems. According to Romero and Ventura (2010), the number of studies concerning educational data mining has grown rapidly since 2000, with the most common tasks or categories being 'providing feedback for supporting instructors', 'recommendations for students', 'predicting student performance' and 'analysis and visualisation of data', followed by 'grouping students', 'student modelling', 'detecting undesirable student behaviours', 'social network

analysis', 'planning and scheduling', 'developing a concept map' and 'construction courseware'. A newer review by Rodrigues et al. (2018) highlights that recent trends in educational data mining studies are:

a reducing the distance between teachers and students

b recommending teaching media more didactic and effective

c identifying similar characteristics of learning or behavioural actions of the student

d improving the process of personalised learning.

The range of topics is wide, and even the studies within the categories differ greatly in terms of what data has been analysed and how.

In general, the goal of learning analytics is to refine the information gained from the learning process to support its development, but a more traditional and commonly used means of achieving the same goal is to gather student feedback. Student evaluations of teaching (SET) are a commonly applied tool in higher education to determine course and teacher effectiveness; over 80% of teachers at European universities report using SET as diagnostic feedback (Nederhand et al., 2022). However, it has been generally noticed that student feedback surveys suffer from low response rates (e.g., Leckey and Neill, 2001; Nair and Adams, 2009; Nederhand et al., 2022). In addition, students are reluctant to answer open-ended questions; for example, in a study by Hujala et al. (2020), only 27–45% of students responding to feedback questionnaires provided answers to open-ended questions, and the answers were often short. When it comes to the quality of student comments, the answers do not necessarily relate to the question asked (Alhija and Fresko, 2009), and feedback can be unconstructive (Ernst, 2014) or even abusive (Tucker, 2014).

A few studies have been devoted to analysing student comments in SET questionnaires. According to Alhija and Fresko (2009), students' comments tend to be general rather than specific, and the authors address three major domains and eight primary content areas in students' comments: the course (content, assignments and general), the instructor (teaching style, personal traits and general) and the context of instruction (scheduling issues and student composition), the last of which is not related to teaching but may negatively affect course feedback. Brockx et al. (2012) studied the topics in positive and negative comments and found that most of the students' positive comments deal with the combination of theory and practice, whereas the relevance and interestingness of the course is the most common topic in negative comments. Stewart (2015) focused on the language that students use and writes that positive comments most typically refer to how supportive and caring staff has been or how approachable staff were, and this praise is often directly targeted at lecturers, whereas criticisms (e.g., concerning teaching skills) objectified teaching as an act. Hujala et al. (2020) found that the most common themes appearing in verbal student feedback were good content, dissatisfaction with personnel or the course, workload, stress and good teaching methods. Lowenthal et al. (2015) focused on online courses and found that students rate online courses lower than face-to-face courses, but they warn that the results should not be generalised to the larger public.

In sum, students' comments are above all about satisfaction or dissatisfaction (which is affected by many things), but they typically do not refer to learning outcomes. As SET are often the sole tool in higher education to determine course and teacher effectiveness,

assessment analytics could play a role in diversifying teaching feedback, not only to support teacher self-reflection but possibly also to provide visibility into learning outcomes at the program level.

2.1 Assessment analytics

Teacher capabilities to plan and implement quality assessment tasks, to interpret assessment outcomes and to engage students in assessment of their own learning have been subjects of considerable research. These capabilities are often referred to as assessment literacy (AL), a concept introduced by Stiggins (1991), who defines AL as "a basic understanding of educational assessment and related skills to apply such knowledge to various measures of student achievement". Assessment literacy is a noteworthy issue in teacher education; for example, Smith et al. (2014) note that student-teachers' thinking and beliefs about assessment are often dominated by their prior experience of formal summative assessment, and moreover, according to Wiggins and McTighe (2007), many teachers imitate the teacher practices they have experienced as a student. This behaviour may limit the diversity of assessment practices, particularly the use of formative assessment, also known as assessment for learning (Kaya et al., 2021).

Xu and Brown (2016) present a conceptual model for AL consisting of six components: knowledge base, teacher conceptions of assessment, institutional and socio-cultural contexts, teacher assessment literacy in practice, teacher learning, and teacher identity (re)construction as assessors. The knowledge base, which is the basis of all other components, consists of

1 disciplinary knowledge and pedagogical content knowledge
2 knowledge of assessment purposes, content and methods
3 knowledge of grading
4 knowledge of feedback
5 knowledge of assessment interpretation and communication
6 knowledge of student involvement in assessment.

As summarised by McMillan (2003), teachers' assessment decision-making is a process by which teachers balance the demands of external factors and constraints with their own beliefs and values. Xu and Brown (2016) claim that teacher assessment literacy in practice (TALiP) consists of various compromises that teachers make. Hence, TALiP is defined as a dynamic, complex entity combining teachers' assessment knowledge, their conceptions of assessment and their responses to the external contexts embedded with actual constraints and affordances in the environment.

Some studies have elevated assessment interpretation to a more central role in assessment literacy. Eyal (2012) states that teachers are measured by their students' performances in tests, but they usually ignore this indicator as a measure of the quality of their own instruction. For example, Popham (2004) discusses assessment literacy as the ability of the teacher to significantly delve into and interpret test results. According to Chan (2018), many teachers expect technology to improve the time efficiency and accuracy of the assessment process, not only in the collection of responses and scoring but also in the analysis and distribution of assessment results. However, practical

implementation requires the development of learning analytics models and methods for this purpose.

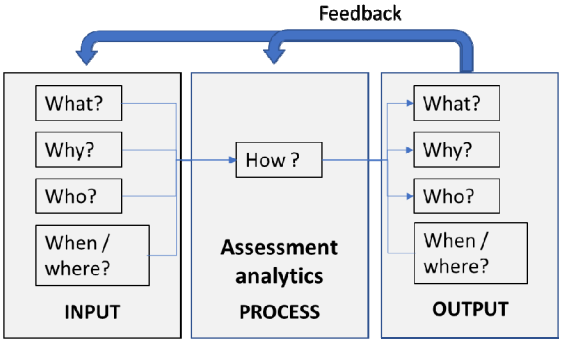
The part of learning analytics that focuses on assessment data can be called assessment analytics, and by itself it is an emerging research field. According to Papamitsiou and Economides (2016), the main objective of assessment analytics is to efficiently and effectively support the improvement of the assessment process. From the teachers' perspective, assessment analytics can be used to facilitate the estimation of students' performances and improve the detection of students at risk, misconceptions and gaps in students' understanding, and guessing or cheating by the students.

The landscape in the domain of assessment analytics is diverse. Some studies focus on the examination results. For example, Badri et al. (2020) examined the factors influencing students' math test scores, and Omorog (2020) analysed MySQL error logs to conclude which tasks pose the most difficulties for students. Another area of focus is the development of assessment practices, for example, the construction of sophisticated measures of assessment (Wilson et al., 2011; Worsley and Blikstein, 2013) and selection of the most appropriate next task during adaptive testing (Barla et al., 2010). Bertheussen (2015) examined the impact of assessment innovations (including spreadsheets within assignments and the final exam) on students' motivation to use spreadsheets in their daily learning activities.

One branch of studies focuses on assessing student behaviour during learning assignments rather than on assignments or results. For example, Pattanasri et al. (2012) use a machine learning approach to assess the comprehensibility of presentation slides on the basis of students' slide-level comprehension reports. Holmes et al. (2018) present a comprehension classification system based on an artificial neural network (ANN); the system detects learner comprehension of on-screen information during e-learning activities, and data gathered by a camera that faces the user and scores given to answers to questions are used to train an ANN. Liu et al. (2015) present a tool that records the intermediate stages of document development and uses this data to measure and visualise learners' engagement in writing assignments. A few studies concern students' mood and emotions during assessment (Chen and Chen, 2009; Moridis and Economides, 2009).

To clarify the concept of assessment analytics, Papamitsiou and Economides (2016) present a general framework for assessment analytics (Figure 1). The input to the assessment analytics system is contextual information related to what should be tracked and assessed, why the assessment is necessary, who the subject and receiver of the assessment (learner or teacher) are, and when and where the assessment takes place. The assessment analytics process itself mostly concerns issues related to how it is applied and which parameters are being exploited during the procedure (methods, resources, instruments, limitations and boundaries, pedagogy and instructional design). The output of the assessment analytics system includes what should be done next, why it should be done and who is the final receiver of the results (e.g., institutions, software developers – beyond students and teachers). Finally, the feedback is related to the delivery of the results to the recipient of the assessment result so that the original context can be changed.

Figure 1 A general framework for assessment analytics (see online version for colours)



Source: Adapted from Papamitsiou and Economides (2016)

However, Ellis (2013) notes that within learning analytics studies, there is relatively little research on assessment analytics and assumes that this 'blind spot' around assessment analytics is most likely because, until relatively recently, the possibility of collecting and collating assessment data at a level of precision that is meaningful and useful has simply been unthinkable. The most significant challenge facing learning analytics is operationalisation – that is, what data should be collected and what should be done with it. Also, Saqr (2017) claims that the potential of assessment data is huge but still underexplored and largely underdeveloped. Nouira et al. (2019) note that the majority of learning analytics models focus on learning traces in general but the models do not take into consideration other types of educational traces, such as assessment traces; they claim that this is due to the lack of specifications in this context and the only learning analytics data model that supports assessment analytics is the xAPI data model, as it contains an optional attribute called *result* that can model assessment results.

Nouira et al. (2019) present an enhanced ontological xAPI data model that supports assessment analytics. The proposed assessment context data are assessment form, assessment type, assessment technique, assessment environment and assessment session. The assessment statement describes the assessment behaviour and experience of the learner during the assessment process. The core of the assessment statement is the assessment result, which contains several types of data: score, rate of completion, duration, attempts, answered questions, unanswered questions, correctly answered questions and wrongly answered questions. The model proposed by Nouira et al. (2019) offers a good basis for collecting assessment data, but it is possible to collect even richer types of data from the answers themselves, depending on the type of task being assessed. More case studies are needed to enhance understanding of the desired outputs of assessment analytics in different contexts.

2.2 Developments in automatic assessment

Recently, automatic assessment of different types of tasks has evolved, and this development increases the need for assessment analytics. Webb et al. (2013) point out that technology enables assessments to be designed with a more user-centred approach, particularly to meet learners' needs, and allows assessments to simultaneously address

assessment *for* learning and assessment *of* learning. In principle, automatic assessment holds the potential for the student to receive more timely feedback to support their learning.

Most state-of-the-art e-learning platforms (e.g., Moodle) provide support for online evaluations and assessments among their features, and in many cases, this support includes automatic exam correction. A good example is multiple-choice tests, where students must choose the correct answer among several possibilities. This type of testing, including variations where students have to match answers from two or more groups or questions requiring short answers, has become very popular because such exams lend themselves to automatic correction (Llamas-Nistal et al., 2013).

Recently, there has been progress in the automatic assessment of short answers (Burrows et al., 2015; McDonald et al., 2017; Siddiqi et al., 2010) – that is, analysing student responses and automatically assigning them to meaningful categories to support formative feedback to students, though it is still a work in progress (Dzikovska et al., 2016) and may be unsuitable for analysing deeper responses (i.e., responses that are indicative of a deep approach to learning (Biggs and Tang, 2011)). Recent examples include studies that use statistical text-mining techniques such as topic modelling (Basu et al., 2013) or k-means clustering (Zesch et al., 2015) or rule-based techniques such as inferencing clustering rules from hand-coded sets of student responses (Willis, 2015). For example, Liu et al. (2017) present a machine learning approach to provide automated feedback of English essays regarding aspects of writing such as grammar, spelling, sentence diversity, and structure.

In other areas, there are improved possibilities for automatic assessment of more complex student works, in which the software scoring has been found to be as consistent as human scoring. For example, Livne et al. (2007) presents a parsing system that produces partial credit scoring of students' constructed responses to mathematical questions, while Baneres et al. (2014) describe a system that provides automated feedback of digital circuit designs. There is a multitude of systems for automatic testing of programming assignments, which are used to supplement teaching in the field of computer science (Amelung et al., 2011; Conejo et al., 2019). Sanna et al. (2012) present a computer vision and image analysis–based tool for automatic assessment of 3D-modelling exams. Other authors (Lamberti et al., 2014; Paravati et al., 2017) present approaches for automatic assessment of 3D animation assignments.

These studies demonstrate that automatic assessment of complex student work is becoming more feasible. As the automatic assessment of assignments becomes increasingly common, it is important that teachers are simultaneously provided with a feedback loop to help interpret students' learning outcomes and adapt their teaching design accordingly. Even though this paper focuses on analysing multiple-choice exam results, which is technically simple, the general feedback loop and its technical implementation options are potentially applicable to more complex automatic assessment cases.

3 Research design

The strategic methodological approach in this study is design science, in which the main objective is to develop tentative solutions to problems (Van Aken, 2004). Design science research (DSR) is conceptualised as a research strategy aimed at gathering knowledge

that can be used in an instrumental way to design and implement actions, processes, or systems to achieve desired outcomes in practice. DSR's core research products are well-tested, well-understood, and well-documented innovative generic designs, dealing with authentic field problems or opportunities. According to van Aken et al. (2016), DSR is the main research strategy in engineering and medicine and is gaining ground in areas such as information systems (Hevner et al., 2004). This approach also seems to be common in learning analytics studies, but different names are used for this research strategy, such as 'design-based research (DBR)' (Rodríguez-Triana et al., 2015).

In practice, the DSR paradigm is about describing and answering real-life problems, which naturally leads to case studies. According to Yin (2014), the case study method is appropriate when one seeks to examine novel and complex phenomena, allowing the phenomenon of interest to be investigated in all its richness and in its natural context. Case studies can be used for answering questions of 'why' or 'how' (Yin, 2014) and are sufficiently flexible to generate holistic knowledge through combining various sources and types of data within the same study (Dubois and Gibbert, 2010; Eriksson and Kovalainen, 2016). This study can be defined as an intensive case study, as the aim is to understand the situation in depth by providing a holistic and contextualised description and interpretation (Eriksson and Kovalainen, 2016).

3.1 Description of the empirical setting

3.1.1 EXAM electronic exam system

The EXAM system stems from the mission of ten Finnish universities to jointly define and produce a completely new exam system that meets the requirements of a digitalising study environment and utilises its potential. The cooperation was started in 2014, and, as of 2020, this platform is in use in 27 universities (almost all in Finland). As exam facilities are shared among universities, it has increased students' opportunities for place-independent studying.

The basic process of the EXAM system is as follows:

1 the teacher creates an exam in the system, adds the desired questions, and publishes it for the desired time

2 the student registers for the exam

3 the student reserves the time and place for the exam space of his or her choice

4 the student enters the exam room with a personal key

5 the student logs in to the system and completes the exam

6 the exam is evaluated by the teacher, and the student receives information about the grading by email.

The exam rooms have computers on which the exam is conducted, as well as a video and audio recording monitoring system. For students, the biggest benefit of electronic exams is the opportunity to take the exam at a time that suits them.

The EXAM system currently does not provide any analytic tools. However, the system is developed with public funding, so the need for analytic tools is under investigation. The approach adopted in development efforts has been such that some example metrics are defined and user preferences for utilising different metrics are

addressed with a survey. The metrics suggested by this survey are the following (the author has not been involved in this choice of metrics): grade distribution, average score/grade, time spent on the exam (minimum, maximum, average), time spent on the exam compared to the grade (graph), number of graded exams as a function of exam period, 'passed' rate per question, and questions with minimum and maximum scores compared to maximum (%).

While it is possible to offer many kinds of metrics, it is unclear without piloting if they provide valuable information to users. We claim that instead of offering a set of general metrics and mapping user opinions of their potential, it is possible to adopt a reverse approach in analytics tool development: first investigate what kind of information is needed, and then determine how this information can be extracted from the available data.

3.1.2 Description of the specific exam

There are a few reasons for focusing on a single course. For privacy reasons, access to the data collected by the system is restricted. Only the system administration and the exam author have access to the exam results. Secondly, there are still very few courses where the exams are automatically evaluated. The course in question is an introduction to supply chain management. The course is included in the Bachelor of Science degree, and the teaching language is Finnish. After completing the course, the learning goal is that students will be able to define the basic concepts of supply chain management, analyse inventory and design methods for inventory management, and roughly evaluate the cost effects of logistical decisions.

The course is offered every year, and approximately 180 students are enrolled in the course. The course assessment consists of several assignments that the students complete individually online. Since a student's identity cannot be confirmed online, it is possible for someone to complete the assignments on behalf of the student. The EXAM system is used to verify that the student can perform independently.

All questions are multiple choice; some have just one correct answer while others have multiple correct answers. A question bank of 48 questions is available, from which 20 questions are allotted to each exam. As a result, each exam is different, which is supposed to reduce potential cheating. The exam is assessed automatically, but the system does require teachers to log in to confirm the assessments. The students can book a time for a retake as soon as the assessment is confirmed, and the best score remains valid. The number of retakes is not restricted but can be controlled with an assessment confirmation schedule.

After the course, students receive a survey, where they can provide their feedback on the pros and cons of the course in an open-ended format. Between 2017 and 2020, 38 comments concerned the electronic exam. Ten of the comments concerned the exam in general – whether if the exam was easy or difficult, or if it was 'good' or not. Fifteen of the comments were positive comments on the way the exam was organised. In short comments, simply 'EXAM' was mentioned as the best part of the course, such as "EXAM was a good thing". More detailed comments mentioned the following: "…you can keep learning after making mistakes", "one learned more while doing the exams", "you could choose the time yourself". Five of the comments were neutral or unclear

comments about the way the exam was organised, such as "multiple-choice exam reduces workload", "the exam could have been also in Moodle", and "permitting retakes was an interesting choice".

Eight of the comments were critical. Four of the critical comments suggested that the exam did not enhance learning or that some other format would encourage better studying: "A traditional paper exam with open-ended questions would force students to really study and understand what they are reading". In four other critical comments, a worry was expressed that some students retake the exam too frequently: "It feels stupid that folks continuously took the exam without studying", "…some people took it probably 10 times to get a good grade", "Does it develop anyone's learning when they take the exam twice a day and 10 times altogether? If there was one exam, everyone would read the required amount at once and get the score they deserve. This, I think, is unfair to those who work throughout the course". In addition, the student union filed a complaint about the assessment procedure as unfair, since traditionally exam retakes are limited to three attempts (after which permission to retake needs to be applied literally), but in the EXAM system, the number of retakes is not restricted.

Thus, it appears that the biggest worry concerning the exam is that it encourages superficial learning, as the exam can be repeated multiple times to receive a good score. In addition, some students directly expressed a fear that people repeating the exam would overload the capacity of the exam space, blocking other users. It is possible to analyse from exam data whether this concern is valid.

3.2 Data collection and analysis

The data for this study were collected from the EXAM platform. As this learning platform does not support downloading data directly from the system, a web scraping method was used. Web scraping, also called 'parsing' or "web harvesting", means extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol or through a web browser. While web scraping can be done manually, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data are gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

In this study, web scraping was conducted using Web Scraper, an extension of Google Chrome. Data rows extracted from the EXAM system included the following information: student ID, start and end time of the exam, questions allotted, students' score per question, and students' answer choices. Data collection with web-scraping was done for the purposes of this study only, web-scraping is not suggested as a tool for everyday use.

Data analysis was performed using a spreadsheet application. The dataset consisted of assessment data from 236 exams completed by 139 students. Prototyping the analysis tools with spreadsheet was done for research purposes, spreadsheets are not suggested as analysis tools for everyday use. The results should be understood as illustrative examples of what descriptive statistics from assessments can look like. The analyses can be categorised in three ways, depending on objective:

1 *Basic descriptive statistics for each question*: Score distribution per question was assessed to determine which questions were easy or difficult for students.

2 *Error analyses for answers*: The incidence of incorrect answer choices was examined to find out if there were any systematic patterns in incorrect answers that could result from common misunderstandings or misleading questions.

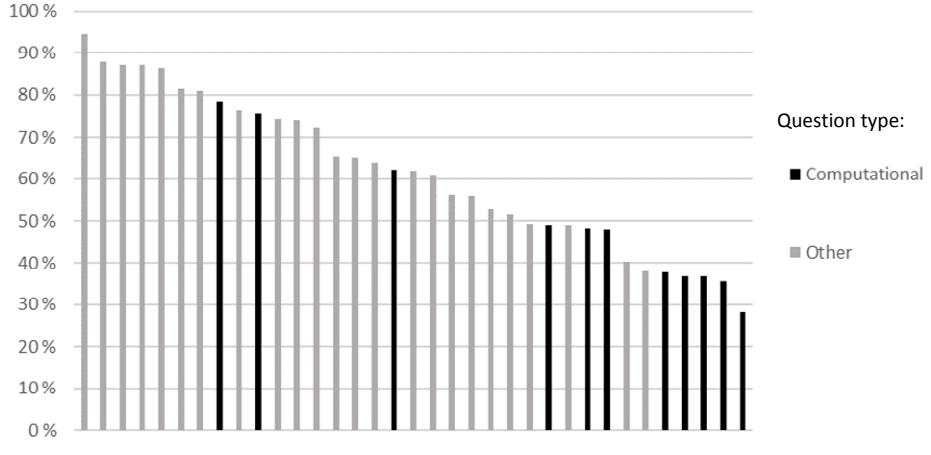
3 *Student exam-taking behaviour*: Distributions for time spent taking the exam, number of exam visits per student, and the relationship between scores and number of retakes were assessed to determine if (and to what degree) it is necessary to control the exam retakes or limit the time reserved for the exam.

4 Results and discussion

4.1 Descriptive statistics of questions

The main purpose of providing descriptive statistics was to evaluate if there were differences in the difficulty level of specific questions and to elucidate which questions were especially difficult or easy. Figure 2 illustrates this situation with questions for which there is only one correct answer. Black bars represent computational questions, and grey bars represent questions where the answer options are in text format. The illustration indicates that in this case, computational questions seem to be a bit more challenging for students than the other questions. Each question was allotted to an exam a minimum of 71 times, so it can be assumed that the differences between question statistics are not explained by random variation.

Figure 2 The proportion of correct answers per question

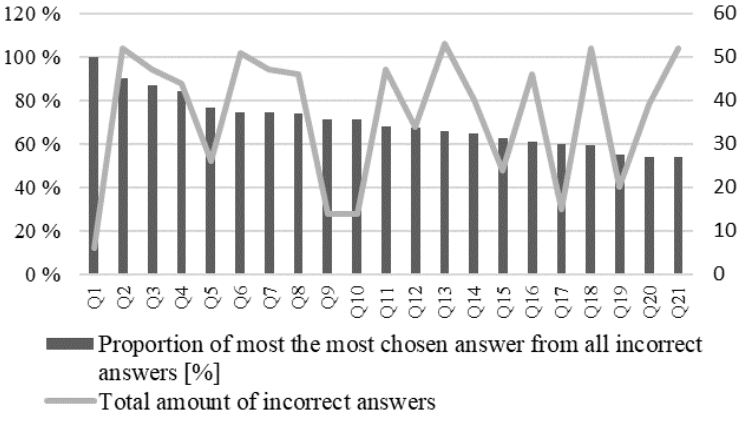


This information can be used when reviewing the questions and when revising the teaching in general. Questions that are too easy can be replaced with more challenging ones, and there is an indication that students may need more support in developing computational skills. The difficulty of questions is not unambiguous, but different questions can be compared against each other. Therefore, it is reasonable to illustrate how the proportion of wrong answers compares to other questions.

4.1.1 Error analyses for answers

To determine if some questions were systematically answered incorrectly, a bias analysis of incorrect answers was conducted. The proportion of the most chosen wrong answers among the total number of wrong answers was calculated. In Figure 3, questions for which the proportion of the most common among all incorrect answers was over 50% are organised by this metric. Questions that are seldom answered incorrectly are more prone to diverge from the average due to random variation. The more observations there are (in this case, the number of incorrect answers), the more relevant the result is. Therefore, the total number of incorrect answers per question is presented in parallel.

Figure 3 Bias analysis of incorrect answers



For example, all students answering Q1 incorrectly chose the same answer option, but the number of wrong answers was relatively low. In contrast, Q2 was answered incorrectly by over 50 students, of which 90% chose the same answer option; for Q2, there is a clear bias in incorrect answers. The content of Q2 addresses synonyms of inventory turnaround time; it is useful to identify that this is a potential subject of common misunderstanding. Q3, Q4, Q6, Q7, Q8, and Q11, which are also clearly biased, are computational questions, so it could be possible to also distinguish the bias in different question categories.

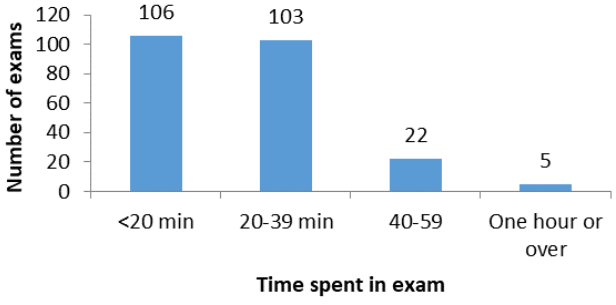
The example demonstrates that bias in incorrect answers can be measured, and the results can be valuable, but developing a generalisable metric for bias identification would require setting threshold values for both reliability (in this case, number of observations) and bias (in this case, the proportion of the most common among all wrong answers). As the number of students and the number of answer options vary by exam, developing a universally generalisable metric can be complicated. Instead, the general idea of bias measurement and the illustration type as presented in Figure 3 could be easily applied in other exams as well.

4.1.2 Students' exam-taking behaviour

Figure 4 shows the frequency distribution of time spent on the exam. Only five (2%) of the students spent 1 h or more in the exam space. Originally, the time reserved per exam was 3 h, but after the first year it was decreased to 2 h. In case of restricted capacity

(which was one of the worries expressed by students), it would be feasible to limit the visit time to 1 h. The number of students needing a longer time to complete the exam is presumably low, but extra time could be arranged for some students, such as those with diagnosed dyslexia or physical disability.

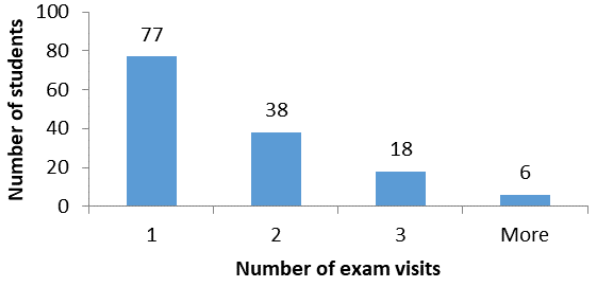
Figure 4 Frequency distribution of time spent on the exam (see online version for colours)



Observing the distribution for the time spent on the exam is more insightful than just looking at the average time spent on the exam, since it is possible to conclude how much time is needed and estimate the number of students that possibly require extended time. However, different bin limits may be relevant for different cases. Therefore, developing an automatic distribution visualisation that would serve all users in the best way possible may be a complex task.

Figure 5 presents the frequency distribution of the number of exam visits. A narrow majority of the students (55%) did not retake the exam, and retakes were typically limited to one or two. Only six (4%) of the students retook the exam more than twice. The exam was retaken a maximum of five times by only one student.

Figure 5 Frequency distribution of exam visits per student (see online version for colours)



The exam scores of five students who repeated the exam most often were analysed further. Four of these five were unable to improve their score from their first attempt, and the final scores of these five students were between 66% and 81% of the maximum score (average final score being 64%). Considering these statistics, worries about students that retake the exam frequently to 'unfairly' improve their grades seem unfounded. This example shows that unexpected analysis needs may emerge from different sources.

4.1.3 Summary and discussion

The previous examples show that while some simple generalisable metrics, such as average score or average time spent on the exam, are useful, other metrics and illustrations can also provide valuable insights from the assessment data. However, while insightful, the example analyses presented in this study are case specific, though they have generalisable features. In general, it can be assumed that there are exam-specific or user-specific analysis needs.

Providing a very broad set of possibly useful metrics for each electronic assessment system user is not necessarily the best solution. The metrics used in this study were relatively simple and can be calculated using a spreadsheet program. In practice, a greater obstacle is gathering and preparing the data for analysis. Even though it is possible to gather needed information using web scraping, it is not a convenient data gathering method for frequent use. Scraping codes need to be modified for each website – and in this case for each exam – and the data requires some cleaning before statistical tools can be used. In addition, web scraping, especially if done by inexpert means, burdens servers and, depending on the service provider, may lead to blocking or banning from a website.

Therefore, a resource-efficient solution to respond to varying analysis needs would be software allowing the user (in this case the author of the exam) access to the exam assessment data in a structured format. For example, the instructor could download complete assessment data in comma separated values (CSV) format and import the data to the software performing the analysis (e.g., spreadsheet application). Alternatively, the software could provide an application programming interface for automatic data transfer from one software or software module to another for analysis.

This study concerned a multiple-choice exam, which in a technical sense is a very simple case. However, it can be expected that the three categories for assessment analyses – descriptive statistics, error analyses, and student time use – are, in principle, also suitable for more complex, automatically evaluated learning activities/exams, such as programming assignments, mathematical assignments, or essays, though the actual metrics are naturally different. Therefore, it is recommended that all automatic evaluation systems are designed to provide the data needed for this type of analysis. In practice, this is not always a matter of course.

5 Conclusions

The case example implies that while learning analytics have recently received greater attention, not all online learning systems provide analysis tools or enable direct access to data, which is a prerequisite for performing analytics in a resource-efficient manner.

This study demonstrates that three kinds of analyses of assessment data can be useful:

1 general metrics of student performance that make it easier to determine the difficulty level of each question and the exam as a whole

2 metrics that delve deeper into specific questions in order to elucidate possible systematics behind incorrect answers

3 understanding students' exam-taking behaviours.

This case study shows that although software providers seem to tend to offer universally generalisable metrics for assessment analysis, there may be a need for case-specific metrics and analyses. A practical conclusion is that all the metrics used in the case presented in this paper are relatively simple to calculate via a spreadsheet application. If ready-made analysis tools are not provided, obtaining exam result data in CSV format would still help significantly.

The limitations of this study are those typical of single case studies. It cannot be concluded from this study how common specific analysis needs are or if there are other assessment analysis needs that are not represented in this study. In addition, this study focused only on multiple-choice questions, which is a relatively simple case of automatic assessment. However, the literature implies that it is possible to extend automatic assessment to more complex assignments. Therefore, further studies are needed to conclude what types of analyses and metrics are useful and applicable in contexts in which more complex tasks are assessed automatically.

References

Alhija, F.N.A. and Fresko, B. (2009) 'Student evaluation of instruction: what can be learned from students' written comments?', *Studies in Educational Evaluation*, Vol. 35, No. 1, pp.37–44.

Amelung, M., Krieger, K. and Rösner, D. (2011) 'E-assessment as a service', *IEEE Transactions on Learning Technologies*, Vol. 4, No. 2, pp.162–174.

Badri, M., Alnuaimi, A., Yang, G. and Rashedi, A., Al (2020) 'Examining the relationships of factors influencing student mathematics achievement', *International Journal of Innovation in Education*, Vol. 6, No. 1, pp.12–32.

Bainbridge, J., Melitski, J., Zahradnik, A., Lauría, E.J.M., Jayaprakash, S. and Baron, J. (2015) 'Using learning analytics to predict at-risk students in online graduate public affairs and administration education', *Journal of Public Affairs Education*, Vol. 21, No. 2, pp.247–262.

Baneres, D., Clariso, R., Jorba, J. and Serra, M. (2014) 'Experiences in digital circuit design courses: a self-study platform for learning support', *IEEE Transactions on Learning Technologies*, Vol. 7, No. 4, pp.360–374.

Barla, M., Bieliková, M., Ezzeddinne, A.B., Kramár, T., Šimko, M. and Vozár, O. (2010) 'On the impact of adaptive test question selection for learning efficiency', *Computers and Education*, Vol. 55, No. 2, pp.846–857.

Basu, S., Jacobs, C. and Vanderwende, L. (2013) 'Powergrading: A clustering approach to amplify human effort for short answer grading', *Transactions of the Association for Computational Linguistics*, Vol. 1, pp.391–402.

Berland, M., Davis, D. and Smith, C.P. (2015) 'AMOEBA: designing for collaboration in computer science classrooms through live learning analytics', *International Journal of Computer-Supported Collaborative Learning*, Vol. 10, No. 4, pp.425–447.

Bertheussen, B.A. (2015) 'Cultivating spreadsheet usage in a finance subject through learning and assessment innovations', *International Journal of Innovation in Education*, Vol. 3, No. 1, pp.1–14.

Biggs, J. and Tang, C. (2011) *Teaching for Quality Learning at University*, 4th ed., Open University Press, Maidenhead, UK.

Brockx, B., Van Roy, K. and Mortelmans, D. (2012) 'The student as a commentator: students' *comments in student evaluations of teaching'*, *Procedia – Social and Behavioral Sciences*, Vol. 69, pp.1122–1133.

Burrows, S., Gurevych, I. and Stein, B. (2015) 'The eras and trends of automatic short answer grading', *International Journal of Artificial Intelligence in Education*, Vol. 25, No. 1, pp.60–117, Springer New York, LLC.

Chan, K.K. (2018) 'The effect of teachers' perceptions on the role of technology in assessment: the case of Macau', *International Journal of Learning, Teaching and Educational Research*, Vol. 17, No. 2, pp.127–137.

Charleer, S., Moere, A., Vande, Klerkx, J., Verbert, K. and De Laet, T. (2018) 'Learning analytics dashboards to support adviser-student dialogue', *IEEE Transactions on Learning Technologies*, Vol. 11, No. 3, pp.389–399.

Chatti, M.A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Fahmy Yousef, A.M. and Schroeder, U. (2016) 'Video annotation and analytics in courseMapper', *Smart Learning Environments*, Vol. 3, No. 1, pp.1–21.

Chen, C.M. and Chen, M.C. (2009) 'Mobile formative assessment tool based on data mining techniques for supporting web-based learning', *Computers and Education*, Vol. 52, No. 1, pp.256–273.

Choi, S.P.M., Lam, S.S., Li, K.C. and Wong, B.T.M. (2018) 'Learning analytics at low cost: at-risk student prediction with clicker data and systematic proactive interventions', *Educational Technology and Society*, Vol. 21, No. 2, pp.273–290.

Conejo, R., Barros, B. and Bertoa, M.F. (2019) 'Automated assessment of complex programming tasks using SIETTE', *IEEE Transactions on Learning Technologies*, Vol. 12, No. 4, pp.470–484.

de Barba, P.G., Kennedy, G.E. and Ainley, M.D. (2016) 'The role of students' Motivation and participation in predicting performance in a MOOC', *Journal of Computer Assisted Learning*, Vol. 32, No. 3, pp.218–231.

Dermo, J. (2009) 'e-assessment and the student learning experience: a survey of student perceptions of e-assessment', *British Journal of Educational Technology*, Vol. 40, No. 2, pp.203–214.

Dipace, A., Fazlagic, B. and Minerva, T. (2019) 'The design of a learning analytics dashboard: eduOpen mooc platform redefinition procedures', *Journal of E-Learning and Knowledge Society*, Vol. 15, No. 3, pp.29–47.

Dubois, A. and Gibbert, M. (2010) 'From complexity to transparency: managing the interplay between theory, method and empirical phenomena in IMM case studies', *Industrial Marketing Management*, Vol. 39, No. 1, pp.129–136.

Dzikovska, M.O., Nielsen, R.D. and Leacock, C. (2016) 'The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications', *Language Resources and Evaluation*, Vol. 50, No. 1, pp.67–93.

Ellis, C. (2013) 'Broadening the scope and increasing the usefulness of learning analytics: the case for assessment analytics', *British Journal of Educational Technology*, Vol. 44, No. 4, pp.662–664.

Eriksson, P. and Kovalainen, A. (2016) *Qualitative Methods in Business Research*, 2nd ed., Sage Publications, Thousand Oaks, California.

Ernst, D. (2014) 'Expectancy theory outcomes and student evaluations of teaching', *Educational Research and Evaluation*, Vol. 20, pp.536–556.

Eyal, L. (2012) 'Digital assessment literacy – the core role of the teacher in a digital environment', *Educational Technology and Society*, Vol. 15, No. 2, pp.37–49.

Ferguson, R. (2012) 'Learning analytics: drivers, developments and challenges', *International Journal of Technology Enhanced Learning*, Vol. 4, Nos. 5/6, pp.304–317.

Florian-Gaviria, B., Glahn, C. and Fabregat Gesa, R. (2013) 'A software suite for efficient use of the European qualifications framework in online and blended courses', *IEEE Transactions on Learning Technologies*, Vol. 6, No. 3, pp.283–296.

Greller, W. and Drachsler, H. (2012) 'Translating learning into numbers: a generic framework for learning analytics', in *Journal of Educational Technology and Society*, Vol. 15, pp.42–57, International Forum of Educational Technology and Society.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) 'Design science in information systems research', *MIS Quarterly: Management Information Systems*, Vol. 28, No. 1, pp.75–105.

Holmes, M., Latham, A., Crockett, K. and andO'Shea, J.D. (2018) 'Near real-time comprehension classification with artificial neural networks: decoding e-learner non-verbal behavior', *IEEE Transactions on Learning Technologies*, Vol. 11, No. 1, pp.5–12.

Holmes, N. (2015) 'Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module', *Assessment and Evaluation in Higher Education*, Vol. 40, No. 1, pp.1–14.

Hujala, M., Knutas, A., Hynninen, T. and Arminen, H. (2020) 'Improving the quality of teaching by utilising written student feedback: a streamlined process', *Computers and Education*, Vol. 157, pp.103965.

Kaya, G., Atasoy, V., Candan-Helvacı, S. and Pektaş, M. (2021) 'The role of science teachers' awareness in their classroom practice of formative assessment', *Egitim ve Bilim*, Vol. 46, No. 205, pp.335–357.

Kitto, K., Lupton, M., Davis, K. and Waters, Z. (2017) 'Designing for student-facing learning analytics', *Australasian Journal of Educational Technology*, Vol. 33, No. 5, pp.152–168.

Lamberti, F., Sanna, A., Paravati, G. and Carlevaris, G. (2014) 'Automatic grading of 3D computer animation laboratory assignments', *IEEE Transactions on Learning Technologies*, Vol. 7, No. 3, pp.280–290.

Lawson, C., Beer, C., Rossi, D., Moore, T. and Fleming, J. (2016) 'Identification of 'at risk' students using learning analytics: the ethical dilemmas of intervention strategies in a higher education institution', *Educational Technology Research and Development*, Vol. 64, No. 5, pp.957–968.

Leckey, J. and Neill, N. (2001) 'Quantifying quality: the importance of student feedback', *Quality in Higher Education*, Vol. 7, No. 1, pp.19–32.

Liu, M., Calvo, R.A., Pardo, A. and Martin, A. (2015) 'Measuring and visualizing students' behavioral engagement in writing activities', *IEEE Transactions on Learning Technologies*, Vol. 8, No. 2, pp.215–224.

Liu, M., Li, Y., Xu, W. and Liu, L. (2017) 'Automated essay feedback generation and its impact on revision', *IEEE Transactions on Learning Technologies*, Vol. 10, No. 4, pp.502–513.

Livne, N.L., Livne, O.E. and Wright, C.A. (2007) 'Can automated scoring surpass hand grading of students' Constructed responses and error patterns in mathematics?', *Journal of Online Learning and Teaching*, Vol. 3, No. 3, pp.295–306.

Llamas-Nistal, M., Fernández-Iglesias, M.J., González-Tato, J. and Mikic-Fonte, F.A. (2013) 'Blended e-assessment: migrating classical exams to the digital world', *Computers and Education*, Vol. 62, pp.72–87.

Lowenthal, P., Bauer, C. and Chen, K.Z. (2015) 'Student perceptions of online learning: an analysis of online course evaluations', *American Journal of Distance Education*, Vol. 29, No. 2, pp.85–97.

Mangaroska, K. and Giannakos, M. (2019) 'Learning analytics for learning design: a systematic literature review of analytics-driven design to enhance learning', *IEEE Transactions on Learning Technologies*, Vol. 12, No. 4, pp.516–534.

Martinez-Maldonado, R., Yacef, K. and Kay, J. (2015) 'TSCL: a conceptual model to inform understanding of collaborative learning processes at interactive tabletops', *International Journal of Human Computer Studies*, Vol. 83, pp.62–82.

McCann, A.L. (2010) 'Factors affecting the adoption of an E-assessment system', *Assessment and Evaluation in Higher Education*, Vol. 35, No. 7, pp.799–818.

McDonald, J., Bird, R.J., Zouaq, A. and Moskal, A.C.M. (2017) 'Short answers to deep questions: supporting teachers in large-class settings', *Journal of Computer Assisted Learning*, Vol. 33, No. 4, pp.306–319.

Melero, J., Hernández-Leo, D., Sun, J., Santos, P. and Blat, J. (2015) 'How was the activity? A visualization support for a case of location-based learning design', *British Journal of Educational Technology*, Vol. 46, No. 2, pp.317–329.

Moridis, C.N. and Economides, A.A. (2009) 'Mood recognition during online self-assessment tests', *IEEE Transactions on Learning Technologies*, Vol. 2, No. 1, pp.50–61.

Nair, C.S. and Adams, P. (2009) 'Survey platform: a factor influencing online survey delivery and response rate', *Quality in Higher Education*, Vol. 15, No. 3, pp.291–296.

Nederhand, M., Auer, J., Giesbers, B., Scheepers, A. and van der Gaag, E. (2022) 'Improving student participation in SET: effects of increased transparency on the use of student feedback in practice', *Assessment and Evaluation in Higher Education*.

Nouira, A., Cheniti-Belcadhi, L. and Braham, R. (2019) 'An ontology-based framework of assessment analytics for massive learning', *Computer Applications in Engineering Education*, Vol. 27, No. 6, pp.1343–1360.

Omorog, C.D. (2020) 'Mining mySQL error logs to map student learning', *International Journal of Innovation in Education*, Vol. 6, No. 1, pp.1–11.

Papamitsiou, Z. and Economides, A.A. (2016) 'An assessment analytics framework (AAF) for enhancing students' *Progress'*, *Formative Assessment, Learning Data Analytics and Gamification: In ICT Education*, pp.117–133.

Paravati, G., Lamberti, F., Gatteschi, V., Demartini, C. and Montuschi, P. (2017) 'Point cloud-based automatic assessment of 3D computer animation courseworks', *IEEE Transactions on Learning Technologies*, Vol. 10, No. 4, pp.532–543.

Pardo, A., Ellis, R.A. and Calvo, R.A. (2015) 'Combining observational and experiential data to inform the redesign of learning activities', *ACM International Conference Proceeding Series,* pp.16–20 March, pp.305–309.

Pattanasri, N., Mukunoki, M. and Minoh, M. (2012) 'Learning to estimate slide comprehension in classrooms with support vector machines', *IEEE Transactions on Learning Technologies*, Vol. 5, No. 1, pp.52–61.

Popham, W.J. (2004) 'Why assessment illiteracy is professional suicide', *Educational Leadership*, Vol. 62, No. 1, pp.82–83.

Pursel, B.K., Zhang, L., Jablokow, K.W., Choi, G.W. and Velegol, D. (2016) 'Understanding MOOC students: motivations and behaviours indicative of MOOC completion', *Journal of Computer Assisted Learning*, Vol. 32, No. 3, pp.202–217.

Rodríguez-Triana, M.J., Martínez-Monés, A., Asensio-Pérez, J.I. and Dimitriadis, Y. (2015) 'Scripting and monitoring meet each other: aligning learning analytics and learning design to support teachers in orchestrating CSCL situations', *British Journal of Educational Technology*, Vol. 46, No. 2, pp.330–343.

Sanna, A., Lamberti, F., Paravati, G. and Demartini, C. (2012) 'Automatic assessment of 3D modeling exams', *IEEE Transactions on Learning Technologies*, Vol. 5, No. 1, pp.2–10.

Saqr, M. (2017) 'Assessment analytics: the missing step', *International Journal of Health Sciences*, Vol. 11, No. 1, pp.1–2.

Siddiqi, R., Harrison, C.J. and Siddiqi, R. (2010) 'Improving teaching and learning through automated short-answer marking', *IEEE Transactions on Learning Technologies*, Vol. 3, No. 3, pp.237–249.

Siemens, G. (2013) 'Learning analytics: the emergence of a discipline', *American Behavioral Scientist*, Vol. 57, No. 10, pp.1380–1400.

Smith, L.F., Hill, M.F., Cowie, B. and Gilmore, A. (2014) *Preparing Teachers to Use the Enabling Power of Assessment*, pp.303–323.

Stewart, M. (2015) 'The language of praise and criticism in a student evaluation survey', *Studies in Educational Evaluation*, Vol. 45, pp.1–9.

Stiggins, R. (1991) 'Assessment literacy', *Phi Delta Kappan*, Vol. 72, No. 7, pp.534–539.

Tervakari, A., Kuosa, K., Koro, J., Paukkeri, J. and Kailanto, M. (2014) 'Teachers' learning analytics tools in a social media enhanced learning environment', *Proceedings of 2014 International Conference on Interactive Collaborative Learning, ICL 2014*, pp.355–360.

Tucker, B. (2014) 'Student evaluation surveys: anonymous comments that offend or are unprofessional', *Higher Education*, Vol. 68, No. 3, pp.347–358.

Van Aken, J.E. (2004) 'Management research based on the paradigm of the design sciences: the quest for field-tested and grounded technological rules', *Journal of Management Studies*, Vol. 41, No. 2, pp.219–246.

Van Aken, J., Chandrasekaran, A. and Halman, J. (2016) 'Conducting and publishing design science research: inaugural essay of the design science department', *Journal of Operations Management*, Vols. 47–48, pp.1–8.

Webb, M., Gibson, D. and Forkosh-Baruch, A. (2013) 'Challenges for information technology supporting educational assessment', *Journal of Computer Assisted Learning*, Vol. 29, No. 5, pp.451–462.

Wiggins, G.P. and McTighe, J. (2007) *Schooling by Design□: Mission, Action, and Achievement*, p.285.

Willis, A. (2015) 'Using NLP to support scalable assessment of short free text responses', *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistic, Denver, Colorado, pp.243–253.

Wilson, K., Boyd, C., Chen, L. and Jamal, S. (2011) 'Improving student performance in a first-year geography course: examining the importance of computer-assisted formative assessment', *Computers and Education*, Vol. 57, No. 2, pp.1493–1500.

Worsley, M. and Blikstein, P. (2013) 'Towards the development of multimodal action based assessment', *ACM International Conference Proceeding Series*, Association for Computing Machinery, New York, United States, pp.94–101.

Xu, Y. and Brown, G.T.L. (2016) 'Teacher assessment literacy in practice: a reconceptualization', *Teaching and Teacher Education*, Vol. 58, pp.149–162.

Yin, R.K. (2014) *Case Study Research: Design and Methods*, 5th ed., SAGE, Los Angeles.

Zesch, T., Heilman, M. and Cahill, A. (2015) 'Reducing annotation efforts in supervised short answer scoring', *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistic, Denver, Colorado, pp.124–132.