Internet and Higher Education 30 (2016) 30–43

Contents lists available at ScienceDirect

Internet and Higher Education

Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments

Dongho Kim [a], Yeonjeong Park [b,c], Meehyun Yoon [a], Il-Hyun Jo [b,c,*]

[a] Department of Career and Information Studies, 850 College Station Rd, University of Georgia, GA, United States
[b] Center for Teaching and Learning, 417 Eodeung Rd, Honam University, Gwangju, Republic of Korea
[c] Department of Educational Technology, 52 Ewha Rd, Ewha Womans University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 29 October 2015
Received in revised form 23 March 2016
Accepted 27 March 2016
Available online 2 April 2016

Keywords:
Learning analytics
Asynchronous online discussion
Proxy variable
Educational data mining
Learner online behavior

ABSTRACT

Although asynchronous online discussion (AOD) is increasingly used as a main activity for blended learning, many students find it difficult to engage in discussions and report low achievement. Early prediction and timely intervention can help potential low achievers get back on track as early as possible. This study presented a data mining process to construct proxy variables that reflect theoretical and empirical evidence and measured the accuracy of a prediction model that incorporated all of the variables for validation. For the empirical study, data were obtained from 105 university students who were enrolled in two blended learning courses that used AOD as their main activity. The results indicated the high accuracy of the prediction model as well as the possibility of early detection and timely interventions. In addition, we examined participants' learning behaviors in the two courses using the proxy variables and provided suggestions for practice. The implications of this study for education data mining and learning analytics are discussed.

1. Introduction

Blended learning is increasingly adopted in higher education due to its flexibility (Irvine, Code, & Richards, 2013). Blended learning has benefitted from using both synchronous and asynchronous delivery modes for its online portion to enhance students' access and engagement. Synchronous delivery modes use a two-way, real-time technologies such as videoconferencing system to support learning that takes place between two or more people at the same time (Butz & Stupnisky, 2016; Hrastinski, 2008). Synchronous learning has been lauded for its engaging nature that helps to reduce students' feeling of isolation in online learning environments. In contrast, asynchronous delivery modes are commonly facilitated by technologies that allow learners to engage in learning at any time. Asynchronous learning makes learners more flexible in terms of reflecting on learning content and refining their contributions (Hrastinski, 2008; Vignare, 2007).

Asynchronous online discussion (AOD) is a popular form of asynchronous learning used to support critical discussion and interaction among individual learners (Hew, Cheung, & Ng, 2010; Liu, Magjuka, Bonk, & Lee, 2007; Loncar, Barrett, & Liu, 2014). AOD offers learners the freedom to exhibit their own learning style without constraints of time and space (Berge & Collins, 1995). Empirical evidence of the effect of AOD has been reported in research on blended learning (Vignare, 2007). Researchers and practitioners acknowledge the potential of AOD to promote social interaction and reflection while allowing for a better understanding of contents (Andresen, 2009; Thomas, 2002). AOD is also recognized as a means to promote a sense of community in online courses in that it facilitates "information sharing, idea exchanges, and mentoring" (Liu et al., 2007, p. 12). Learners, however, often encounter challenges in deeply engaging in discussion topics (Balaji & Chakrabarti, 2010; Mason, 2011) and sustaining course-related endeavors (Hew et al., 2010; Wise, Speer, Marbouti, & Hsiao, 2013). Many learners superficially participate in AOD without contributing to the discussions (Hernández-García, González-González, Jiménez-Zarco, & Chaparro-Peláez, 2015; Mason, 2011; Nagel, Blignaut, & Cronjé, 2009). Those challenges can result in shallow interactions and fragmented communications, which impede development of a sense of online community (Liu et al., 2007).

Data-mining techniques have been used to address these issues through examining students' online learning behaviors (Lara, Lizcano, Martínez, Pazos, & Riera, 2014). Online learning behaviors are often analyzed using student log data in the discpline of educational data mining and learning analytics. Log data provide information about how students participate in various online learning activities. We can obtain, for example, information about how much time students spent on a particular online activity. Despite some concerns about translating students' log data into their actual learning behaviors, a large body of literature has provided empirical evidence of strong correlation between them (Hung & Zhang, 2008; Jo, Kim, & Yoon, 2015; Lara et al., 2014; Romero, López, Luna, & Ventura, 2013). In addition to learners'

* Corresponding author.
E-mail addresses: dongho@uga.edu (D. Kim), ypark@honam.ac.kr (Y. Park), meehyun@uga.edu (M. Yoon), ijo@ewha.ac.kr (I.-H. Jo).

http://dx.doi.org/10.1016/j.iheduc.2016.03.002
1096-7516/

visible behaviors, their psychological characteristics, such as an interest in a particular topic, may be reflected in their log data (Woolf et al., 2009).

Given the enormous amount of data that is generated in online discussions with regard to learner participation (Loncar et al., 2014; Romero et al., 2013), automated examination is of benefit to those who teach a blended course that adopts AOD (Dringus & Ellis, 2005). If instructors could track the status of students' learning, it would be possible for them to implement timely interventions (Macfadyen & Dawson, 2010; Zacharis, 2015), which would lead to students' perseverance and the successful completion of online courses (Lykourentzou, Giannoukos, Nikolopoulos, Mpardis, & Loumos, 2009).

If students are identified as potential low achievers during a course, instructors can then encourage them to participate through the use of facilitative strategies (e.g. Hew et al., 2010; Hou, 2011), incentives (Gilbert & Dabbagh, 2005), and relevant learning materials (Yeh & Van Buskirk, 2005)

The current study aimed to propose proxy variables for different AOD settings, and the variables were constructed to reflect theoretical and empirical evidence from prior research. This solid basis can contribute to the high applicability and generalizability of the generated variables to address variations in AOD context settings (e.g., discussion types and interaction patterns), instruction strategies (Dykman & Davis, 2008; Hou, 2011; Schworm & Gruber, 2012), and student attributes (Duque, Gómez-Pérez, Nieto-Reyes, & Bravo, 2015; Hernández-García et al., 2015; Topcu & Ubuz, 2008)

The objectives of this study were as follows: (a) to present a data mining process for constructing proxy variables that represent the specific behavioral and psychological characteristics of high achievers in asynchronous online discussion; (b) to empirically validate the variables in two blended courses that adopt AOD as their main activity; and (c) to provide suggestions for practice.

In this study, the learning analytics approach was used to address the following questions:

1. How accurately can the proxy variables predict low and high achievers? and
2. What learning behaviors are observed in AOD through the lens of the proxy variables?

Prior to addressing the research questions, the following section describes the process of constructing the proxy variables as indicators of success in AOD environments.

2. Constructing the proxy variables

Proxy variables are those that are alternatively used when the direct measurement of conceptual variables is difficult because of access or feasibility (Jo et al., 2015; Wickens, 1972). The concept of proxy variables is often used in the field of social science to create prediction models (Durden & Ellis, 2003). Here, we developed a method for extracting such proxy variables to represent key factors that have been identified in the literature on AOD. Three steps were taken to transform the key factors into proxy variables (see Fig. 1). First, four AOD success factors were identified through an extensive literature review: *Active participation in the course*, *Engagement with discussion topics*, *Consistent effort and awareness*, *and Interaction*. Second, specific behavioral and psychological characteristics of high achievers (e.g., regular study and many postings) were identified for each factor based on prior research that explored indicators of high academic performance in AOD. Lastly, 10 proxy variables, each of which represents one of the four factors, were calculated. To allow for automated prediction, we included only quantitative variables that did not require manual assessment from instructors. For example, we purposefully did not analyze the structure of discussion threads because they were likely confined to particular topics. The remainder of this section provides details of how the proxy variables were constructed within the four key factors.

2.1. Active participation

Active participation is a crucial measure of student engagement, and it typically leads to high academic performance in online discussions

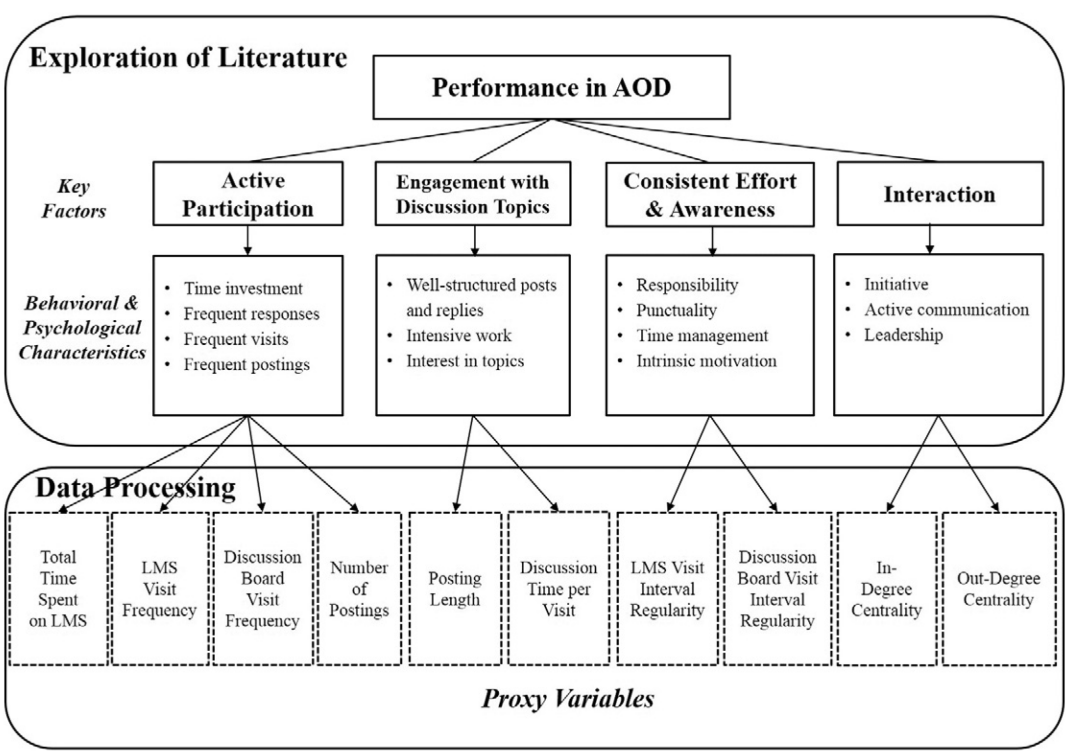


Fig. 1. Process of extracting proxy variables.

(Chan & Chan, 2011; Hrastinski, 2009; Mason, 2011; Shana, 2009; Wise et al., 2013). Previous research indicated that active participation is highly associated with meaningful discussion (Mason, 2011). Richardson and Ice (2010) reported that students who showed high levels of achievement actively participated in online discussions: "exploring the relevant information, and exchanging opinions, and responding to others' posts" (p. 54). Studies adopting a learning analytics approach also reported similar results. For example, Macfadyen and Dawson (2010) found that students' activeness was a crucial factor that directly affected their achievement. In that study, students' participation, represented by the numbers of postings, emails, and completed assessments, was revealed to predict students' final grades. Zacharis (2015) recognized the correlation between students' academic performance and active participation as measured by LMS login data such as the total time spent online and number of discussion messages read and posted.

2.1.1. Total time spent on learning management systems and discussion boards

Online discussion takes a large amount of time to explore the different learning resources as well as discussion topics (e.g., Richardson & Ice, 2010). Learners discuss topics to challenge each other's views and make logical arguments. For course activities, learners need to clarify topics and organize their thoughts with sufficient evidence (Gao, 2014). Furthermore, they are often required to internalize course content by reading course materials or other relevant resources before participating in discussions (Jung & Gilson, 2014; Lara et al., 2014). Thus, we recognized that student time spent on learning management systems (LMS) and discussion boards implies learners' readiness for discussions.

Learners usually do not show initiative at the early stages of online discussions, instead of moving toward meaningful learning over a sufficient amount of time as they become familiar with discussion topics (Richardson & Ice, 2010). The need for sufficient time in discussions is emphasized in numerous studies that investigated students' cognitive presence in online learning. For example, Garrison and Arbaugh (2007) argued that an online discussion progresses through multiple phases: identifying issues, exploring the issues, constructing meaning, and applying new ideas. The transition process necessitates students' time investment and cognitive engagement. Details on calculating reliable total times are provided in the data processing section below.

2.1.2. LMS and discussion board visit frequency

Lara et al. (2014) reported that the number of times someone accesses a virtual classroom is a strong indicator of their active participation in AOD. Particularly, the study noted that student participation, measured by the number of online classroom visits and interactions with resources, was positively correlated with students' perseverance in online courses.

It is assumed that the more frequently learners visit online systems, the more useful information they can obtain in a timely manner (Cruz-Benito, Therón, García-Peñalvo, & Lucas, 2015), and this assumption is supported by research that reported a positive relationship between active participation and number of visits (Hung & Zhang, 2008; Webb, Jones, Barker, & van Schaik, 2004). In addition, students' frequent visits suggest their need to obtain information, read others' posts, or leave posts themselves (Dennen, 2008). Whether students post or not, their visits to a course page or discussion board can be seen as a learning-relevant activity. Those variables were calculated using the total number of visits to the LMS and discussion boards.

2.1.3. Number of postings

Uploading postings is necessary for online discussions to be sustained. Actively participating learners engage in "producing a written contribution and inserting it into a topic and comment structure" (Topcu & Ubuz, 2008, p. 3). Prior studies provided empirical evidence that number of postings is a significant predictor of learner achievement. (Chan & Chan, 2011; Hung & Zhang, 2008; Topcu & Ubuz, 2008; Webb et al., 2004; Xie, 2013). For example, Webb et al. (2004) recognized number of postings as representing the degree of students' contribution to the productivity of discussions. It was also found to be significantly predictive of students' final grades. Macfadyen and Dawson (2010) found that number of postings was positively correlated not only with students' final grades but also with how their learning progressed. More postings were found to lead to more meaningful learning. Schellens and Valcke (2005), in a study that investigated the effects of AOD on student knowledge construction, reported that discussion groups that generated more postings achieved qualitatively greater knowledge gains. Despite empirical evidence of positive correlation between the number of postings and learner performance in AOD, some researchers suggested that only limited interpretation should be given to the correlation because the number does not uncover qualitative aspects of postings (Dringus & Ellis, 2005). Andresen (2009) also pointed out the limitation and stressed the importance of considering other variables such as cognitive engagement. In this regard, the current study just viewed this variable as one that constitutes the active participation factor rather than relating it to discussion quality. This variable was calculated for each student by counting each one's numbers of postings and replies.

2.2. Engagement with discussion topics

Although quantified time investments can serve as indicators of active participation, more factors are needed to explain productive discussions (Thomas, 2002). Many students do not meet expectations in terms of quality regardless of the amount of time they spend or how frequently they access online discussions (Nagel et al., 2009). Even the act of posting does not necessarily mean that students are engaged in discussion topics when they only do so to meet course requirements or teachers' expectations (Dennen, 2008). Deep learning in AOD is achieved when students deeply engage with discussion topics because discussion is an authentic task that requires an in-depth understanding of the topic (Mason, 2011; Wise et al., 2013). Low engagement with AOD discussion topics causes learners to have fragmented conceptions and incomplete knowledge about the topics (Bliuc, Ellis, Goodyear, & Piggott, 2010). A number of studies also recognize engagement as essential to knowledge construction (e.g., deNoyelles, Zydney, & chen, 2014; Richardson & Ice, 2010). For example, Schellens and Valcke (2005) observed that knowledge construction in AOD is fostered when students' cognitive processing is actively engaged. In line with this study, Shukor, Tasir, Van der Meijden, and Harun (2014) noted that students' high levels of knowledge construction involved high-level "evaluation type of comment[s]" (p. 219). Based on the evidence, the proxy variables in this study for student engagement with discussion topics indirectly measure learners' intensive time use and the amount of information contained in postings.

2.2.1. Posting length

Long postings typically imply that learners have invested a considerable amount of time in structuring and articulating their thoughts (Hewitt, Brett, & Peters, 2007). Length of postings was reported to correlate with engagement in discussions (Xie, 2013). This variable was constructed with the assumption that students' cognitive engagement is demonstrated to support their arguments with sufficient evidence. Prior studies emphasize that providing evidence in postings leads to high-quality discussions. For example, Shukor et al. (2014) noted that students exhibit cognitive engagement while "arguing, justifying, explaining, and providing evidence" during discussions (p, 225). In another study, Ekahitanond (2013) found that students who demonstrated higher levels of critical thinking provided more supporting reasons in their postings. Such cognitive activities result in betterelaborated postings; Hara, Bonk, and Angeli (2000) contended that the lengths of postings indicated students' engagement levels; in the study, longer messages were found to have more supporting evidence and implicit clues, consistent indicators of in-depth information processing. This variable was calculated by counting the total number of letters used for all postings made by each student divided by the total number of postings.

2.2.2. Discussion time per visit

This variable was proposed based on the assumption that students who spent longer amounts of time per visit would be more apt to participate in in-depth discussions because those who stayed longer had more time to read and write carefully. In fact, discussion visit duration is positively associated with interaction with posts, which can lead to reflective efforts to gain an in-depth understanding of discussion topics (Wise, Marbouti, Hsiao, & Hausknecht, 2012; Wise et al., 2013). Ekahitanond (2013) asserted that students invested time intensively in order to gain an in-depth understanding of what others were saying The study found that high achievers in online discussions "spent extended amounts of time examining each other's peer comments questions, or recommendations" (p. 260). Intensive work is also found when students upload postings. For example, Shukor et al. (2014) argued that knowledge construction is achieved through externalizing thoughts that involve arguments, justification, and decision making, which is assumed to result in more intensive time spent on discussion boards. Students are expected to work intensively so that their postings will have complete logical structures. This variable was calculated using the total time spent on a discussion board divided by the number of visits to the board.

2.3. Consistent effort and awareness

2.3.1. Visit interval regularity

Students can have enhanced learning experiences through regular participation in discussions (Andresen, 2009). Regular learning is known to be associated with consistent effort (Jo et al., 2015). Darabi, Liang, Suryavanshi, and Yurekli (2013) conducted a meta analysis of 72 prior studies to examine the effectiveness of discussion strategies. They identified regular interactions between an instructor and students as one of the most important indicators of success in online discussions. Regularity of learning also implies the awareness of learning in that learners' regular participation results from their purposefully planning to study. Successful learners are therefore expected to adopt a regular routine to attain academic achievement (Blaxter & Tight, 1994).

To measure the extent to which students showed consistent effort and awareness in the courses, the LMS and discussion board visit interval regularities were calculated using the standard deviations of the intervals (Jo et al., 2015). Therefore, these variables technically mean the irregularity of the intervals.

In Fig. 2, the length of the course is the gap between A and B, and $\Delta t_{1,2}$ is the interval between the login (visit) point t_1 and t_2. Likewise, all the other intervals were obtained. The mean of each learner's login (visit) interval was used to calculate the standard deviation. Because standard deviation is only a reflection of the central tendency of the login (visit) interval, the regularity can be the same with different login frequencies or login durations (Fig. 3).

2.4. Interaction

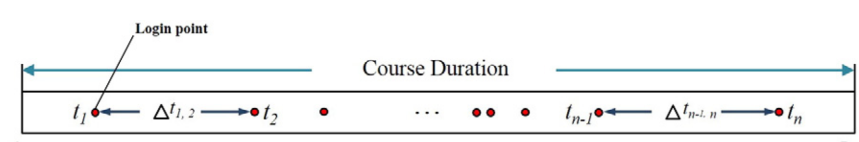
Students' interaction data are a strong indicator of how they connect with others through discussions (De Laat, Lally, Lipponen, & Simons, 2007). The absence of interaction data can often result in an incomplete understanding of students' participation in discussions (Macfadyen & Dawson, 2010). In order to generate proxy variables for quantifying student interactions, degree centrality was used as an indicator of the students' prominence and engagement in discussions (Dawson, Macfadyen, Lockyer, & Mazzochi-Jones, 2011). Centrality measures are useful to indicate communicative interactions that occur in computer supportive environments (Dawson, 2008). Furthermore, they are known to indicate the extent to which an individual learner is "active, prestigious, powerful, and visible" (Cho, Gay, Davidson, & Ingraffea, 2007, p. 6). Interactions are commonly analyzed using a set of actors or network members (Dawson, 2008). In this study, in-degree centrality and out-degree centrality were calculated by counting the ties each student had established with their peers (Dawson et al., 2011). An actor was a reply either given or received by a student during the discussions (see Fig. 4).

2.4.1. In-degree centrality

In-degree centrality was indicative of how many replies were received by a student; it represented the extent to which each student triggered subsequent postings. Higher in-degree centrality implies a higher degree of prominence of the actor in the network (Zemljič & Hlebec, 2005). Romero et al. (2013) indicated that degree centrality is a significant predictor of students' achievement in online discussions. Similarly, Russo and Koesten (2005) found that this measure is positively associated with student course grade points. In the current study, this variable was calculated as the normalized number of replies each student received from others.

2.4.2. Out-degree centrality

Here, out-degree centrality referred to how many replies each student generated in response to previous postings. High out-degree centrality implied that the students eagerly responded to others' postings. Out-degree centrality has been reported to have a positive correlation with student cognitive learning outcomes (Russo & Koesten, 2005). The measure also indicates the extent to which an actor in the social network contribute to online communities (Curran & Abidi, 2007). This variable was calculated as the normalized number of replies each student wrote to others.



Mean of learning interval $$\overline{\Delta t_i} = \frac{\sum_{i=1}^{n-1} \Delta t_i}{n}$$

Standard deviation $$S_i = \sqrt{\frac{\sum_{i=1}^{n-1}(t_i - \overline{\Delta t_i})^2}{n-1}}$$

Fig. 2. Calculating the login interval and regularity.

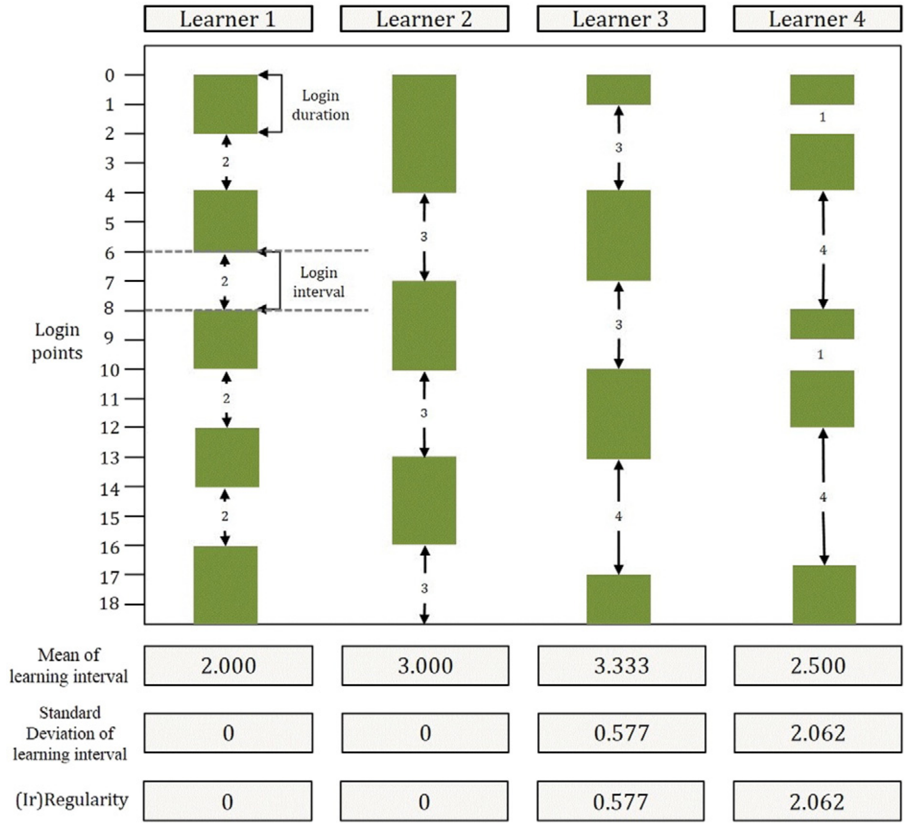


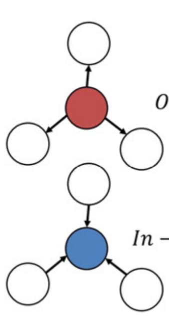
Fig. 3. Example of calculating login regularity (Jo et al., 2015).

Student interaction analysis using degree centrality can be included in automated prediction because network values are readily obtained from posting-relevant activities, which leave visible records (Xie, 2013). Without manual work from instructors, the system can compute the network variables using the data recorded about postings and replies.

3. Methods

3.1. Research context

To validate the constructed proxy variables, two 16-week blended courses were analyzed. The courses were offered to undergraduate students at a private women's university in South Korea and used AOD as their main activity. Each course was taught by a different instructor, who addressed different content areas in social science (see Table 1). The instructors had face-to-face meetings every week with students,



$$Out-degree\ centrality = \frac{Number\ of\ out-links}{n-1}$$

$$In-degree\ centrality = \frac{Number\ of\ in-links}{n-1}$$

Fig. 4. Calculating network degree centrality.

and the students participated in AOD after regular class time. Although instructors gave a new discussion topic every week, students could always participate in previous weeks' discussions.

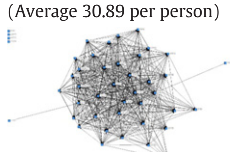
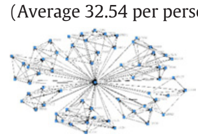
The primary reason we chose the two courses was that they used two distinct types of AOD: whole-class discussion (WCD) (e.g. Deryakulu & Olkun, 2007; Hara et al., 2000; Macfadyen & Dawson, 2010; Richardson & Ice, 2010) and team-based discussion (TBD) (e.g. Ekahitanond, 2013; Hsieh & Tsai, 2012; Schellens & Valcke, 2006; Shukor et al., 2014). Although there are wide variations in AOD, most can be categorized into one of the two types.

Students in WCD courses (in this study, Course X) participate in the same forum, whereas students in TBD courses (Course Y) participate in separate group discussions. Consequently, the sociogram of Course X presents a complex network, whereas Course Y presents a star-shaped network with one central node that indicates the instructor.

Although participation in discussions was not largely reflected in the final grades, it was important for both courses because all other assignments and tests were organized around the discussion topics. As shown in Table 1, the discussion topics were derived from students' individual essays or an issue reflected in a TV drama. These courses were students' major field of study and considered interesting to the students Therefore, they could be intrinsically motivated to participate in the discussions. The students could either initiate a discussion by uploading the first posting on the topic or reply to existing discussions. The total number of messages was 1298 in Course X and 2018 in Course Y.

In this study, low achievers were defined as those who received grades of C or below, belonging to the bottom 33%. High achievers were those who received grades of A or B. Although the term 'low achiever' was an operational definition, it was decided that the final grade could serve as the obvious outcome of student achievement. Data were collected within the Moodle LMS, an open source system for course management.

Table 1
Descriptions of the two AOD environments.

	Course X	Course Y
Course topic	Administration, law, and politics	Woman policy
Discussion type	Whole-class discussion	Team-based discussion
Number of participants	43	62
Weights for final grade	Attendance 5% AOD participation 15% Individual essays 20% Midterm and final exam 60%	Attendance 5% AOD participation 15% Team reports 20% Individual essays 5% Midterm and final exam 55%
Discussion topic	Selected weekly from students' individual essays	Selected weekly from a TV drama entitled "Can We Marry?"
Instructor gender and experience	A male professor with over 20 years of teaching and 3 years of online teaching experience	A female professor with over 15 years of teaching and 3 years of online teaching experience
Role of instructor	To review students' essays, provide individual feedback, and select the discussion topics	To observe team discussion and occasionally provide feedback during the course
Number of discussion group(s)	1	11
Facilitation for AOD participation and length of posting	Students get one point from one posting with 300–600 words and get two points with more than 601 words. The accumulated points were calculated for their AOD participation. Simple responses (e.g., I agree with you.) or greetings were not counted to get the point.	There was no specific length limitation. The following students were rewarded through special gifts by instructor: 1) team leaders, 2) students who posted the first, 50th, 100th, 150th, and 200th times, 3) 5 students who posted the most in class for discussion and collaboration.
Comment types in the system	2 layers (initial posting and reply) ※Technically, the same word limit (65,535 words) applies to the two layers	
# of Postings	Total 1298 messages (Average 30.89 per person)	Total 2018 messages (Average 32.54 per person)
Sociogram		

3.2. Data processing

Educational data mining and learning analytics are emerging disciplines, "concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students, and the settings which they learn in" (Berland, Baker, & Blikstein, 2014, p. 209). The disciplines encompass areas of research that use various types of educational data for prediction, knowledge discovery and decision making (Baker & Yacef, 2009). The current study attempted to perform prediction and discover knowledge about learning behaviors in AOD.

In the process of constructing the proxy variables, data mining techniques were used to control for possible prediction errors. Log data sets contain too much information, including meaningless cases as well as missing data that must be properly managed based on the researchers' decisions. Although the proxy variables were intended to be used for completely automated assessments, the log data set was manually retrieved in order to identify possible issues with regard to computing the data. In data mining, it is important to have explicit criteria for determining how to work with deficient data sets (Dasu & Johnson, 2003).

To elicit reliable results, the sixteenth week, at which time the participants took the final examination, was excluded from the analysis because there were no topics to discuss. The data from that week differed from those obtained in the previous weeks.

Student log data were accumulated throughout the course durations, resulting in the collection of increasing amounts of information towards the end of the courses. For example, the number of postings for the third week was the sum of postings from the first to third weeks. In addition to the student log data, the mid-term test score was included in the prediction from the eighth week because the midterm exam was held on the first day of the eighth week in both courses.

The diverse information on the discussion boards was converted into a database format through the following data preprocessing procedures: extracting data, de-identifying student IDs, categorizing the postings (e.g., initial posting and reply), counting the numbers of letters, and calculating the aforementioned diverse time-related variables using the time stamps on the discussion boards. Lastly, the data set was reorganized to address possible issues (e.g., missing values). For example, students whose log data were no longer recorded after a certain time point were excluded because this indicated that the students had dropped out of the courses.

Another issue was determining how accurately the total time spent on the LMS should be calculated from the myriad of login points. Login time is often used to measure learning time in online learning contexts. It can be calculated simply by using the login duration (e.g. Hung & Zhang, 2008; Hwang & Wang, 2004). However, login time could inevitably include meaningless pauses without participation in learning activities. To exclude the time in which no actual learning took place, the times spent on each learning task were combined, such as the time spent reading materials, downloading files, or uploading postings. This is a more reliable way to obtain the actual learning time than simply using the total login duration.

Another data manipulation technique was also employed given that many learners may log out without using the 'sign out' button, which could have resulted in the absence of a logout record and the overestimation of login duration. This issue has been discussed in some prior studies and usually replaced with the average time spent for other activities (e.g., Kovanović, Gašević, Joksimović, Hatala, & Adesope, 2015). This study also replaced the last activity time with the average time spent on the other activities within the same visit (Fig. 5).

3.3. Classification technique

To measure how well the proxy variables classified the low and high achievers in terms of accuracy, precision, recall, and specificity, this study employed the random forest (RF) technique proposed by Breiman (2001). RF is well-suited for developing prediction models with relatively small samples and a large number of predictors (Bureau et al., 2005). A begging process enables RF to achieve better generalization by decreasing the variance compared with other classification methods, such as a single decision tree or logistic regression (Qi, 2012).

As a supervised learning technique based on the tree-based algorithm, RF is a popular machine learning method that features random sampling and ensemble strategies. The basic principle of RF is to combine many decision trees that are grown on multiple bootstrap samples from a learning sample (Genuer, Poggi, & Tuleau-Malot, 2010). RF is performed based on generated decision trees that were calculated using random subsets of the data. Each node in the trees is then split using the best split among all variables (see Fig. 6) (Breiman, 2001; Liaw & Wiener, 2002). The recursive partitioning enables determination of the contribution and importance of each predictor (Strobl, Malley, & Tutz, 2009).

For each tree generated on the assigned bootstrap sample, the error rate can be calculated based on the observations that were left out of the bootstrap sample. This rate is called the out-of-bag (OOB) error rate, and it is estimated using one-third of the data that were left out in the training process. In essence, the rate indicates the overall percentage of

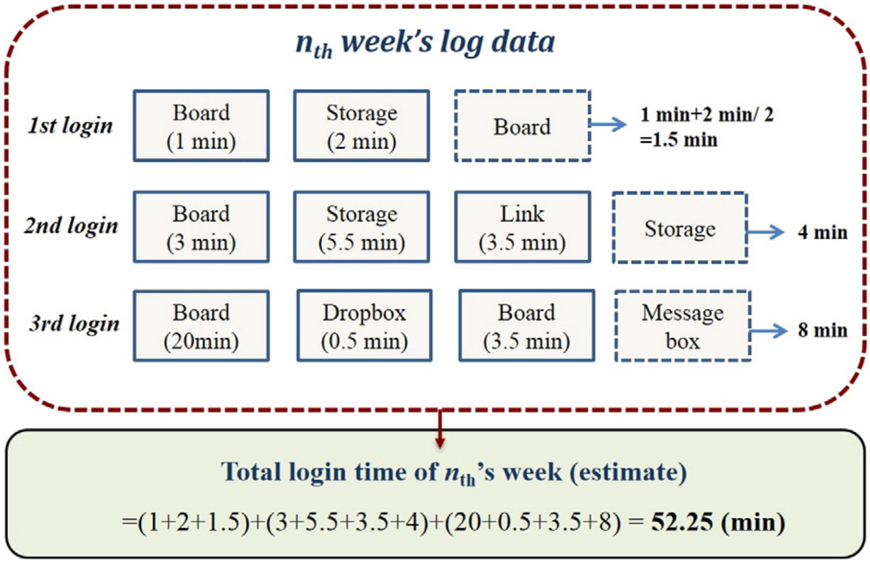


Fig. 5. Calculating total login time for students who exit without clicking the logout button.

misclassifications of the leftover samples. Because OOB estimates are used as an ingredient for providing as accurate an estimate of the generalization error as possible using a test and training set, there is no need to employ separate cross-validation in RF (Breiman, 2001). Accordingly, all collected data were analyzed without having separate sets for validation.

Where t is the number of generated trees, the variable importance (VI) can be estimated for each tree by calculating errOOB$_t$, the error of a single tree t on the OOB$_t$ sample (Genuer et al., 2010). The values of X^j in OOB$_t$ are permuted, but all other variables are unchanged to obtain a perturbed sample (OOB_t^j). The error of predictor t on the perturbed sample ($errOOB_t^j$) is shown in Fig. 7.

4. Results

4.1. Research question 1: how accurately can the proxy variables predict low and high achievers?

Fig. 8 shows the accuracy of the prediction model in the two courses. Accuracy refers to the proportion of correctly predicted cases divided by all possible cases. For Course X, the overall accuracy was already over 70% by the second week, which means that over 70% of the students could be correctly classified at a very early stage of the course. This is a high rate given the results from previous research (e.g. Lara et al., 2014; Romero et al., 2013; Thammasiri, Delen, Meesad, & Kasap, 2014). It is notable that the accuracy reached 88.37% during the sixth week, two weeks before the midterm exam, and it showed an increasing trend until the end of the courses.

The prediction model also showed high accuracy for Course Y (see Fig. 9). It was over 70% from the first week, and it reached 93.55% during the eighth week after showing steady increases. The model maintained its stable accuracy (over 90%) through the end of the course. The accuracy showed a dramatic increase in the eighth week when the mid-term test score was included. In fact, the OOB error rate for Course Y indicated that the mid-term test score was the most important variable in the eighth week but was not as influential for Course X. It is inferred that individual differences in the mid-term test was greater in Course Y than in Course X.

As shown in Table 2, we determined how many students were correctly classified during the middle phase (eighth week), which is considered to be a critical point at which students still have time to get back on track with their studies (Lara et al., 2014). In both courses, only two low achievers were incorrectly classified as high achievers, and another two high achievers were incorrectly classified as low achievers. In this study, specificity indicated how many low achievers were classified as such in the prediction results.

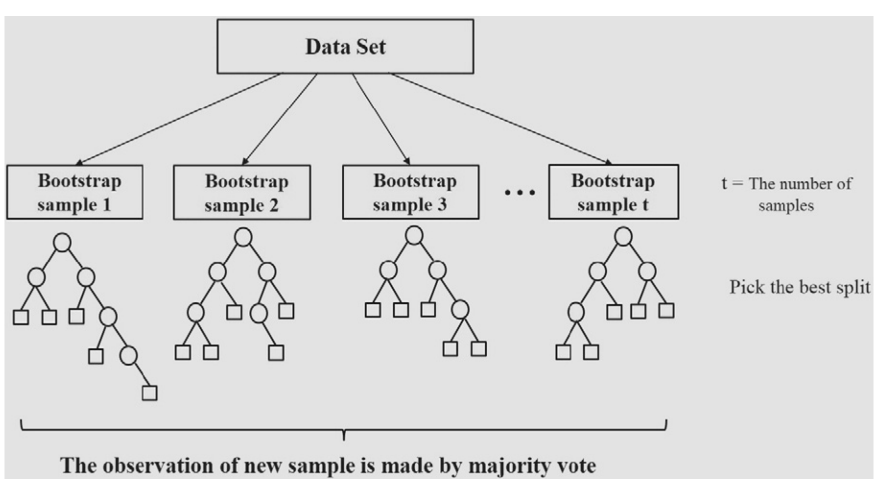


Fig. 6. Bootstrapping in random forest classification.

$$\mathrm{VI}(X^j) = \frac{1}{ntree} \sum_t \left(err\widetilde{OOB}_t^{\,j} - errOOB_t \right)$$

Fig. 7. Calculating errOOB.

centrality was again most frequent (7), and length (6) and number (5) of postings were newly added.

In Phase 1 of Course Y, discussion board visit frequency (6), discussion board visit interval regularity (4), and LMS visit frequency (4) were chosen. In Phase 2, discussion board visit frequency (7) and discussion board visit interval regularity (8) still appeared to be important, and LMS visit interval regularity (5) was newly added.

Table 2
Confusion matrix of local models in the eighth week.

	Course X ($N = 43$)				Course Y ($N = 62$)			
	Predicted							
Actual	High achievers		Low achievers		High achievers		Low achievers	
High achievers	28		2		43		2	
Low achievers	2		11		2		15	
Eighth week	Precision[a]	Recall[b]	Specificity[c]	Accuracy[d]	Precision	Recall	Specificity	Accuracy
	93.3%	93.3%	84.6%	90.7%	95.6%	95.6%	88.2%	93.6%

Note. True negative (TN): Case was negative and predicted as negative; True positive (TP): Case was positive and predicted as positive; False negative (FN): Case was positive but predicted as negative; False positive (FP): Case was negative but predicted as positive.

[a] Precision: TP / (TP + FP).
[b] Recall: TP / (TP + FN).
[c] Specificity: TN / (TN + FP).
[d] Accuracy: (TN + TP) / (TP + TN + FP + FN).

4.2. Research question 2: what learning behaviors are observed in AOD through the lens of the proxy variables?

To determine which proxy variables were important in the two phases of the two courses (before and after the mid-term tests), we first listed important variables from each week (Table 3). The degree of importance was obtained from the RF module, as discussed in the previous section. We interpreted the degree of importance as a relative ranking because it is not absolute and should not be compared over different weeks that have different sets of log data (Strobl et al., 2009)

After obtaining three important variables each week, we listed another three important variables in each phase based on how many times they appeared to be important each week. For example, in Phase 1 of Course X, out-degree centrality was important in that it appeared six times. We excluded the mid-term test scores in the process because this analysis aimed to identify online learning behaviors using the proxy variables.

In Phase 1 of Course X, out-degree centrality, LMS visit interval regularity, and total time spent on the LMS were shown to be important. Specifically, out-degree centrality appeared most frequently (6 times) in Phase 1, followed by LMS visit interval regularity (5) and total time spent on the LMS (4). In Phase 2, out-degree centrality, number of postings, and length of postings were shown to be important. Out-degree

5. Discussion

5.1. Prediction model performance

The prediction model that incorporated the proxy variables showed commendable performance in terms of accuracy, precision, specificity, and recall. The model's accuracy was stable throughout the two courses, with an increasing trend through the end.

It is noteworthy that the prediction model obtained high accuracy (~70%) from the first week. The accuracy was over 90% at the middle stage, and the fact that more than 90% of low achievers could be correctly classified at the middle phase of the courses reconfirms the possibility of employing timely intervention for students and for course improvement. Using proxy variables can be greatly beneficial because they play a role not only as predictors but also as indicators that provide instructors with specific information on students' online behaviors. In addition to knowing who are likely to be low achievers, instructors can see what specific behaviors the students are lacking. Subsequent instructional measures can be taken based on the indicators. This also

Table 3
The patterns of the importance of the proxy variables.

Course	Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	Importance	TTL	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC	ODC
		DVF	DIR	DTV	LIR	LIR	LIR	LIR	DVF	LOP	LOP	NOP	NOP	IDC	NOP	NOP
		LVF	LIR	TTL	TTL	TTL	DTV	DVF	LOP	DVF	DVF	LOP	IDC	NOP	LOP	LOP
	Phase	1 (7 weeks)								2 (8 weeks)						
	Frequency	ODC (6)								ODC (8)						
		LIR (5)								LOP (6)						
		TTL (4)								NOP (5)						
Course	Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	Importance	DTV	LIR	DIR	DIR	DVF	DVF	DVF	DVF	DVF	DTV	DIR	DIR	DVF	DVF	DVF
		TTL	DVF	LIR	LVF	LVF	LVF	DIR	DIR	DIR	LIR	LIR	LIR	LVF	LVF	LIR
		DVF	DTV	DVF	DVF	DIR	LIR	LVF	LIR	LIR	DIR	DVF	DVF	DIR	DIR	DIR
	Phase	1 (7 weeks)								2 (8 weeks)						
	Frequency	DVF (6)								DIR (8)						
		DIR (4)								DVF (7)						
		LVF (4)								LIR (5)						

Note. TTL: total time spent on LMS; LVF: LMS visit frequency; DVF: discussion board visit frequency; NOP: number of postings; LOP: length of postings; DTV: discussion time per visit; LIR: LMS visit interval regularity; DIR: discussion board visit interval regularity; IDC: in-degree centrality; ODC: out-degree centrality.

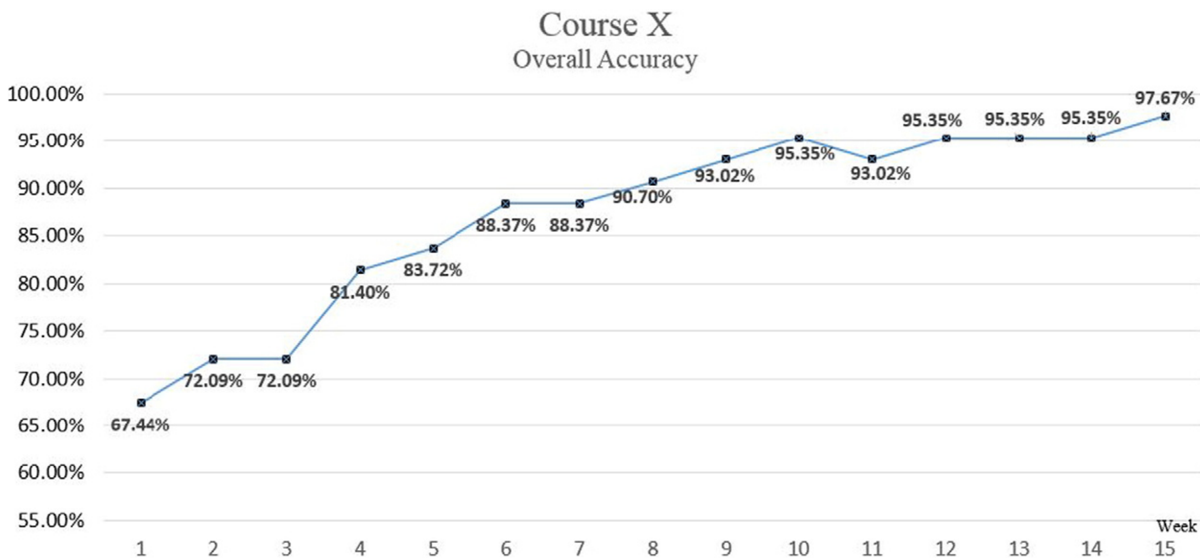


Fig. 8. Overall prediction accuracy for Course X.

implies that such variables will allow instructors to provide personalized interventions as well as to change their instructional strategies to promote AOD. Particularly, more attention should be paid to the specificity of the prediction model in regard to the importance of detecting low achievers. Classifying high achiever students as low achievers is relatively acceptable because giving special attention to potential high achievers may not be detrimental to them. However, providing no intervention for struggling students can be problematic because they are falling further behind. In this regard, it is encouraging that specificity appeared to be higher than 80% before the mid-phase (eighth week) in both courses. This means that the model resulted in false negative errors at a rate of less than 20% (classifying low achievers as not at risk) (Macfadyen & Dawson, 2010). The rate is considered to support the possibility of early detection and intervention.

Furthermore, it should be noted that the model's accuracy was sufficiently high even though the discussion score only accounted for 15% of the final score in both courses. This demonstrates the validity of the proxy variables that were employed as well as their potential to be applied in other AOD environments for which discussion only counts for a small portion of the final grade.

The results also indicate the reliability of the model. As aforementioned, this model yielded good prediction results despite distinct variations in the two AOD environments. Although the two courses produced different interaction patterns among instructors and students, the model showed similar prediction performance. We can expect the proxy variables to be applied to other AOD environments.

5.2. Four themes from additional analysis and suggestions for practice

From the investigation of the importance of the variables in the two phases, two themes emerged for the whole-class discussion course (Course X), and a further two emerged for the team-based discussion course (Course Y)

5.2.1. Whole-class discussion

5.2.1.1. Intellectual relationships through postings and replies. For Course X, which adopted the WCD format, the out-degree centrality appeared to be important throughout the course. That is, students who actively

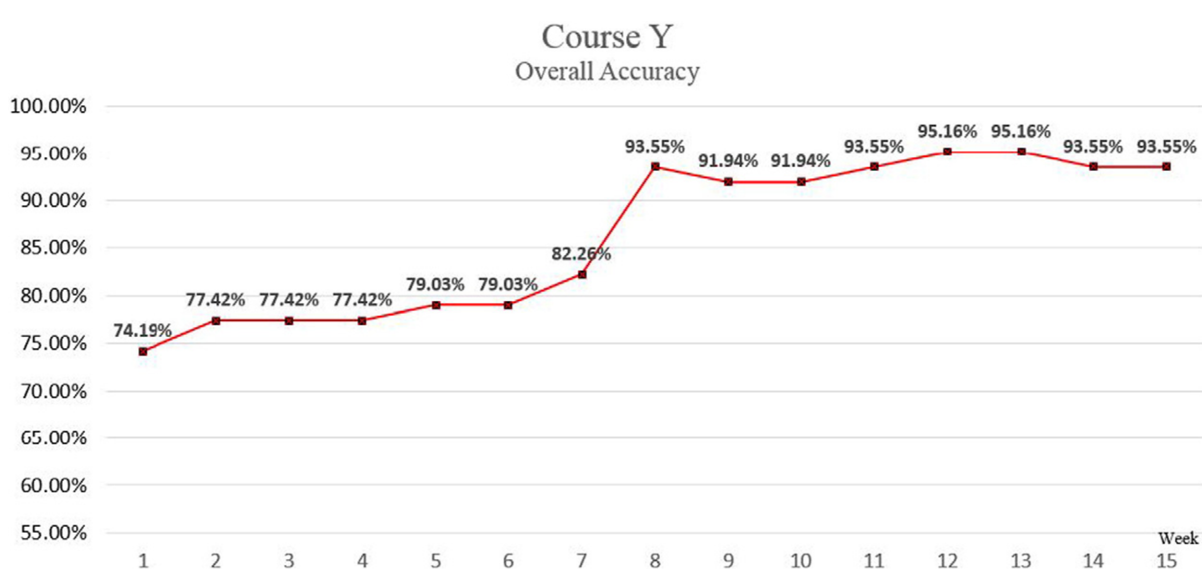


Fig. 9. Overall prediction accuracy for Course Y.

engaged in replying to others' postings demonstrated high academic achievement. This result is consistent with previous reports of a positive relationship between number of replies and achievement. Macfadyen and Dawson (2010) investigated student log data in an online WCD context and recognized the individual student networks as representing with whom the students were forming communicative relationships. In that study, high-performing students were found to have established more complex networks than their low-performing peers. Gao (2014) also noted the intellectual relationships students had established with their peers, specifically that meaningful learning and knowledge construction occurred when learners discussed to "negotiate meanings, and reconsider and refine previous thinking" (p. 4). Such activities necessarily involve communicating with others and are manifested by outgoing edges in a network's structure.

The whole-class discussion format, in which students do not have a specific team to interact with, could affect the importance of out-degree centrality. Macfadyen and Dawson (2010) found that in WCD courses, low-performing students tended to have limited networks that comprised low-performing peers. Unless they are encouraged to establish wide connections with each other, students are likely to maintain inactive or superficial interactions with limited numbers of people. This phenomenon raises the need for students to increase their "network density and diversity of relationships" (Macfadyen & Dawson, 2010, p. 596).

Separately, posting and replying are recognized as important for articulating ideas to argue, justify, explain, and provide evidence (Shukor et al., 2014). Such activities are positively correlated with cognitive engagement (Richardson & Ice, 2010). From this point of view, we can explain why the in-degree centrality did not appear to be as important as the out-degree centrality. In-degree centrality does not necessarily imply student productivity even though it represents the student's prominence within a network structure. Even if students' postings receive much attention from peers and trigger subsequent replies and even postings, this cannot be directly interpreted as their engagement in knowledge construction.

5.2.1.2. Suggestion: encourage students to reply to others. Again, it is worthy to note that out-degree appeared to be more important than in-degree centrality. This implies that there should be interventions to encourage students to reply to others' postings beyond general facilitation strategies. Although both posting and replying contribute to discussion, replying is related to paying attention to others' messages. This particular behavior is essential to establish a mutual communication network among participants toward in-depth discussion (Dawson, 2008; Deryakulu & Olkun, 2007). The emphasis on replying is thus important to design an effective intervention.

Instructors need to draw individual learners' attention to others' ideas on a discussion topic. For example, establishing a minimum reply requirement can be a way to make learners review and reply to others' ideas. Instructor can intervene in discussion with a question that invites all students to respond to the original posting. It is important to phrase a question in a way that motivates as many students as possible (Durrington, Berryhill, & Swafford, 2006).

Given the heavy workload of instructors in WCD settings, student-moderated discussion can also be useful to promote consistent learner-to-learner interaction (Durrington et al., 2006). A student in charge of moderating discussion defines what he/she expects of peers. For example, a moderator can state a criterion for an expected response to a discussion topic so that peers reply without confusion.

When students have little knowledge about how to initiate dialogue, more structured strategies (e.g., discussion prompts) can be employed. Darabi et al. (2013) found that students who used pre-structured prompts generated more messages triggering subsequent messages. The result indicates the potential of structured strategies to guide novice learners at an initial stage of discussion. As noted earlier, students participating in WCD do not have a specific group or person to discuss with. Encouraging replying will serve to trigger subsequent discussions to construct mutual communication networks among students.

5.2.1.3. A shift from course to discussion. For Course X, the length and number of postings appeared to be important in Phase 2. Those two variables were associated with active participation and engagement with discussion topics, respectively. We need to note that in Phase 2, the two LMS login variables (total time spent on LMS, LMS visit interval regularity) became less important. Rather, students' direct participation and engagement in discussions became more crucial predictors. This transitional process from exploring the course to in-depth discussions has been reported in previous studies. For example, Richardson and Ice (2010) examined students' critical thinking in an online WCD setting in the following four phases: triggering, exploration, integration, and resolution. They found that students' demonstration of critical thinking depended on their comfort level with the course and online discussion format. Students need time to become familiar with a course rather than immediately engaging in discussions. We can infer that potential high achievers in Phase 2 were those who successfully moved from gaining familiarity to engaging in discussions.

It is noteworthy that the transition was found only in the whole-class discussion format (Course X), in which students needed considerable time to engage in discussion topics. Without teams to interact with from the beginning, students in WCD courses need to explore overall course settings, available resources, and discussion topics, and students who familiarize themselves with the learning environment can better engage with discussion topics. Students need "to gain comfort and confidence in the online discussion format" to promote their cognitive engagement (Shea & Bidjerano, 2009, p. 551).

5.2.1.4. Suggestion: allow students to progress toward in-depth discussion. As evidenced by prior studies, learners participating in AOD move through several phases toward in-depth discussion. Interventions thus need to be provided adaptively to the different phases (Darabi, Arrastia, Nelson, Cornille, & Liang, 2011). Although there can be various ways to do so, the current study recognizes scaffolding as a useful strategy.

Scaffolding strategies are particularly beneficial when learners need to become increasingly independent and competent as they familiarize themselves with a discussion topic (Belland, Glazewski, & Richardson, 2008). For example, an instructor can provide a temporal support for learners to gain familiarity with a course and discussion topic at the beginning and gradually remove the support (i.e., fading). Different types of scaffolds can apply as students progress through the different phases. For example, instructors can offer content-related comments (e.g., conceptual scaffolds) in the early phases to guide them while using facilitative comments (e.g., strategic scaffolds) in the later phases to leave enough room for their intellectual involvement. Using scaffolding strategies for AOD has the great potential because scaffolds can be provided in computer-based environments in many different forms such as texts, visual aid, or videos.

5.2.2. Team-based discussion

5.2.2.1. Social cohesion in a small-group activity. In both phases of Course Y, LMS login variables did not appear to be as important as they were in Course X. LMS visit frequency and visit interval regularity were shown to be the third most important variables in both phases, whereas discussion-related variables (discussion visit frequency and discussion visit interval regularity) emerged as more important even from Phase 1. Considering that visits to the discussion board are essential for keeping up with discussion topics and progress, the analysis result indicates that successful students were those who made an early effort to participate in discussions. This early participation in team-based discussion is reported in some prior studies (e.g., Hsieh & Tsai, 2012).

Early participation may be critical in team-based discussion settings given that the discussion format assigns students to small discussion groups from the beginning of courses. For example, the students in Course Y in this study were more likely ready to participate in discussions from the first week because they did not need as much time to explore who they needed to communicate with as the Course X students did. Instead, in team-based discussions, it is more important to focus on interactions within small discussion teams as early as possible to share deeper thoughts through subsequent discussions.

The sense of intimacy in a team-based discussion is explained by the term 'social cohesion' in a previous study by Schellens and Valcke (2005). The study explained that social cohesion is formed through social messages exchanged among discussion group members. From this viewpoint, we can reconfirm the reason that establishing a network with as many students as possible was a crucial indicator of high achievement in the whole-class discussion setting, in which social messages are rarely found. In this regard, Dennen and Wieland (2007) noted the importance of 'social acknowledgement' and found that the group that used social acknowledgement (e.g., mentioning a name) demonstrated a higher level of engagement and intellectual interactions than the other group that used little social acknowledgement.

5.2.2.2. Suggestion: support students' cognitive engagement from the beginning. Students in TBD experience earlier social presence than those in WCD as they are assigned to a specific group. Research adopting the community of inquiry approach claims that social presence, defined as an ability to project personal characteristics as a real person, serves as a support for cognitive presence (Garrison, 1997). The claim provides rationale for early cognitive interventions in TBD because students participating in TBD are ready to engage in cognitive interactions with team members at an early stage.

There are various instructional strategies that can be used to promote early cognitive engagement. For example, Ekahitanond (2013) reported the effect of peer feedback on university students' reflection on discussion topics. Specifically, exchanging ideas, learning together, and comparing peers' responses resulted in students' enhanced understanding. Cognitive engagement can also be fostered through instructional materials. Gao (2014) revealed the effect of instructional guide on participants' discussion strategies. The result showed that participants who practiced the written guide describing discussion strategies prior to discussion (e.g., how to organize ideas) demonstrated more frequent use of the strategies. Role assignments at the beginning can also be a good strategy to enhance team members' cognitive engagement and understanding of a topic (McLaughlan, 2007). Student responsibility and comprehensive knowledge level can also be improved with a specific role when they bring domain-specific knowledge to team discussions (Darabi et al., 2011).

5.2.2.3. Regularity as representing awareness and consistency. In both phases of Course Y, visit interval regularity appeared to be important. Notably, in Phase 2, regular visits to the discussion boards became more important, appearing every week, and regular visits to the LMS were newly included rather than LMS visit frequency. The results indicated that regular participation in discussions was more prominent in the TBD course than it was for whole-class discussion. Team-based discussions that have small numbers of participants may make limited progress without everyone's contributions. In contrast, in a whole-class discussion setting, there are a sufficient number of students who can continue discussions on particular topics even if others contribute little.

Our interpretation of this theme is compatible with a study by Schellens and Valcke (2006). In that study, students' interactions in team discussions were observed to examine how their participation in AOD influenced their knowledge construction. The result indicated that task-oriented interactions occurred throughout the course. The fact that students engaged in task-oriented communication from the beginning to the end indicates the need for a consistent team effort to construct team knowledge.

In a previous study (Jo et al., 2015), regular login intervals were found to predict positive learning outcomes in an online learning environment. Regularity variables, unlike other login-related variables such as total login time, represent a consistent effort, and regularity is also assumed to represent the awareness of learning because regular learners are thought to have the intention to study regularly. For example, learners who study on the same day every week are likely to show a regular login pattern and thus to contribute to their discussion teams throughout a course.

5.2.2.4. Suggestion: help students sustain consistent engagement. As discussed, TBD is largely influenced by individual team members' contribution. Instructors thus need to encourage a team to sustain their participation throughout discussion. Incomplete intellectual network stemming from the unbalanced contribution of members can lead to an insufficient understanding.

Research on the community of inquiry argues that teaching presence contributes to learners' consistent cognitive engagement and critical inquiry. Garrison, Anderson, and Archer (1999) suggested three ways to increase teaching presence in AOD. First, individual contributions should be acknowledged to arouse student motivation. Second, effective group consciousness should be created to enable information sharing, reaching consensus, and understanding among team members. Last, explanatory feedback should be offered to promote learner reflection and understanding.

Teaching presence emerges via various channels. Using simple methods such as little guidance or messages acknowledging students' contribution can be a way (Tagg & Dickinson, 2008). One important strategy for teaching presence is that facilitating interventions need to be dispersed throughout discussions rather than being clustered at a particular point in time (Hew & Cheung, 2008).

Encouraging student reflection on their own posting also promotes consistent engagement of learners. For example, asking students a relevant question with their posting can encourage students to undertake further research and reflection (Durrington et al., 2006).

Aligning discussion topics with participants' interest is also an effective way to keep them engaged in discussion (Wu & Hiltz, 2004). Students become intrinsically motivated when discussion topics are suitably selected to match their background and interest. Similarly, choosing a topic that can be discussed from multiple perspectives increases student interactivity and engagement throughout discussion (Durrington et al., 2006).

6. Conclusion

In the current study, we provided a process for constructing proxy variables to be applied to AOD contexts, as well as empirical evidence that supported the validity of the variables. This study shows the potential of evidence-based learning analytics. To determine whether our application of the proxy variables was consistent with evidence reported in previous studies, we conducted further analysis to examine the prediction results for each week. This can be considered an additional validation process because we compared the patterns that appeared in the real data with existing evidence.

The present study is in line with research that has adopted learning analytics and data mining to discover knowledge from data (Romero et al., 2013). It will make the following contributions to research and practice in the areas of online learning and learning analytics.

First, this study suggested unique procedures for transforming conceptual constructs into quantifiable, and quantified, variables. The presence of theoretical and empirical evidence prior to our data analysis made it possible to obtain more reliable and applicable proxy variables. Such evidence-based approaches will also help to diversify research topics and address a wide range of education issues. To date, the majority of research in learning analytics has relied heavily on employing data, with little theoretical consideration (Kumari, 2011); in many cases, predictors were selected largely based on data availability. However, new discoveries in the field of learning analytics cannot be achieved without exhaustive examination of the existing knowledge. This field is searching for a new role, apart from the commercial area, as an independent academic discipline that addresses education problems (Siemens & Long, 2011). In fact, some research studies have aimed to determine how to balance theoretical considerations and data-driven approaches (e.g. Angeli & Valanides, 2013; Kumari, 2011; Wise et al., 2013). The data processing approach suggested herein will be conducive for refining the methodology in this field.

Second, this study showed the possibility of using proxy variables to improve instruction practices. Proxy variables can serve not only as predictors of students' online behaviors but also as indicators that provide directions for intervention to support students. Proxy variables that are constructed on solid evidence will help instructors decide what they need to improve their teaching as well as student learning. However, although our findings can help in understanding how the proxy variables captured students' online behaviors in real AOD settings, we are cautious about using this as a prescriptive solution. Because we constructed the proxy variables adopting a top–down approach, starting from empirical and theoretical evidence, applying the variables could result in obtaining different learning behaviors according to course settings. In particular, it should be noted that even though some variables were not listed among the top three variables, we still regarded them as important. Again, the primary purpose of the additional analysis was to confirm that the patterns of the proxy variables, as a result of applying real data, represented the existing evidence. For example, the fact that the length of postings was not among the top three variables in Phase 2 of Course Y does not mean that student engagement was not important. Rather, posting length was merely not among the more prominent predictors.

Third, the use of proxy variables can expand to a wide range of online and blended courses without requiring additional work by the instructor for assessment. Furthermore, early classification of students will enable effective and unobtrusive formative assessments. Proxy variables have great potential for courses that target large numbers of people, such as massive open online courses, because it is nearly impossible for instructors to manually assess their learners while teaching massive courses. The use of proxy variables will not only help instructors to use effective strategies but also help students to persevere in learning even when sufficient interaction is difficult.

Finally, we can expect proxy variables to be integrated into larger systems to function as comprehensive solutions for a variety of education issues, including predicting low achievers, assessing specific behaviors, and providing automated recommendations. Although many studies have attempted to identify better prediction models for student classification, few have focused on employing integrated systems for simultaneous prediction and action (e.g., Lara et al., 2014). Proxy variables can be combined with programming tools and data mining solutions to serve education purposes. Emerging techniques that are compatible with the database will contribute to the discovery of more meaningful knowledge from data.

7. Limitations and future research

This study has limitations that should be addressed in future research. First, although we chose two courses using two distinct types of AOD, the proxy variables need be tested with more AOD contexts. For example, the prediction result might be different in a hybrid course that employs both team-based and whole-class discussions. The limited number of samples is partly the result of the availability of student log data. Future studies can be considered to validate the proxy variables in more diverse course settings. Especially, the findings of this study should not be generalized to full asynchronous online courses without a face-to-face delivery mode. As noted by Liu et al. (2007), students have a lower sense of community in fully asynchronous courses than those enrolled in blended learning courses. Accordingly, implications and instructional suggestions in this study may translate to full AOD environments in a limited manner.

Second, we sought to analyze quantifiable behaviors (e.g., number of postings, degree centrality, posting length, etc.) rather than actual discussion content. Although this is a suitable methodology for automated assessment, the current study could have benefited from using non-overlapping qualitative data to better explain learner behaviors. We thus suggest that future research further investigate actual interactions and content of postings to provide strong evidence. In fact, research using a data-mining approach is increasingly looking to assess more complex learner behaviors by using advanced data mining algorithms (Berland et al., 2014). Accumulating knowledge obtained from qualitative analysis will be beneficial to research and practice in the data-mining and learning analytics fields.

Third, although we performed proper data processing to obtain reliable learning time, it is still possible that participants performed multiple tasks during learning time rather than one activity that they were thought to perform. For example, we cannot determine whether a participant fully engaged in a discussion topic based solely on the discussion time per visit variable if he or she was actually doing irrelevant tasks on other windows in their web browser (e.g., email checking and web-surfing). Such limitation of data-mining research using quantified log data has been revealed in prior studies (Dringus & Ellis, 2005). However, recent data tracking techniques and advanced tools are expected to enable more accurate data mining. For example, an increasing number of LMS provides more advanced customization to examine learners' specific online behaviors (e.g., multi-tasking) other than just intermittent clicks. We anticipate that future research will take advantage of fast-advancing technologies to obtain more reliable data-mining results.

Fourth, participants' level of experience with AOD was not taken into account despite the possibility of its effect on participants' online learning behaviors. In fact, there is some evidence that students' education level is positively correlated with performance in AOD (Darabi et al., 2013). Future research may elicit different findings by examining students' log data taking into consideration their prior experience and skills in online discussions.

Fifth, participants represent only one gender since this study was conducted at a woman's university. Although we assumed that gender differences would be minimally reflected in participants' online behaviors, there might have been possible gender effects. Future research that includes both genders would be able to provide more generalizable findings for research and practice without potential bias.

Last, this study did not provide details on how instruction interventions should be implemented in an integrated system. Although we noted the potential of a system to enable early detection and personalized interventions, more details need to be provided for system design purposes. Simply displaying significant predictors cannot be a solution to online learning issues. To elicit meaningful changes in instruction and learning, future research is merited to provide principles for designing a comprehensive system.

Acknowledgment

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5B6036244).

References

Andresen, M. A. (2009). Asynchronous discussion forums: Success factors, outcomes, assessments, and limitations. *Journal of Educational Technology & Society*, *12*(1), 249–257.

Angeli, C., & Valanides, N. (2013). Using educational data mining methods to assess field-dependent and field-independent learners' complex problem solving. *Educational Technology Research and Development*, *61*(3), 521–548.
Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–16.
Balaji, M. S., & Chakrabarti, D. (2010). Student interactions in online discussion forum: Empirical research from "media richness theory" perspective. *Journal of Interactive Online Learning*, *9*(1), 1–22.
Belland, B. R., Glazewski, K. D., & Richardson, J. C. (2008). A scaffolding framework to support the construction of evidence-based arguments among middle school students. *Educational Technology Research and Development*, *56*(4), 401–422.
Berge, Z., & Collins, M. (1995). *Computer-mediated communication and the online classroom in distance learning. Vol. 2*, Cresskill, NJ: Hampton Press.
Berland, M., Baker, R. S. J. D., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, *19*, 205–220.
Blaxter, L., & Tight, M. (1994). Juggling with time: How adults manage their time for lifelong education. *Studies in the Education of Adults*, *26*(2), 162–179.
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.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., ... Van, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, *28*, 171–182.
Butz, N. T., & Stupnisky, R. H. (2016). A mixed methods study of graduate students' self-determined motivation in synchronous hybrid learning environments. *The Internet and Higher Education*, *28*, 85–95.
Chan, C. K. K., & Chan, Y. -Y. (2011). Students' views of collaboration and online participation in knowledge forum. *Computers & Education*, *57*(1), 1445–1457.
Cho, H., Gay, G., Davidson, B., & Ingraffea, A. (2007). Social networks, communication styles, and learning performance in a CSCL community. *Computers & Education*, *49*(2), 309–329.
Cruz-Benito, J., Therón, R., García-Peñalvo, F. J., & Lucas, E. P. (2015). Discovering usage behaviors and engagement in an Educational Virtual World. *Computers in Human Behavior*, *47*, 18–25.
Curran, J. A., & Abidi, S. S. R. (2007). Evaluation of an online discussion forum for emergency practitioners. *Health Informatics Journal*, *13*(4), 255–266.
Darabi, A., Arrastia, M. C., Nelson, D. W., Cornille, T., & Liang, X. (2011). Cognitive presence in asynchronous online learning: A comparison of four discussion strategies. *Journal of Computer Assisted Learning*, *27*(3), 216–227.
Darabi, A., Liang, X., Suryavanshi, R., & Yurekli, H. (2013). Effectiveness of online discussion strategies: A meta-analysis. *American Journal of Distance Education*, *27*(4), 228–241.
Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning.* Hoboken, NJ: Wiley-Interscience.
Dawson, S. (2008). A study of the relationship between student social networks and sense of community. *Journal of Educational Technology & Society*, *11*(3), 224–238.
Dawson, S., Macfadyen, L., Lockyer, L., & Mazzochi-Jones, D. (2011). Using social network metrics to assess the effectiveness of broad based admission practices. *Australasian Journal of Educational Technology*, *27*(1), 16–27.
De Laat, M., Lally, V., Lipponen, L., & Simons, R. -J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 87–103.
Dennen, V. P. (2008). Pedagogical lurking: Student engagement in non-posting discussion behavior. *Computers in Human Behavior*, *24*(4), 1624–1633.
Dennen, V. P., & Wieland, K. (2007). From interaction to intersubjectivity: Facilitating online group discourse processes. *Distance Education*, *28*(3), 281–297.
deNoyelles, A., Zydney, J. M., & Chen, B. (2014). Strategies for creating a community of inquiry through online asynchronous discussions. *Journal of Online Learning and Teaching*, *10*(1), 153–165.
Deryakulu, D., & Olkun, S. (2007). Analysis of computer teachers' online discussion forum messages about their occupational problems. *Journal of Educational Technology & Society*, *10*(4), 131–142.
Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, *45*(1), 141–160.
Duque, R., Gómez-Pérez, D., Nieto-Reyes, A., & Bravo, C. (2015). Analyzing collaboration and interaction in learning environments to form learner groups. *Computers in Human Behavior*, *47*, 42–49.
Durden, G., & Ellis, L. (2003). Is class attendance a proxy variable for student motivation in economics classes? An empirical analysis. *International Social Science Review*, *78*(1&2), 42–46.
Durrington, V. A., Berryhill, A., & Swafford, J. (2006). Strategies for enhancing student interactivity in an online environment. *College Teaching*, *54*(1), 190–193.
Dykman, C. A., & Davis, C. K. (2008). Online education forum: Part two-teaching online versus teaching conventionally. *Journal of Information Systems Education*, *19*(2), 157–164.
Ekahitanond, V. (2013). Promoting university students' critical thinking skills through peer feedback activity in an online discussion forum. *Alberta Journal of Educational Research*, *59*(2), 247–265.
Gao, F. (2014). Exploring the use of discussion strategies and labels in asynchronous online discussion. *Online Learning*, *18*(3), 1–19.
Garrison, D. R. (1997). Computer conferencing: The post-industrial age of distance education. *Open Learning*, *12*(2), 3–11.
Garrison, D. R., & Arbaugh, J. B. (2007). Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education*, *10*, 157–172.
Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, *2*(2–3), 87–105.
Genuer, R., Poggi, J. -M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.
Gilbert, P. K., & Dabbagh, N. (2005). How to structure online discussions for meaningful discourse: A case study. *British Journal of Educational Technology*, *36*(1), 5–18.
Hara, N., Bonk, C. J., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science*, *28*(2), 115–152.
Hernández-García, Á., González-González, I., Jiménez-Zarco, A. I., & Chaparro-Peláez, J. (2015). Applying social learning analytics to message boards in online distance learning: A case study. *Computers in Human Behavior*, *47*, 68–80.
Hew, K. F., & Cheung, W. S. (2008). Attracting student participation in asynchronous online discussions: A case study of peer facilitation. *Computers & Education*, *51*(3), 1111–1124.
Hew, K. F., Cheung, W. S., & Ng, C. S. L. (2010). Student contribution in asynchronous online discussion: A review of the research and empirical exploration. *Instructional Science*, *38*(6), 571–606.
Hewitt, J., Brett, C., & Peters, V. (2007). Scan rate: A new metric for the analysis of reading behaviors in asynchronous computer conferencing environments. *The American Journal of Distance Education*, *21*(4), 215–231.
Hou, H. -T. (2011). A case study of online instructional collaborative discussion activities for problem-solving using situated scenarios: An examination of content and behavior cluster analysis. *Computers & Education*, *56*(3), 712–719.
Hrastinski, S. (2008). Asynchronous and synchronous e-learning: A study of asynchronous and synchronous e-learning methods discovered that each supports different purposes. *EDUCAUSE Quarterly*, *31*(4), 51–55.
Hrastinski, S. (2009). A theory of online learning as online participation. *Computers & Education*, *52*, 78–82.
Hsieh, Y. -H., & Tsai, C. -C. (2012). The effect of moderator's facilitative strategies on online synchronous discussions. *Computers in Human Behavior*, *28*(5), 1708–1716.
Hung, J. -L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *Journal of Online Learning and Teaching*, *4*(4).
Hwang, W. -Y., & Wang, C. -Y. (2004). A study of learning time patterns in asynchronous learning environments. *Journal of Computer Assisted Learning*, *20*(4), 292–304.
Irvine, V., Code, J., & Richards, L. (2013). Realigning higher education for the 21st-century learner through multi-access learning. *Journal of Online Learning and Teaching*, *9*(2), 172–186.
Jo, I. H., Kim, D., & Yoon, M. (2015). Constructing proxy variables to measure adult learners' time management strategies in LMS. *Educational Technology & Society*, *18*(3), 214–225.
Jung, J., & Gilson, T. A. (2014). Online threaded discussion: Benefits, issues, and strategies. *Kinesiology Review*, *3*(4), 241–246.
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.
Kumari, M. (2011). Data driven data mining to domain driven data mining. *Global Journal of Computer Science and Technology*, *11*(23), 65–68.
Lara, J. A., Lizcano, D., Martínez, M. A., Pazos, J., & Riera, T. (2014). A system for knowledge discovery in e-learning environments within the European higher education area—Application to student data from Open University of Madrid, UDIMA. *Computers & Education*, *72*, 23–36.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Liu, X., Magjuka, R. J., Bonk, C. J., & Lee, S. (2007). Does sense of community matter? *The Quarterly Review of Distance Education*, *8*(1), 9–24.
Loncar, M., Barrett, N. E., & Liu, G. -Z. (2014). Towards the refinement of forum and asynchronous online discussion in educational contexts worldwide: Trends and investigative approaches within a dominant research paradigm. *Computers & Education*, *73*, 93–110.
Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950–965.
Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599.
Mason, R. B. (2011). Student engagement with, and participation in, an e-forum. *Journal of Educational Technology & Society*, *14*(2), 258–268.
McLaughlan, R. G. (2007). Instructional strategies to educate for sustainability in technology assessment. *International Journal of Engineering Education*, *23*(2), 201.
Nagel, L., Blignaut, A. S., & Cronjé, J. C. (2009). Read-only participants: A case for student communication in online classes. *Interactive Learning Environments*, *17*(1), 37–51.
Qi, Y. (2012). Random forest for bioinformatics. *Ensemble Machine Learning*, *307*.
Richardson, J. C., & Ice, P. (2010). Investigating students' level of critical thinking across instructional strategies in online discussions. *The Internet and Higher Education*, *13*(1), 52–59.
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.
Russo, T. C., & Koesten, J. (2005). Prestige, centrality, and learning: A social network analysis of an online class. *Communication Education*, *54*(3), 254–261.
Schellens, T., & Valcke, M. (2005). Collaborative learning in asynchronous discussion groups: What about the impact on cognitive processing? *Computers in Human Behavior*, *21*(6), 957–975.

Schellens, T., & Valcke, M. (2006). Fostering knowledge construction in university students through asynchronous discussion groups. *Computers & Education*, *46*(4), 349–370.

Schworm, S., & Gruber, H. (2012). e-Learning in universities: Supporting help-seeking processes by instructional prompts. *British Journal of Educational Technology*, *43*(2), 272–281.

Shana, Z. (2009). Learning with technology: Using discussion forums to augment a traditional-style class. *Journal of Educational Technology & Society*, *12*(3), 214–228.

Shea, P., & Bidjerano, T. (2009). Community of inquiry as a theoretical framework to foster "epistemic engagement" and "cognitive presence" in online education. *Computers & Education*, *52*(3), 543–553.

Shukor, N. A., Tasir, Z., Van der Meijden, H., & Harun, J. (2014). Exploring students' knowledge construction strategies in computer-supported collaborative learning discussions using sequential analysis. *Journal of Educational Technology & Society*, *17*(4), 216–228.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, *46*(5), 30–32.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

Tagg, A., & Dickinson, J. (2008). Tutor messaging and its effectiveness in encouraging student participation on computer conferences. *International Journal of E-Learning & Distance Education*, *10*(2), 33–55.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330.

Thomas, M. J. (2002). Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning*, *18*(3), 351–366.

Topcu, A., & Ubuz, B. (2008). The effects of metacognitive knowledge on the pre-service teachers' participation in the asynchronous online forum. *Journal of Educational Technology & Society*, *11*(3), 1–12.

Vignare, K. (2007). Review of literature blended learning: Using ALN to change the classroom—Will it work? In A. G. Picciano, & C. D. Dziuban (Eds.), *Blended learning: research perspectives* (pp. 37–63). Needham, MA: Sloan Center for Online Education.

Webb, E., Jones, A., Barker, P., & van Schaik, P. (2004). Using e-learning dialogues in higher education. *Innovations in Education and Teaching International*, *41*(1), 93–103.

Wickens, M. R. (1972). A note on the use of proxy variables. *Econometrica*, *40*(4), 759–761.

Wise, A. F., Marbouti, F., Hsiao, Y. -T., & Hausknecht, S. (2012). A survey of factors contributing to learners' "listening" behaviors in asynchronous online discussions. *Journal of Educational Computing Research*, *47*(4), 461–480.

Wise, A. F., Speer, J., Marbouti, F., & Hsiao, Y. -T. (2013). Broadening the notion of participation in online discussions: Examining patterns in learners' online listening behaviors. *Instructional Science*, *41*(2), 323–343.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, *4*(3–4), 129–164.

Wu, D., & Hiltz, S. R. (2004). Predicting learning from asynchronous online discussions. *Journal of Asynchronous Learning Networks*, *8*(2), 139–152.

Xie, K. (2013). What do the numbers say? The influence of motivation and peer feedback on students' behaviour in online discussions. *British Journal of Educational Technology*, *44*(2), 288–301.

Yeh, H. -T., & Van Buskirk, E. (2005). An instructor's methods of facilitating students' participation in asynchronous online discussion. In C. Crawford, R. Carlsen, I. Gibson, K. McFerrin, J. Price, R. Weber, & D. Willis (Eds.), *Proceedings of society for information technology & teacher education international conference 2005. Chesapeake, VA: Association for the Advancement of Computing in Education (AACE)*. (pp. 682–688).

Zacharis, N. Z. (2015). A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, *27*, 44–53.

Zemljič, B., & Hlebec, V. (2005). Reliability of measures of centrality and prominence. *Social Networks*, *27*(1), 73–88.