Yang, C. C. Y., Chen, I. Y. L., & Ogata, H. (2021). Toward Precision Education: Educational Data Mining and Learning Analytics for Identifying Students' Learning Patterns with Ebook Systems. *Educational Technology & Society*, *24* (1), 152–163.

Toward Precision Education: Educational Data Mining and Learning Analytics for Identifying Students' Learning Patterns with Ebook Systems

Christopher C. Y. Yang[1*], Irene Y. L. Chen[2] and Hiroaki Ogata[3]
[1]Graduate School of Informatics, Kyoto University, Japan // [2]Department of Accounting, National Changhua University of Education, Taiwan // [3]Academic Center for Computing and Media Studies, Kyoto University, Japan // yang.yuan.57e@st.kyoto-u.ac.jp // irene@cc.ncue.edu.tw // hiroaki.ogata@gmail.com
[*]Corresponding author

ABSTRACT: Precision education is now recognized as a new challenge of applying artificial intelligence, machine learning, and learning analytics to improve both learning performance and teaching quality. To promote precision education, digital learning platforms have been widely used to collect educational records of students' behavior, performance, and other types of interaction. On the other hand, the increasing volume of students' learning behavioral data in virtual learning environments provides opportunities for mining data on these students' learning patterns. Accordingly, identifying students' online learning patterns on various digital learning platforms has drawn the interest of the learning analytics and educational data mining research communities. In this study, the authors applied data analytics methods to examine the learning patterns of students using an ebook system for one semester in an undergraduate course. The authors used a clustering approach to identify subgroups of students with different learning patterns. Several subgroups were identified, and the students' learning patterns in each subgroup were determined accordingly. In addition, the association between these students' learning patterns and their learning outcomes from the course was investigated. The findings of this study provide educators opportunities to predict students' learning outcomes by analyzing their online learning behaviors and providing timely intervention for improving their learning experience, which achieves one of the goals of learning analytics as part of precision education.

Keywords: Precision education, Learning analytics, Educational data mining, Learning pattern, Ebook learning log

1. Introduction

Precision education (Yang, 2019) has been recognized as a new challenge of applying artificial intelligence, machine learning, and learning analytics to improve both learning performance and teaching quality. To facilitate precision education, digital learning platforms have been reported to play an important role to collect the educational records of student's online and offline learning behaviors, performances, and other types of interactions (Maldonado-Mahauad et al., 2018). Utilizing the increasing volume of education data collected from various digital learning platforms, Siemens and Long (2011) identified two areas, educational data mining (EDM) and learning analytics (LA), which enable the integration and investigation of big data capabilities in education. Consequently, analyzing students' online learning behavior and identifying subgroups of students with certain learning patterns on various digital learning platforms have drawn considerable attention from the LA and EDM research communities.

LA has been identified as a conceptual framework for the analysis of student behavior and includes the prediction of students' learning performance, the process development of education data analysis (Hwang, Chu & Yin, 2017), data collection, and timely interventions (Hwang, 2014). One of the goals of LA as a part of precision education has been indicated to the prediction of students' learning outcomes from the course by analyzing their behaviors of online learning and providing timely intervention for improvement of their learning experience (Lu et al., 2018). LA approaches typically rely on educational data collected from users' interactions with information and communication technologies, such as Learning Management System (LMS) or social media (Gašević et al., 2016). The relevant literature has indicated that the development of education technologies that utilize education data gathered from learners has enabled various LA approaches to be used to help students succeed in various educational contexts (Wong, 2017; Kumar & Kumar, 2018). The analytics results will not only increase benefits for educators and learners but also present considerable potential for optimizing institutional processes (Colvin et al., 2015). Data mining techniques are commonly applied to identify patterns among students based on education data (Baker & Inventado, 2014). The interpretation of these identified patterns provides insight into learning and teaching processes, predicts students' learning outcomes from the course, suggests supportive interventions, and facilitates decision-making on resource allocation (Gašević et al., 2016). LA and EDM constitute a well-established framework that utilizes certain methods for successively gathering, processing, reporting, and acting on machine-readable data from learners to advance the educational

environment (Papamitsiou & Economides, 2014) and demonstrate considerable potential for understanding and optimizing the learning process (Baker & Inventado, 2014).

However, without a series of data analytics methods for the identification of subgroups of students and their learning patterns using LA and EDM approaches, it has been demonstrated in previous studies that students have difficulty with accurately identifying and describing how they studied and what strategies they applied to previous learning activities (Zhou & Winne, 2012). Furthermore, the relevant literature indicates that students often fail to adjust the learning strategies they previously adopted to better address changing learning situations (Lust, Elen & Clarebout, 2013). Consequently, students tend to employ suboptimal learning strategies (Winne & Jamieson-Noel, 2003) when encountering unfamiliar learning situations.

To better understand how students learn and behave in these learning environments, in this study, the authors applied data analytic methods using the LA and EDM approaches to identify subgroups of students with various learning patterns utilizing an ebook system. The aim of the study was to explore student learning while an ebook system was used for both teaching and learning support. A clustering approach was employed to identify subgroups of students with different patterns of ebook learning behavior. Moreover, a statistical analysis was performed to investigate the associations between the identified subgroups of students and their learning outcomes from the course. Accordingly, the authors aimed to answer the following research questions:

RQ1: How many subgroups of students with different patterns of learning can be identified when they utilize an ebook system?

RQ2: What is the association between these subgroups of students and their learning outcomes from the course when learning with an ebook system?

2. Literature review

2.1. LA and EDM

Higher education has seen rapid development with the integration of the internet and various web-based technologies. With the rapid development and use of digital learning platforms, such as OpenCourseWare (OCW) and massive open online courses (MOOCs), education has experienced substantial growth in the volume of data derived from students' interactions with technology and their personal and academic profiles (Ferguson, 2012). The data collected from various learning activities have been used to develop predictive models of user behavior with increasing frequency in areas such as marketing, financial markets, sports, and health. Given the richness of the data that are collected in various educational settings, a growing number of educational institutions has been applying LA and EDM to support learners' strategic planning and decision-making (Chen et al., 2020). In recent years, many postsecondary educational institutions have started utilizing the data from learning activities they designed for their courses, combining LA approaches with data mining algorithms to better understand students' learning processes and their successes during the course (Pardo, Han & Ellis, 2016).

The uses of LA and EDM can provide potentially useful information from large volumes of unstructured data (Chen et al., 2020). The application of LA and EDM can promote and improve the design of online learning platforms, learning materials, and activities to ensure greater educational effectiveness and optimize learning environments (Greller & Drachsler, 2012). The obtained analysis results provide opportunities for course instructors and students to examine and adjust their strategies of teaching and learning, respectively. Consequently, with the constant adoption of new strategies, new education data will continuously be generated, leading to new analysis results for course instructors and students (Hwang, Chu & Yin, 2017).

The data collected from various digital learning platforms are now being utilized within various supportive environments. Student behavior modeling has received considerable attention in the field of EDM (Papamitsiou & Economides, 2014). Regarding LA and EDM approaches, the clustering and classification of education data seem to be the techniques most frequently used to measure and interpret student online learning behaviors when they are interacting with a digital learning platform. For example, Krumm et al. (2014) proposed an early warning system to detect at-risk students who might fail a course using students' learning data gathered through their interactions with various digital learning platforms. Hu, Lo and Shih (2014) developed an early warning system using a decision tree classifier. Corrin & De Barba (2015) propose a reporting system that presents visualized information to support students and instructors by reflecting on students' learning processes during a given period. Romero et al. (2013) used a sequential minimal optimization classification approach and students'

learning behavioral data before a midterm exam to achieve the highest accuracy for predicting students' final learning performance.

The analysis of education data promotes the prediction of academic performance, the implementation of LA, and the identification and improvement of students' behavioral patterns and performance. These research fields focus on students and emphasize the role of contemporary technologies in improving their learning experience and performance (Chen et al., 2020). Nevertheless, most related researchers have extracted data from OCW, MOOCs, or LMS; very few have investigated students' learning behavioral data from ebook systems in particular.

2.2. Identification of students' learning patterns

Although applying digital learning platforms to support teaching has become common in institutions of higher education, it is still difficult for most stakeholders using these digital platforms in their classes to effectively and precisely follow how students actually interact with the given online learning materials, how they behave under a given learning activity, or how they adjust their learning behavioral patterns when engaging in certain learning activities. One of the emerging issues in the research fields of LA and EDM has been pointed to finding appropriate methods of extracting meaning from the education data (Chen et al., 2020). Most digital learning platforms do not automatically include the advanced tools required for applying LA or EDM approaches, and stakeholders have indicated that utilizing these tools is too complex as they have features that are well beyond the scope of what teachers require (Romero et al., 2016). Therefore, the need to employ new LA and EDM approaches and develop the corresponding tools for stakeholders to simply observe the behaviors and interaction patterns of students conducting certain online learning activities has been raised as a critical issue (Juhaňák, Zounek & Rohlíková, 2019). Recently, many studies have proposed different combinations of analytics methods that can be adopted to explore students' learning patterns using LA and EDM techniques in various educational settings.

Yang et al. (2019) applied the k-means clustering technique to explore the learning patterns of 1,326 undergraduate students, utilizing eight learning features that were extracted from the students' interactions with an ebook system. In the study, five subgroups of students are identified based on the eight online learning features. The differences between the subgroups in terms of their learning behaviors and learning outcomes were reported.

Similarly, in the context of using an ebook system for learning support, Yin et al. (2019) apply LA and EDM approaches in an undergraduate course to discover how learning patterns differ across 98 students taking online courses using an ebook system. In their study, based on the results of the k-means clustering technique, four learning patterns were identified, representing the four patterns of learning with the ebook system. The authors named the patterns according to groups as follows: "preview and diligent group," "diligent group," "efficient group," and "poor performance group." Moreover, the differences between the identified learning patterns of the students in terms of their learning outcomes and the distribution of each learning variable were investigated and discussed. Finally, the correlation between students' online learning variables and their learning outcomes are presented.

In correlating the education data with the students' self-reported data and learning patterns, Maldonado-Mahauad et al. (2018) present a bottom-up approach that mines students' behavioral patterns (process mining and clustering) using the traditional top–down approach that utilizes the validated self-reported measurement (a self-regulated learning questionnaire). In their study, three learning patterns were identified from the sequential behaviors of 3,458 online learners based on their interactions with a MOOC lecture: "sampling learners," "comprehensive learners," and "targeting learners." Furthermore, their study investigated the differences between the identified learning patterns of students in terms of their self-regulated learning (SRL) profiles and the learning outcomes of the lecture.

Compared with the literature discussed herein, the current study focused on identifying the subgroups of students with different learning patterns when learning using an ebook system in an undergraduate course. Moreover, the association between the identified subgroups of students with different patterns of learning and their learning outcomes from the course when learning with an ebook system were explored and discussed.

3. Study context and data collection

3.1. Context of the exploratory study and participants

The data used in the present study were collected from an undergraduate course called Accounting Information Systems taught at a university. A total of 113 undergraduate students enrolled in the course. These students were from the Department of Accounting. To support the instructor and students, an ebook system called BookRoll developed by Ogata et al. (2015), was used. Students enrolled in this course were allowed to study the learning material using the BookRoll system at any time. Students' online learning logs, created while interacting with BookRoll, were tracked and recorded in a database.

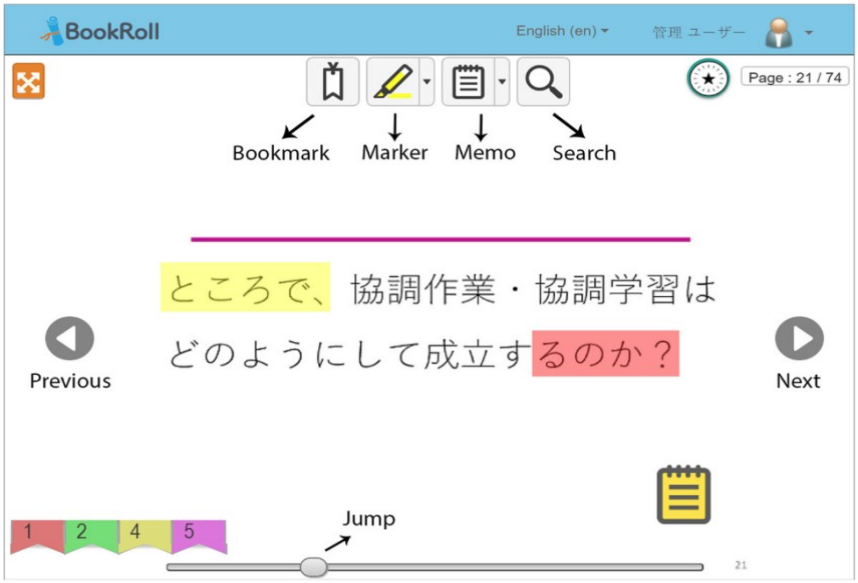

Figure 1. An example of the user interface of BookRoll.

Table 1. Example of the BookRoll learning behavioral data collected in this study

User_ID	Content_ID	Operation_Name	Operation_Date
15910	ed645f3821e	OPEN	2019/3/3 10:03:52
15910	ed645f3821e	ADD MEMO	2019/3/3 10:04:32
15923	ed645f3821e	OPEN	2019/3/3 10:07:03
15926	ed645f3821e	OPEN	2019/3/3 11:27:14
15926	ed645f3821e	NEXT	2019/3/3 11:27:20

Table 2. Number of occurrences of each BookRoll learning behavior

Learning behavioral data	Number of occurrences
OPEN	2,029
NEXT	35,988
PREV	10,167
ADD MARKER	9,125
ADD MEMO	7,284
ADD BOOKMARK	1,429
DELETE MARKER	659
DELETE MEMO	749
DELETE BOOKMARK	101
CLOSE	1,375
Total	68,906

An example of the user interface of BookRoll is displayed in Figure 1. An example of the students' learning logs collected in BookRoll is presented in Table 1. BookRoll allows users to browse the digital learning materials at any time and place after they are uploaded. Several functions are available such as page turning, marker drawing, memo taking, and page jumping. Data on the students' learning behaviors while using BookRoll are stored on its internal database. For the present study, students enrolled in the course were encouraged to use BookRoll to freely browse the digital learning material during various periods of learning, including in-class learning and out-

of-class learning. The descriptions of the functions of BookRoll are detailed in Ogata et al. (2015). For this study, 68,906 ebook learning logs from 113 undergraduate students were collected using BookRoll, and the number of occurrences of each BookRoll learning behavior is given in Table 2.

3.2. Indicators

In this study, several indicators were extracted to measure the students' learning behaviors based on their uses of the ebook system BookRoll as well as their learning outcomes from the course. The collected learning behavioral data from a total of 113 undergraduate students using BookRoll is described in Table 3. Moreover, a final exam was conducted to measure the students' learning outcomes from the course. The final exam scores were determined by the course instructor at the end of the course. The collection and analysis of the indicators used to measure the students' ebook learning behaviors and their learning outcomes are detailed as follows. Indicators of learning behavior and learning outcomes from the course:

- Backtrack reading rate (BRR): Students' BRR has been proven to positively correlate with their learning outcomes from the course (Yin et al., 2019). In this study, to analyze the students' BRR, the learning behavioral data NEXT and PREV were used together to measure the BRR for the course materials throughout the course. BRR was defined as the total number of times the student reviewed the previous page divided by the total number of times the student advanced to the next page throughout the course (# of PREV /# of NEXT). For example, a BRR value of 0.5 would correspond to a case where a student reviewed the previous page 50 times and advanced to the next page 100 times in one lecture throughout the course.
- Reading time (RT): The students' RT spent has been proven in the relevant literature to positively correlate with their learning outcomes and can be used to effectively identify subgroups of students learning with an ebook system (Yin et al., 2019; Yang et al., 2019). In this study, to analyze the students' RT over an hour, time-stamp data was collected and summed up to measure students' RT with ebook learning materials throughout the course.
- Adding annotation (AN): It is suggested that the use of annotations can facilitate students' learning activities during the course by providing support through directing attention and building both internal and external connections (Du, 2004). Accordingly, Yeh and Lo (2009) demonstrate a positive correlation between the use of annotations and students' academic performance, as the students who learned with an online annotation system that allowed them to add and delete online annotations achieved better academic performance in the course than students who did not. In this study, in order to analyze the students' behaviors in terms of AN, the total volume of learning behavioral data, ADD MARKER, ADD MEMO, and ADD BOOKMARK, was summed together to measure the students' behaviors in AN related to the ebook learning materials throughout the course.
- Deleting annotation (D-AN): Yeh and Lo (2009) demonstrated a positive correlation between the tendency to delete annotations and students' academic performance in a course that was part of the same experiment as the AN study. In the present study, to analyze the students' behaviors in terms of D-AN, the total volume of learning behavioral data, DELETE MARKER, DELETE MEMO, and DELETE BOOKMARK, was summed together to measure the students' behaviors in D-AN on the ebook learning materials throughout the course.
- Learning outcome (LO): In school education, students' learning outcomes in the course or academic performances have been reported to strongly correlate with their learning engagement (Yang et al., 2020). The students' scores in the final exam issued by the course instructor were recorded at the end of the experiment. The exam comprised 40 multiple-choice items, with a perfect score of 100. Moreover, the Kuder-Richardson Formula 20 value was 0.61, showing acceptable internal consistency of the final exam (Cortina, 1993).

Table 3. Description of the collected learning behavioral data from BookRoll

Learning behavioral data	Description of the learning behavioral data
ADD MARKER	Students added a marker on the ebook learning material
ADD MEMO	Students added a memo to the ebook learning material
ADD BOOKMARK	Students added a bookmark to the ebook learning material
DELETE MARKER	Students deleted a marker on the ebook learning material
DELETE MEMO	Students deleted a memo from the ebook learning material
DELETE BOOKMARK	Students deleted a bookmark from the ebook learning material
NEXT	Students advanced to the next page of the ebook learning material
PREV	Students returned to the previous page of the ebook learning material

3.3. Clustering analysis

To analyze the ebook learning logs collected from 113 undergraduate students in the present study, we followed the analysis procedure illustrated in Figure 2. After collecting the students' learning behavioral data and extracting the indicators of learning behavior, the authors first addressed the common scale of the dataset values. All the learning behavioral data were previously normalized to a value in the range of [0, 1] by min-max normalization for further clustering analysis. Nevertheless, raw data (before data normalization) were used when presenting the analysis results detailed in the following sections.

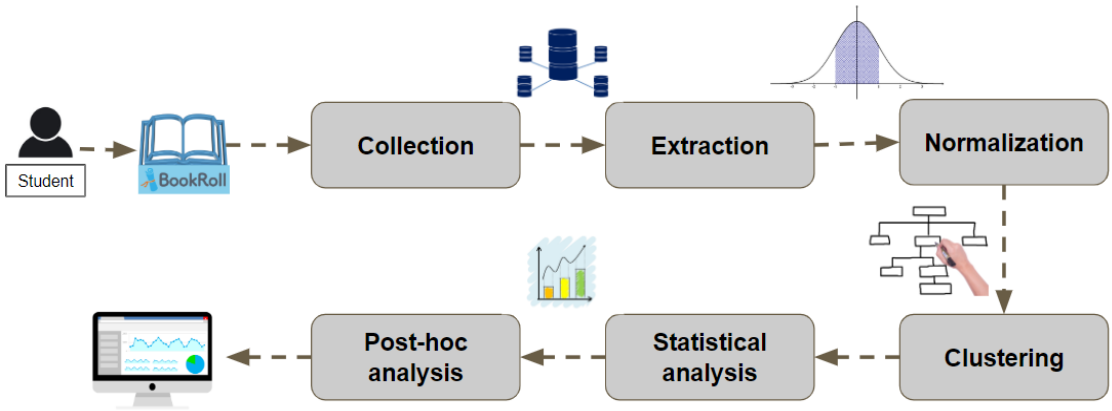

Figure 2. Analysis procedure

Next, to better understand how many subgroups of students with different learning patterns when learning with BookRoll can be identified using the four indicators of learning behavior as well as the distribution of each learning behavior between those subgroups, an agglomerative hierarchical clustering method based on Ward's method was applied. This clustering technique is recommended for identifying student subgroups in a given online learning context (Kovanović et al., 2015). The result of the cluster dendrogram displayed in Figure 3 led to the selection of three subgroups of students as the best options in this study. Thus, the students were categorized into three subgroups, which were labeled as Comprehensive learning group (CLG), Reflective learning group (RLG), and Selective learning group (SLG), based on the distribution of the four indicators of learning behavior.

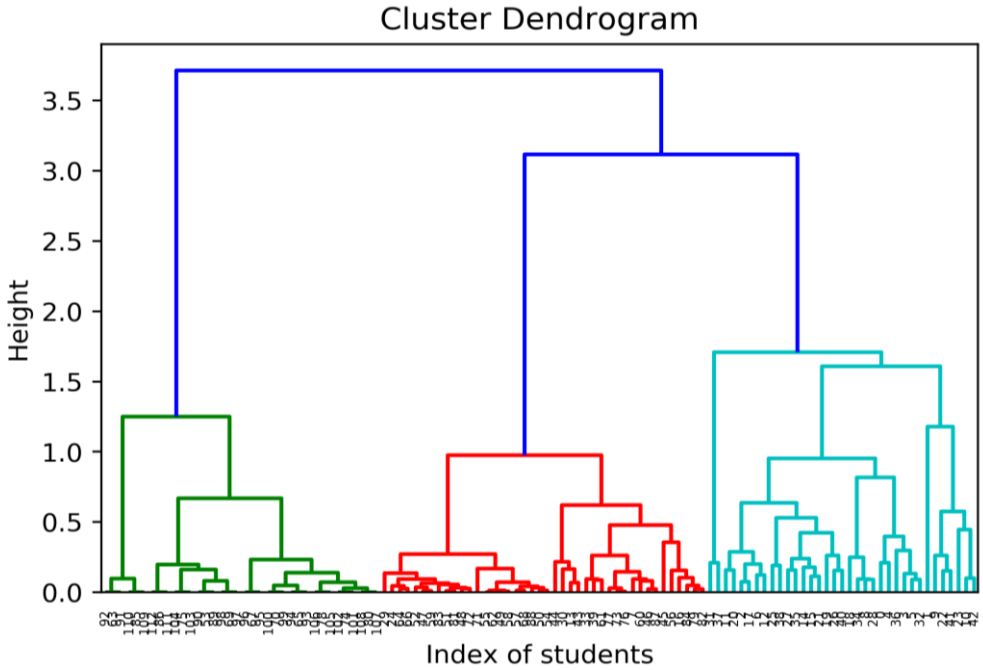

Figure 3. Dendrogram plot of clustering

4. Results

4.1. Analysis of the four indicators of learning behavior between the identified subgroups of students

After the agglomerative hierarchical clustering method was applied, the Kruskal–Wallis test was conducted to compare the identified subgroups of students in terms of the four indicators of learning behavior because the values did not have homogeneity of variance assumptions, which is required for a parametric test. Moreover, a nonparametric post-hoc test was performed using the Mann–Whitney U test to verify whether the difference between each pair of subgroups of students was statistically significant. The Kruskal-Wallis and Mann-Whitney tests have been evident with the statistical performances when the scale of data is nonparametric as well as the possibility to be applied for data analytic approaches in other face-to-face and on-line courses (Ahammed & Smith, 2019). The increasing applications of Kruskal-Wallis and Mann-Whitney U tests were identified in several research fields such as engineering applications, medicine, biology, psychology, and education (Ostertagová, Ostertag & Kováč, 2014).

According to the results listed in Table 4, significant differences were observed between the three identified subgroups of students in terms of BRR ($H = 76.87$, $p < .001$), RT ($H = 66.27$, $p < .001$), AN ($H = 73.04$, $p < .001$), and D-AN ($H = 23.81$, $p < .001$). The three identified subgroups of students are recorded as follows:

Table 4. Kruskal–Wallis test and post-hoc test (Mann–Whitney U) results of the learning behaviors between the (a) CLG, (b) RLG, and (c) SLG subgroup: median (25th percentile, 75th percentile)

	CLG ($n = 35$) (a)	RLG ($n = 36$) (b)	SLG ($n = 42$) (c)	H	Post-hoc test (Mann-Whitney U)
BRR	0.26 (0.17, 0.33)	0.61 (0.53, 0.72)	0.11 (0.05, 0.22)	76.87^{***}	a > c, b > a, b > c
RT	39.67 (27.61, 58.95)	14.52 (13.11, 15.29)	14.24 (12.56, 20.53)	66.27^{***}	a > b, a > c
AN	61.5 (47.45, 75.5)	16 (11.75, 24.25)	24 (13.25, 56.75)	73.04^{***}	a > b, a > c, c > b
D-AN	13 (4, 20.5)	4 (0.75, 5.75)	2 (1, 4)	23.81^{***}	a > b, a > c, b > c

Note. $^{***}p < .001$.

Comprehensive learning group (CLG): Students in this subgroup were categorized as CLG students ($N = 35$; 30.97% of the students) as the students in this subgroup exhibited the highest values for most of the indicators of learning behavior, representing a comprehensive way of learning as identified in Maldonado-Mahauad et al. (2018). For the BRR, CLG students exhibited higher BRR values (median = 0.26, 25th percentile = 0.17, 75th percentile = 0.33) compared with Selective learning group (SLG) students (median = 0.11, 25th percentile = 0.05, 75th percentile = 0.22). This implies that CLG students were more engaged in terms of BRR than SLG students at a statistically significant level. For RT, CLG students exhibited higher RT values (median = 39.67, 25th percentile = 27.61, 75th percentile = 58.95) compared with the Reflective learning group (RLG) students (median = 14.52, 25th percentile = 13.11, 75th percentile = 15.29) and SLG students (median = 14.24, 25th percentile = 12.56, 75th percentile = 20.53). This implies that CLG students spent more RT on learning the ebook learning materials than RLG and SLG students at a statistically significant level. For AN, CLG students exhibited higher values of AN (median = 61.5, 25th percentile = 47.45, 75th percentile = 75.5) compared with RLG students (median = 16, 25th percentile = 11.75, 75th percentile = 24.25) and SLG (median = 24, 25th percentile = 13.25, 75th percentile = 56.75). This implies that CLG students added more annotations to the ebook learning materials than RLG and SLG students at a statistically significant level. For D-AN, CLG students exhibited higher values of D-AN (median = 13, 25th percentile = 4, 75th percentile = 20.5) compared with RLG students (median = 4, 25th percentile = 0.75, 75th percentile = 5.75) and SLG students (median = 2, 25th percentile = 1, 75th percentile = 4). This implies that CLG students deleted more annotations from the ebook learning materials than RLG and SLG students at a statistically significant level.

Reflective learning group (RLG): Students in this subgroup were categorized as RLG students ($N = 36$; 31.86% of the students) as the students in this subgroup exhibited the highest values of BRR and higher values of D-AN compared with SLG students, representing a reflective way of learning as identified in Brinton et al. (2016). For the BRR, RLG students exhibited higher BRR values (median = 0.61, 25th percentile = 0.53, 75th percentile = 0.72) than CLG students (median = 0.26, 25th percentile = 0.17, 75th percentile = 0.33) and SLG students (median = 0.11, 25th percentile = 0.05, 75th percentile = 0.22). This implies that RLG students were more engaged in terms of BRR than CLG and SLG students at a statistically significant level. For D-AN, RLG students exhibited higher values of D-AN (median = 4, 25th percentile = 0.75, 75th percentile = 5.75) compared with SLG students (median = 2, 25th percentile = 1, 75th percentile = 4). This implies that RLG students deleted more annotations from the ebook learning materials than SLG students did at a statistically significant level.

Selective learning group (SLG): Students in this subgroup were categorized as SLG students ($N = 42$; 37.17% of the students) as the students in this subgroup exhibited higher values of AN compared with RLG students but exhibited the lowest values of most of the indicators of learning behavior on the other side. This represents a selective way of learning as identified in Maldonado-Mahauad et al. (2018). For AN, the SLG students had higher values of AN (median = 24, 25th percentile = 13.25, 75th percentile = 56.75) compared with RLG students (median = 16, 25th percentile = 11.75, 75th percentile = 24.25). This implies that SLG students added more annotations to the ebook learning materials than did RLG students at a statistically significant level.

4.2. Analysis of learning outcomes between the identified subgroups of students

To investigate the association between the identified subgroups of students with different patterns of learning and their learning outcomes from the course, the Kruskal–Wallis test was conducted to compare the subgroups of students in terms of their learning outcomes from the course as the values did not satisfy homogeneity of variance assumptions, which is required for a parametric test. According to the results listed in Table 5, a significant difference was observed between the three subgroups of students in terms of their learning outcomes from the course (LO; $H = 14.32$, $p < .001$). Moreover, a nonparametric post-hoc test was performed using the Mann–Whitney U test to verify whether the difference between each pair of subgroups of students was statistically significant.

The CLG students had higher values of LO (median = 83, 25th percentile = 78, 75th percentile = 85) than the SLG students did (median = 76, 25th percentile = 73, 75th percentile = 84), indicating that CLG students obtained significantly higher scores in the final exam than did SLG students. This implies that CLG students achieved better learning outcomes from the course than SLG students at a statistically significant level. Furthermore, RLG students had higher LO values (median = 84, 25th percentile = 82, 75th percentile = 86) than the SLG students (median = 76, 25th percentile = 73, 75th percentile = 84), indicating that the RLG students obtained significantly higher scores in the final exam than did SLG students. This implies that RLG students achieved better learning outcomes from the course than did SLG students at a statistically significant level.

Table 5. Kruskal–Wallis test and post-hoc test (Mann–Whitney U) results of the learning outcomes from the course between the (a) CLG, (b) RLG, and (c) SLG subgroup

Student subgroup	*N*	Median	25th percentiles	75th percentiles	*H*	Post-hoc test (Mann-Whitney U)
CLG (a)	35	83	78.13	85.85	14.32^{***}	a > c
RLG (b)	36	84.3	82.05	86.85		b > c
SLG (c)	42	76.25	73.03	84.05		

Note. $^{***}p < .001$.

5. Discussion and practical implications

5.1. Identifying subgroups of students with different patterns of learning using an ebook system

In addressing the first research question, the current authors used an agglomerative hierarchical clustering approach based on Ward's method to identify subgroups of students with different learning patterns using an ebook system. In this study, three subgroups of students with different learning patterns were identified. To better understand the difference in the learning patterns between the subgroups of students, the authors considered the difference in each indicator of learning behavior between the subgroups using Kruskal–Wallis and Mann–Whitney U tests.

The results of these tests provide evidence that the CLG students tended to apply a comprehensive learning approach when learning with an ebook system as they showed the highest tendency among almost all the indicators of learning behavior between the three subgroups, which represents a more comprehensive and engaged learning approach that combines several learning strategies when learning with an ebook system. This finding is consistent with those of Maldonado-Mahauad et al. (2018), where the students who adopted a comprehensive learning approach usually utilized a combination of learning strategies.

RLG students tended to apply a reflective learning approach when using an ebook system as they did not spend as much RT as the CLG students did. Furthermore, RLG students did not add as many annotations to the ebook learning materials as the CLG and SLG students did. Instead, the RLG students had the highest BRR values of

the three subgroups and had a higher tendency to delete annotations from ebook learning materials compared with SLG students, which represents a reflective learning strategy for an ebook system. This finding is consistent with previous studies where one subgroup of students had a higher BRR as they tended to frequently return to the previous page of the ebook learning material for review instead of going ahead to the next page (Yin et al., 2019) but with a lower tendency to add annotations to ebook learning materials (Yang et al., 2019) when learning with an ebook system. Another finding in this study demonstrates that this learning strategy also led to a higher tendency to delete annotations from ebook learning materials as the students in this subgroup typically reflected on the annotations they had made previously after returning to the previous page of the ebook learning material. This finding is consistent with the relevant literature, where the learning behavior of backtrack reading is usually connected to a reflection learning strategy of associating current knowledge with previous knowledge (Costa & Kallick, 2008).

SLG students tended to apply a selective learning approach when using an ebook system as they did not have a BRR as high as that of CLG and RLG students. Furthermore, the SLG students did not spend as much RT as the CLG students did or delete as many annotations from the ebook learning materials as the CLG and RLG students. Instead, the SLG students showed a higher tendency to add annotations to ebook learning materials compared with RLG students, which represents a selective learning strategy for using an ebook system. This finding is consistent with that of a previous study; a subgroup of students had a higher tendency to add annotations to ebook learning materials when learning with an ebook system (Yang et al., 2019).

The findings of this study in identifying three subgroups of students provide insights into the true ebook learning patterns of students. The results of the clustering analysis are consistent with the relevant literature; students can be categorized into several subgroups according to their patterns of learning with ebook systems, such as BRR, annotation usage, and RT (Yin et al., 2019; Yang et al., 2019). The students were therefore provided opportunities to accurately identify and describe how they studied and what strategies they applied to previous learning activities after receiving this information from either educators or digital learning platforms. This issue is highlighted in the literature (Zhou & Winne, 2012).

5.2. Association between the identified subgroups of students with different patterns of learning and their learning outcomes from the course when learning with an ebook system

In addressing the second research question and better understanding the association between the identified subgroups of students with different patterns of learning and their learning outcomes from the course, the current authors considered the difference in the students' learning outcomes from the course across the three subgroups of students, again using the Kruskal–Wallis and Mann–Whitney U tests.

The results of these tests demonstrate that the students in CLG achieved better learning outcomes from the course than the students in SLG at a statistically significant level, as the final exam scores of the students in CLG were significantly higher than those of the students in SLG. Therefore, it is evident that the students who adopted a comprehensive learning approach that combined several types of learning strategies achieved better learning outcomes from the course than those who adopted a selective learning approach at a statistically significant level. This finding is consistent with the relevant literature, where students who follow a comprehensive or deep learning approach achieve stronger academic performance (Bliuc et al., 2010; Ellis et al., 2008), whereas students who adopt a surface or selective learning approach achieve poorer academic performance, as such learning approach was negatively correlated with their academic performance (Richardson, Abraham & Bond, 2012).

Moreover, the results demonstrate that the RLG students achieved better learning outcomes from the course than the SLG students at a statistically significant level, as the final exam scores of the RLG students were significantly higher than those of the SLG students. Therefore, students who adopted a reflective learning approach and exhibited a higher BRR and tendency to delete annotations from ebook learning materials as well as a lower tendency to add annotations to ebook learning materials can achieve better learning outcomes from the course than students who adopted selective learning approach that involves a higher tendency to add annotations to ebook learning materials and a lower BRR and tendency to delete annotations from ebook learning materials, at a statistically significant level. This finding is consistent with the literature; students who exhibited a higher BRR achieved better learning outcomes than students with a lower BRR when using ebook systems (Yin et al., 2019). This finding can also be connected to the relevant literature; backtrack reading was positively correlated to a review learning strategy of allotting time to commit knowledge from the learning

materials to students' long-term memory (Lindsey et al., 2014), thereby leading to improved student learning outcomes from the course, as proven in the current study.

The findings noted herein provide insight into the association between students' learning patterns and their learning outcomes from the course when learning with an ebook system. The researchers and educators are therefore provided opportunities to predict students' learning outcomes by analyzing their online learning behaviors and providing timely intervention for improving their learning experience, which achieves one of the goals of learning analytics as part of precision education (Lu et al., 2018). Moreover, the teachers at every education level are provided opportunities to apply the experimental results to serve as a basis for adjusting their teaching strategies or materials for achieving personalized learning in the course. Furthermore, the students are provided opportunities to adjust the learning strategies they had previously adopted to better address changing learning situations and learning goals on receiving this information from either educators or digital learning platforms; this issue has been highlighted in the previous literature (Lust, Elen & Clarebout, 2013).

5.3. Limitations

This study has several limitations that must be considered. First, the sample size of this study was 113 students. The results, although significant, may not be generalized to students from other institutions of higher education. A more general analytical model is required to suit students from different institutions using the same analytics method. Next, this study focused exclusively on identifying subgroups of students with different learning patterns using ebook learning logs. Consequently, this highlights a limitation in the types of student learning behaviors that can be identified when applying the LA and EDM approaches. Therefore, various digital learning platforms can be integrated to obtain a wider range of student learning behaviors when applying similar analytic methods in future studies. Moreover, a clustering approach was applied once to identify subgroups of students with different learning patterns. It is possible to combine the clustering approach with other techniques such as process mining or sequential data mining, which may be expected to provide deeper insights into students' learning patterns regarding an ebook system.

6. Conclusion

The rising volume of students' learning behavioral data gathered by virtual learning environments provides opportunities for mining students' patterns of learning (Yang et al., 2019). Consequently, the associations and patterns between students' learning behaviors and learning outcomes can be used to trigger a learning process and thereby reach specific goals for both learning and teaching (Reimann, 2016). However, many learning patterns and strategies cannot be easily identified from the system log data without the application of data analytics methods or data mining techniques. Hence, without a series of data analytics methods for the identification of students' learning strategies and their learning patterns using LA and EDM approaches, students often have difficulty adjusting the learning strategies they previously adopted to better address changing learning situations (Lust, Elen & Clarebout, 2013).

Thus, in this study, data analytic methods for examining the learning patterns of students while learning with an ebook system were applied. An agglomerative hierarchical clustering approach was applied to identify subgroups of students with different learning patterns when learning with an ebook system in an undergraduate course for one semester. Several subgroups were identified and categorized as Comprehensive learning group (CLG), Reflective learning group (RLG), and Selective learning group (SLG) based on the different learning strategies the students adopted, such as the higher/lower values of BRR and RT and a tendency toward AN and D-AN across ebook learning materials. Moreover, the association between the subgroups of students with different patterns of learning and their learning outcomes from the course was investigated in this study. It is therefore suggested that by applying the combined LA and EDM approaches to identify and analyze the subgroups of students learning with an ebook system, the instructor may have an opportunity to not only to improve their method of teaching during the course but also to support students in taking suitable actions with recommendations to achieve particular learning goals based on the information derived from the learning patterns of students. Moreover, the authors hope that this study will motivate other researchers to use LA and EDM approaches more often and explore their possibilities for future research. The findings of this study provide not only a detailed demonstration of applying a series of data analytics methods to identify subgroups of students with different patterns of learning throughout a semester in an undergraduate course but also offer insights into the association between students' learning patterns and outcomes from the course when learning with an ebook system, which are expected to make contributions on facilitating the practices of precision education.

Acknowledgement

This work was partially supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

References

Ahammed, F., & Smith, E. (2019). Prediction of students' performances using course analytics data: A Case of water engineering course at the university of south Australia. *Education Sciences*, *9*(3), 245. doi:10.3390/educsci9030245

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). New York, NY: Springer.

Bliuc, A. M., Ellis, R., Goodyear, P., & Piggott, L. (2010). Learning through face-to-face and online discussions: Associations between students' conceptions, approaches and academic performance in political science. *British Journal of Educational Technology*, *41*(3), 512-524.

Brinton, C. G., Buccapatnam, S., Chiang, M., & Poor, H. V. (2016). Mining MOOC clickstreams: Video-watching behavior vs. in-video quiz performance. *IEEE Transactions on Signal Processing*, *64*(14), 3677-3692.

Chen, N. S., Yin, C., Isaias, P., & Psotka, J. (2020). Educational big data: extracting meaning from data for smart education. *Interactive Learning Environments, 28*(2), 142-147. doi:10.1080/10494820.2019.1635395

Colvin, C., Rogers, T., Wade, A., Dawson, S., Gašević, D., Buckingham Shum, S., Karen J, N., Shirley, A., Lori, L., Gregor, K., Linda, C., & Fisher, J. (2015). *Student retention and learning analytics: A Snapshot of Australian practices and a framework for advancement.* Canberra, Australia: Office of Learning and Teaching, Australian Government.

Corrin, L., & De Barba, P. (2015). How do students interpret feedback delivered via dashboards? In *Proceedings of the fifth international conference on learning analytics and knowledge* (pp. 430-431). New York, NY: ACM.

Cortina, J. M. (1993). What is coefficient alpha? An Examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Costa, A. L., & Kallick, B. (Eds.). (2008). *Learning and leading with habits of mind: 16 essential characteristics for success*. Alexandria, VA: ASCD.

Du, M. C. (2004). *Personalized annotation management for web based learning service* (Unpublished master thesis). National Central University, Chungli, Taiwan.

Ellis, R. A., Goodyear, P., Calvo, R. A., & Prosser, M. (2008). Engineering students' conceptions of and approaches to learning through discussions in face-to-face and online contexts. *Learning and Instruction*, *18*(3), 267-282.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The Effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68-84.

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A Generic framework for learning analytics. *Educational Technology & Society*, *15*(3), 42-57.

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469-478.

Hwang, G. J. (2014). Definition, framework and research issues of smart learning environments-a context-aware ubiquitous learning perspective. *Smart Learning Environments, 1*(1), 4. doi:10.1186/s40561-014-0004-5

Hwang, G. J., Chu, H. C., & Yin, C. (2017). Objectives, methodologies and research issues of learning analytics. *Interactive Learning Environments, 25*(2), 143-146.

Juhaňák, L., Zounek, J., & Rohlíková, L. (2019). Using process mining to analyze students' quiz-taking behavior patterns in a learning management system. *Computers in Human Behavior*, *92*, 496-506.

Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *27*, 74-89.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). A Learning management system-based early warning system for academic advising in undergraduate engineering. *Learning analytics* (pp. 103-119). New York, NY: Springer.

Kumar, K., & Kumar, V. (2018). Advancing learning through smart learning analytics: A Review of case studies. *Asian Association of Open Universities Journal*, *13*(1), 1-12.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological science*, *25*(3), 639-647.

Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society*, *21*(2), 220-232.

Lust, G., Elen, J., & Clarebout, G. (2013). Regulation of tool-use within a blended course: Student differences and performance effects. *Computers & Education*, *60*(1), 385-395.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior*, *80*, 179-196.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In *International Conference on Computer in Education (ICCE 2015)* (pp. 401-406). Hangzhou, China: Asia-Pacific Society for Computers in Education.

Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal-Wallis test. In *Applied Mechanics and Materials, 611,* 115–120. doi:10.4028/www.scientific.net/amm.611.115

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A Systematic literature review of empirical evidence. *Educational Technology & Society*, *17*(4), 49-64.

Pardo, A., Han, F., & Ellis, R. A. (2016). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, *10*(1), 82-92.

Reimann, P. (2016). Connecting learning analytics with learning research: The Role of design-based research. *Learning: Research and Practice*, *2*(2), 130-142.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A Systematic review and meta-analysis. *Psychological bulletin*, *138*(2), 353-387.

Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: A Tutorial and case study using moodle data sets. *Data mining and learning analytics: Applications in educational research*, 1-28.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30.

Winne, P. H., & Jamieson-Noel, D. (2003). Self-regulating studying by objectives for learning: Students' reports compared to a model. *Contemporary Educational Psychology*, *28*(3), 259-276.

Wong, B. T. M. (2017). Learning analytics in higher education: An Analysis of case studies. *Asian Association of Open Universities Journal*, *12*(1), 21-40.

Yang, C. C., Chen, I. Y., Huang, A. Y., Lin, Q. R., & Ogata, H. (2020). Can self-regulated learning intervention improve student reading performance in flipped classrooms? *International Journal of Online Pedagogy and Course Design (IJOPCD)*, *10*(4), 1-13.

Yang, C. C., Flanagan, B., Akçapınar, G., & Ogata, H. (2019). Investigating subpopulation of students in digital textbook reading logs by clustering. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge (LAK'19)* (pp. 465-470). Tempe, AZ: Society for Learning Analytics Research (SoLAR).

Yang, S. J. H. (2019). Precision education: New challenges for AI in education [conference keynote]. In *Proceedings of the 27th International Conference on Computers in Education (ICCE)* (pp. XXVII-XXVIII). Kenting, Taiwan: Asia-Pacific Society for Computers in Education (APSCE).

Yeh, S. W., & Lo, J. J. (2009). Using online annotations to support error correction and corrective feedback. *Computers & Education*, *52*(4), 882-892.

Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K., & Ogata, H. (2019). Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A Human factor approach. *International Journal of Human–Computer Interaction*, *35*(4-5), 313-322.

Zhou, M., & Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. *Learning and Instruction*, *22*(6), 413-419.