Research in Learning Analytics and Educational Data Mining to Measure Self-Regulated Learning: A Systematic Review

Amr Adel ElSayed
Arab Academy for Science, Technology, & Maritime Transport
P.O. Box 2033, Elhorria, Cairo, Egypt
amradel.ecommerce@aast.edu

Manuel Caeiro-Rodríguez, Fernando A. Mikic-Fonte, Martín Llamas-Nistal
University of Vigo
E.E. Telecomunicación, Campus-Universitario S/N 36312 Vigo, Spain
{mcaiero; mikic; martin}@gist.uvigo.es

ABSTRACT
Up to date, most of the research to measure Self-Regulated Learning in students has primarily utilized self-report instruments. Recently, there has been a growing tendency towards using other assessment tools; particularly in the context of Learning Analytics and Educational Data Mining. However, there is a gap in the literature to review the application of new techniques used in these domains related to data analytics. To address this gap, we conducted a systematic literature review focusing on the measurement of Self-Regulated Learning features and behaviours in students based on the analysis of tracking and log data using techniques such as cluster analysis, regression or classification; either solely, or associated with self-report instruments. This review aims to categorize the data used in the different papers to measure Self-Regulated Learning and to recognize the behaviours/features/components measured. In addition, it also analyses the most frequently used tools and the application disciplines. This systematic literature review surveys the literature for an eight-year time span from 2011 to 2019, following the general guidelines of systematic reviews with clearly established eligibility criteria. After applying the eligibility criteria, 109 studies were identified as relevant. The findings show an increasing interest in the use of Learning Analytics and Educational Data Mining for assessing Self-Regulated Learning in students, and the tendency to associate the new data analysis techniques with other self-reported measures to obtain data triangulation.

Author Keywords
Self-Regulated Learning, Learning Analytics, Educational Data Mining, Systematic Review, Assessment Tools

INTRODUCTION
[Zimmerman 2002] defined Self-Regulated Learning (SRL) as a “self-directive process by which learners transform their mental abilities into academic skills”. Self-regulated learners are learners who ‘plan, set goals, organize, self-monitor and self-evaluate’ [Zimmerman 1989]. Such learners are self-aware, knowledgeable and capable of deciding their learning approach [Corno 1986]. Moreover, these learners can effectively select and use metacognitive and motivational strategies; can take the initiative to select, structure and create an advantageous learning environment; and can actively participate in selecting the form and amount of required instruction, particularly, asking for help when needed.

SRL is a comprehensive term that includes variables which affect learning such as self-efficacy, volition, cognitive, meta-cognitive, and motivational regulatory components [Panadero 2017]. The core ideas of SRL are motivation and learning strategies adopted by learners to reach their learning objectives. Metacognitive strategies include processes such as the learners’ capability of planning, scheduling, and assessing their advancement in the learning process [Kuo et al. 2013]. In any case, different SRL models exist. For a review of these models refer to [Panadero 2017].

Learning Analytics and Educational Data Mining
According to [Long & Siemens 2014] Learning Analytics (LA) is defined as “... measurement, collection, analysis and presentation of data on students and their contexts, for the purposes of understanding and optimization of learning and the environments in which it takes place”. LA is based on the capture and analysis of the huge amount of individual student interaction data produced by blended or online learning environments (such as Learning Management Systems (LMS). The captured data can be utilized to predict the performance of learners, recognize their study traces, enable learners to reflect about their actual learning activities precisely while the course is progressing, allow the extraction and discovery of educational patterns, etc. More specifically, LA can be used to support learners in their development as SRL students providing them with options towards more productive utilization of learning strategies during different SRL phases in an authentic learning environment [Karlen 2016; Jivet et al. 2018; Koç 2017; Nguyen, Gardner, & Sheridan 2018; Soffer & Cohen 2019; Cicchinelli et al. 2018; Roll & Winne 2015; Chatti & Muslim 2019].

Educational Data Mining (EDM) is a subset of LA [Chatti et al. 2013] whose main concern is to “...analyze, develop, and research the vast amount of educational data automatically in order to distinguish their patterns” [Winne & Baker 2013; Umbleja & Ichino 2016]. While LA focuses on developing methods for utilizing educational data sets to boost the learning process, EDM is more focused on exploring and analyzing the educational data to gain better understanding of the students’ learning settings [Chatti et al. 2013].
Both LA and EDM use methods taken from data analytics fields (e.g. data mining, machine learning) in order to analyze data. There are a large number of methods that can be classified into several categories [Baker 2010]: prediction (e.g. classification, regression, and density estimation), clustering, relationship mining (e.g. association rule mining, correlation mining, sequential pattern mining, causal data mining), discovery with models, and distillation of data for human judgment.

**Measuring Self-Regulated Learning using Self-report Instruments**
A large amount of prior research has focused on the use of self-report instruments in SRL measurement. Self-report instruments can be classified into four main self-report instruments [Roth, Ogrin, & Schmitz 2016]: Questionnaires, Interviews, Think-Aloud Protocols (TAP), and Learning Diaries.

Self-report measurements of SRL view self-regulation as an aptitude possessed by the learners. Such measurements depend primarily on information provided by learners as a self-assessment. Therefore, such information may be biased and inaccurate as it is affected by the learners’ memory retrieval capacity, as well as their own perceptions and viewpoints. In addition, these measurements don’t present how learners dynamically adapt and modify their learning tactics and strategies during the actual learning process [Kinnebrew, Loretz, & Biswas 2013; Panadero, Klug, & Järvelä 2016; Berger & Karabenick 2016; Fincham et al. 2018; Kovanović et al. 2015].

**Measuring Self-Regulated Learning using Learning Analytics and Educational Data Mining techniques**
Due to the above-mentioned shortcomings of measuring SRL using self-report instruments only, there has been a growing tendency towards utilizing other assessment tools, especially used in the LA and EDM research. It is important to notice a main difference: now the goal is to track and directly analyze the actual SRL experience of learners, instead of depending on learners’ perceptions [Jovanović et al. 2017]. Thus, these works aim to assess SRL as an event (time-based and task-related to known start and end) rather than an aptitude (a steady feature of the learner) [Panadero, Klug, & Järvelä 2016]. The measurement of the learners’ behavioural traces is considered more essential than measuring their self-assessments because LA can use such traces to create learning environments that can improve the students’ usage of learning strategies to enhance their learning [Van Laer & Elen 2018; Zhang et al. 2018; Reimann, Markauskaite, & Bannert 2014; Cicchinelli et al. 2018]. Particularly, contextualized environments and mobile devices have already been employed for gathering and monitoring LA for self-regulation [Tabuencía et al. 2015].

SRL measurement using data analytics techniques is taking two approaches: either alone, or with other self-report measures to obtain data triangulation [Panadero, Klug, & Järvelä 2016]. This review investigates both approaches.

**Objectives**
The goal of this paper is to introduce a systematic review about the use of data analytics techniques for the assessment of SRL features and behaviours in students. We consider the different approaches: either solely, or associated with self-report instruments. To achieve this goal, this review will discuss the following issues:

1. What are the commonly used techniques of data analytics in SRL assessment?
2. What are the types of data used to assess SRL?
3. What are the frequently measured SRL behaviours/ features/ components?

**METHODOLOGY**
This review selected previous research published in the period from January 2011 to March 2019. The search took place during the period from the 5th of February to the 1st of April 2019. The review was conducted using a four-step-approach depicted in Figure 1. This review followed the guidelines of [Kitchenham 2004].

![Figure 1 Review Procedure](image)

Search queries and databases
Search queries included the following terms and combinations: “Assess Self-Regulated Learning (skills)”, Measure Self-Regulated Learning (skills)”, “Learning Analytics Self-Regulated Learning (skills)”, “Neural networks Self-Regulated Learning (skills)”, “Data mining Self-Regulated Learning (skills)”, “Process Mining Self-Regulated Learning (skills)”, “Data Analytics Self-Regulated Learning (skills)”. We searched for the above-mentioned search queries in the following list of established databases: EBSCOhost Academic Search Complete, Emerald e-journals premier collection, Education Resources Information Center (ERIC), Google Scholar, IEEE Xplore, Science Direct, Scopus, SpringerLink, Taylor & Francis, and Wiley Online Library. An initial search of the selected databases using the selected search queries resulted in 1476 results.

**Filtering search results according to eligibility criteria**
This stage involved reading the abstracts, conclusions, and methodology sections of the retrieved research. The remaining studies are those that satisfied all of the eligibility criteria mentioned in Table 1.
RESULTS
109 studies fulfilled the eligibility criteria. The majority (exactly 83) of the selected studies were published in scientific journals and 26 studies were published in international conferences.

<table>
<thead>
<tr>
<th>The publication year is restricted between 2011 and 2019 (to exclude outdated research)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The language is English</td>
</tr>
<tr>
<td>The reviewed research includes only peer-reviewed papers &amp; articles</td>
</tr>
<tr>
<td>Research to which the researchers have full text access; whether via open access or provided by the academic networks of the University of Vigo, Arab Academy for Science, Technology, &amp; Maritime Transport, and the Egyptian Knowledge Bank</td>
</tr>
<tr>
<td>Research that applies LA and EDM techniques, either solely or combined with self-report instruments, to measure SRL and its skills</td>
</tr>
</tbody>
</table>

Table 1 Eligibility Criteria

More than half (exactly 59) of the selected studies used data analytics in association with other self-report instruments. The remaining 51 studies utilized data analytics techniques alone.

From Figure 2, it can be noticed that the number of publications studying the utilization of LA and EDM tools in SRL assessment was gradually increasing during the period from 2011 to 2018 except for a slight decline in 2014 and 2017. We found 8 publications from the 1st of January to the 1st of April 2019.

![Figure 2 Distribution of the selected studies by the year of publication](image)

The findings indicate the existence of five disciplines where studies assess SRL using data analytics techniques. These disciplines are: undergraduate university education, school education, open online education, post-graduate education, and workplace settings. The majority of studies (61.82%) were devoted towards assessing SRL in undergraduate university education; followed by school education (22.73%), open online education (7.27%), post-graduate education (5.45%), and workplace (2.73%) respectively.

What are the commonly used techniques of data analytics in SRL assessment?
A large proportion of studies depends on the analysis of trace data generated in log files. Log files generated by LMS can include the behaviour changes of learners during course progression, their interaction with their peers, and their performance in assessments [You 2016]. Log files can also be generated by the means of video annotation software [Gašević, Mirriaži, & Dawson 2014; Gašević et al. 2017; Pardo et al. 2015] or an induction-mining algorithm [Poitras, Doleck, & Lajoie 2018; Poitras et al. 2017]. Other researchers used multimodal trace data (including physiological data such as heart rate, eye-tracking data, step count, facial expressions, and weather conditions; in addition to learning activities) [Bouchet et al. 2013; Di Mitri et al. 2016; Trevors et al. 2016; Di Mitri et al. 2017; Tempelaar, Rientes, & Nguyen 2018; Taub & Azevedo 2019].

Log file analysis includes investigation of trace data left by learners which can lead to a better understanding and discovery of the learners’ behavioural patterns [Li et al. 2018].

Traces obtained from log files enable the provision of consistent information regarding the learners’ usage of cognitive tools during the learning process [Malmberg, Järvelä, & Kirschner 2014]. [Siadaty, Gašević, & Hatala 2016] presented a trace-based methodology for SRL measurement.

The next subsections present a brief description regarding the data mining techniques used for SRL assessment. Each technique and the number of studies using it are presented in Table 2.

Cluster Analysis
Clustering is an unsupervised technique; typically used to categorize the learner behavioural profiles into subset clusters [Romero & Ventura 2010; Segedy, Kinnebrew, & Biswas 2015]. In the same paper, one or more clustering method can be used. Overall, clustering is the most popular technique. It has been used in 42 studies. This popularity is because clustering
enables the provision of more adaptive scaffolding either through agent-based Intelligent Tutoring Systems or Virtual classrooms [Bouchet et al. 2013].

The most commonly used cluster analysis technique in this systematic review is Agglomerative Hierarchical Clustering (22 studies) followed by K-means clustering (13 studies). Latent Profile Analysis (LPA) was applied in 6 studies. 3 studies applied Expectation-Maximization (EM) clustering algorithm, while clustering using Tamhane’s Post Hoc Test and K-medoids clustering were applied in only one study for each.

**Classification**

Classification algorithms are supervised algorithms which predict the membership of individual data instances into groups relying on quantitative information concerning the feature(s) of these instances; according to a training set of formerly labeled items [Phyu 2009; Romero & Ventura 2010]. In the same paper, one or more classification methods can be used. Overall, classification has been used in 19 studies.

Regression analysis algorithms are the most frequently utilized classification algorithms in the selected studies; they were used in 17 studies. Bayesian classifiers were used in 6 studies. Neural networks were exploited in 5 studies. 4 studies utilized Decision tree classification. The same number used support vector machines. A single study exploited k-Nearest Neighbor algorithm.

**Temporal Data mining techniques (Sequence mining and Process mining)**

EDM techniques that use temporal patterns of SRL examine the connections among recurrent SRL events according to their order and temporality. They include sequence mining and process mining (workflow mining) [Reimann, Markauskaite, & Bannert 2014; Malmberg et al. 2015]. In the same paper, one or more temporal data mining technique can be used. Overall, temporal data mining has been used in 26 studies.

Sequence mining examines event behavioural patterns, involving the occurrence frequency of these patterns [Taub & Azevedo 2019]. Sequence mining methods are implemented either separately, or combined (differential sequence mining methodology) to differentiate between frequent patterns among learners [Kinnebrew, Loretz, & Biswas 2013]. 15 studies have employed sequence mining.

The concept of Process Mining is to create a process model to comprehensively describe and discover the processes in an event log file. Events in the event log can be time-stamped or fully-ordered sequences [Sonnenberg & Bannert 2015; Reimann, Markauskaite, & Bannert 2014]. The exploited process mining methods include fuzzy miner, different versions of ProM software, PM² method, bupaR process mining, Markov models, heuristic miner, and conformance checking. 13 studies have employed process mining.

**Other Analytics**

This review indicates the use of less frequent tools for SRL assessment, include Social network analysis (SNA) [Gewerc, Rodríguez-Groba, & Martínez-Piñeiro 2016], Burst-detection [Piotrkowicz, Dimitrova, & Roberts 2018], Mobile learning analytics [Tabuenca et al. 2015], Anomaly detection data mining technique [Taub et al. 2014], Summarization technique [Taub et al. 2014], and Principle component analysis [Umbleja & Ichino 2016].

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Number</th>
<th>Classification</th>
<th>Number</th>
<th>Temporal data mining</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agglomerative Hierarchical</td>
<td>22</td>
<td>Regression analysis</td>
<td>17</td>
<td>Sequence Mining</td>
<td>15</td>
</tr>
<tr>
<td>K-means</td>
<td>13</td>
<td>Bayesian Classifiers</td>
<td>6</td>
<td>Process Mining</td>
<td>13</td>
</tr>
<tr>
<td>Latent Profile Analysis</td>
<td>6</td>
<td>Neural Networks</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expectation-Maximization</td>
<td>3</td>
<td>Decision trees</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tamhane’s Post Hoc Test</td>
<td>1</td>
<td>Support Vector Machines</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-medoids</td>
<td>1</td>
<td>K-Nearest Neighbor</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>Total</td>
<td>19</td>
<td>Total</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2 Distribution of data mining techniques among studies

**What are the types of data used to assess SRL?**

The data used in different papers can be classified into seven categories. The selected papers utilized one or more of those data types (cf. Figure 3):

1. Log data (used in 77 studies): It emerges as a result of learner’s interaction in learning environments and tools such as LMS, Intelligent Tutoring Systems (ITS), nStudy tool, gStudy tool, mobile devices, and video annotation software.
2. Assessment results (used in 27 studies): It includes scores of exams, quizzes, pre- & post-tests, assignments, mind maps, as well as Workplace-Based Assessment (WBA).
3. Multimodal data (used in 9 studies): It includes physiological data such as heart rate, eye-tracking data, step count, facial expressions, and weather conditions; in addition to learning activities.

49
4. Demographic data (used in 9 studies): It includes data such as gender, age, class level, faculty, & study mode [Broadbent & Fuller-Tyszkievič 2018].
5. Chat and forum conversations (used in 5 studies): It involves online discussions among learners [Järvelä, Malmberg, & Koivuniemi 2016; Wise & Hsiao 2018].
6. Self-reported data from learners via instruments including various types of questionnaires, interviews, surveys, SRL quizzes, TAP, and Motivated Strategies for Learning Questionnaire (MSLQ) and its subsets (used in 59 studies).
7. Video recordings (used in 5 studies).

What are the frequently measured SRL behaviours/features/components?
In this review, we utilized most of the SRL categories as classified by [Garcia, R., Falkner, K., & Vivian, R. 2018] (in addition to the category of learning strategies). The selected studies measure at least one of those categories (cf. Figure 4):

1. Goal-setting, planning, and time management (measured in 57 studies): The learner sets precise goals for him/herself [Zimmerman 2002].
2. Keeping, reviewing records, and monitoring (measured in 36 studies): It refers to the learners’ monitoring of their progression toward achieving goals [Zimmerman 2002].
3. Emotion regulation (measured in 21 studies).
4. Learning strategies (measured in 19 studies): It includes behaviours such as reading, repeating, elaboration, judgment of relevance, taking notes, summarizing, coordinating different information sources, activating prior knowledge, peer-learning, processing, questioning, and problem-solving.
5. Self-evaluation (measured in 18 studies): It refers to “comparisons of self-observed performances against some standard, such as one's prior performance, another person's performance, or an absolute standard of performance” [Zimmerman 2002].
6. Seeking information and social help (measured in 16 studies): It pertains learners searching for assistance and information from others [Zimmerman 2002].
7. Organizing & transforming (measured in 6 studies).
8. Environmental structuring (measured in 4 studies): It involves the learners’ usage of e-learning platforms to alter the digital area for achieving their learning objectives [Garcia, R., Falkner, K., & Vivian, R. 2018].
9. Rehearsing & memorizing (measured in 4 studies).

STUDY LIMITATIONS
This study has some limitations. Only research published in English was selected. Furthermore; we found a lot of studies that utilized statistical methods. It was hard to distinguish between studies using statistical methods to provide an analysis of SRL and those that just provide a statistical measure. This study is exploratory and quantitative. It didn’t focus on the quality of measurement provided by every studied assessment tool. Finally, the main focus of this review was on the LA & EDM tools used for SRL assessment. This review didn’t analyze in depth neither the nature of SRL processes and components being assessed, nor the detailed structure of the data utilized for SRL assessment via LA and EDM tools.

CONCLUSION
This review tries to close a literature gap by providing an overview of the research in the LA and EDM domains for the measurement of SRL; either solely, or in association with self-report instruments. It reveals the allocation of data analytics tools applied, determine the most frequently used ones, and their application disciplines. The review also classifies the data used for SRL measurement, determines the behaviours/features/components measured in each study, identifies the
frequently used LA and EDM tools, and finds the applicable disciplines where LA and EDM tools are commonly employed to assess SRL. The main conclusion of this review is that there is a real interest in this field and a great variety of tools and approaches to be followed. Also, it is quite clear that there is no specific solution.

The attained results reveal that the introduction of LA and EDM instruments has not removed the use of self-report instruments. About 54% of the selected studies exploited a combination of self-reported instruments and LA. These findings are compatible with [Panadero, Klug, & Järvelä 2016] and [Pardo, Han, & Ellis 2017] who mentioned that self-reported measures are used in association with other measures to obtain data triangulation and acquire a more complete understanding.

As observed, time management is the most frequently measured SRL behaviour. This result is compatible with the findings of [Tabuenca et al. 2015] who found that mobile devices can be used to track the devoted learning time in mobile learning; and subsequently; improve time management skills of students and support their SRL during online courses.

REFERENCES


Umbleja, K., & Ichino, M. (2016). Predicting Students’ Behavior During an E-Learning Course Using Data Mining. *Proceedings of International Conference on Interactive Collaborative Learning (pp. 175-189).* Springer, Cham. DOI:10.1007/978-3-319-50340-0_14


