Tech Know Learn (2014) 19:221–240
DOI 10.1007/s10758-014-9226-4

WORK-IN-PROGRESS

Development and Validation of a Learning Analytics Framework: Two Case Studies Using Support Vector Machines

Dirk Ifenthaler · Chathuranga Widanapathirana

Published online: 11 May 2014

Abstract Interest in collecting and mining large sets of educational data on student background and performance to conduct research on learning and instruction has developed as an area generally referred to as learning analytics. Higher education leaders are recognizing the value of learning analytics for improving not only learning and teaching but also the entire educational arena. However, theoretical concepts and empirical evidence need to be generated within the fast evolving field of learning analytics. The purpose of the two reported cases studies is to identify alternative approaches to data analysis and to determine the validity and accuracy of a learning analytics framework and its corresponding student and learning profiles. The findings indicate that educational data for learning analytics is context specific and variables carry different meanings and can have different implications across educational institutions and area of studies. Benefits, concerns, and challenges of learning analytics are critically reflected, indicating that learning analytics frameworks need to be sensitive to idiosyncrasies of the educational institution and its stakeholders.

Keywords Learning analytics · Student profile · Learning profile · Study success · Machine learning · Support vector machines

1 Introduction

Massive administrative, systems, academic, and personal data within educational settings are becoming more and more available. This vast amount of educational information

D. Ifenthaler (✉)
Deakin University, Melbourne, Australia
e-mail: dirk@ifenthaler.info
URL: http://www.deakin.edu.au

C. Widanapathirana
Open Universities Australia, Melbourne, Australia
e-mail: chath.widanapathiran@open.edu.au
URL: http://www.open.edu.au

requires well-established data management, analysis, and interpretation (Long and Siemens 2011). Three concepts are linked to processing such educational information: Educational data mining, academic analytics, and learning analytics.

Educational data mining (EDM) refers to the process of extracting useful data out of a large collection of complex educational datasets (Romero et al. 2011). Academic analytics (AA) is the identification of meaningful patterns in educational data in order to inform academic issues (e.g., retention, success rates) and produce actionable strategies (e.g., budgeting, human resources) (Campbell et al. 2010). Learning analytics (LA) uses dynamic information about learners and learning environments, assessing, eliciting and analyzing it, for real-time modeling, prediction and optimization of learning processes, learning environments, as well as educational decision-making (Ifenthaler in press; Lockyer et al. 2013; Johnson et al. 2013).

All three concepts (EDM, AA, LA) refer to processing massive educational data, however, only the LA concept does emphasize the optimization of learning processes and learning environments in real-time. Further, learners' needs and their predispositions are multidimensional and quickly change over time (Ashby 1992; Ifenthaler and Seel 2013). Numerous approaches for understanding these complex patterns of learning and predicting their future developments for automating instruction have been challenged repeatedly in the past (Ifenthaler et al. 2010). Applications of LA presupposes a seamless and system-inherent analysis of learner's progression in order to continuously adapt the learning environment (Azevedo et al. 2005; Kalyuga 2006; Lin et al. 2013). Additionally, LA provides the pedagogical and technological background for producing real-time interventions at all times during the learning process.

The purpose of this study is to address two major challenges of LA: (1) Explore different approaches for data analysis for LA and (2) determine the validity of profiles based on a LA framework. The following section introduces a LA framework, its related profiles, and Support Vector Machines as an alternative approach for data analysis. Next, two case studies for validating student and learning profiles of the LA framework using support vector machines are presented. The general discussion critically reflects on the results, suggests implications, and addresses concerns as well as further challenges of LA. The final section concludes with a general comment towards future applications of LA.

2 Learning Analytics

2.1 Holistic Framework

As the field of LA is growing, several frameworks have been proposed which focus on available data, instruments for data analysis, involved stakeholders, and limitations (Greller and Drachsler 2012; Ferguson 2012). d'Aquin et al. (2014) argue for a closer relationship between LA and linked data with a particular emphasis on semantic web technologies. By connecting online available educational resources this approach, however, does not include valuable information of learner's background information as well as curricular requirements. Other frameworks focus on social learning analytics (SLA) in which discussion activities are visualized using data mining and visualization tools (Schreurs et al. 2014; Buckingham Shum and Ferguson 2012). The proposed tools have the potential to provide rich information about learning processes in discussion activities in real-time. Greller and Drachsler (2012) introduce six critical dimensions of a LA framework including stakeholders, objectives, data, instruments, internal and external

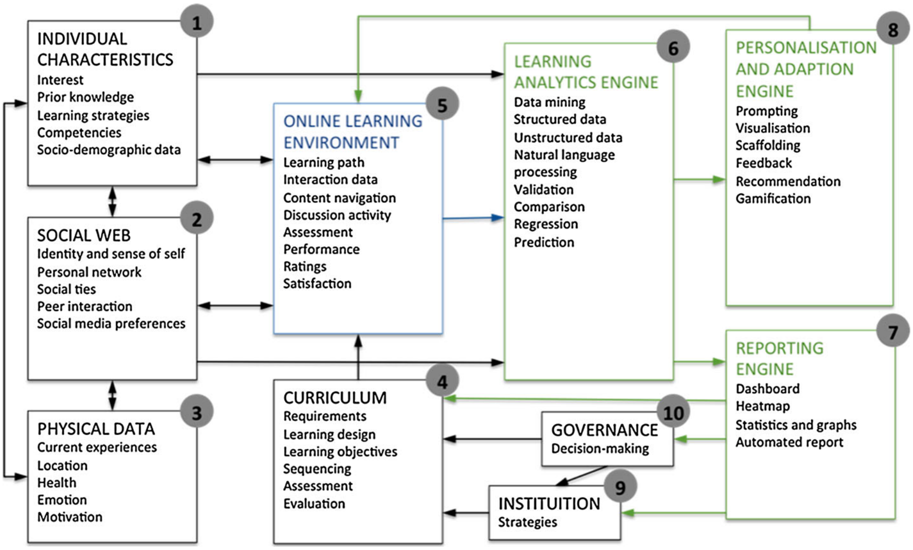

Fig. 1 Components and relations of the LA framework

constraints. These dimensions are critical when designing and implementing LA applications and therefore provide a valuable guideline for LA projects. Still, elaborated and more importantly empirically validated LA frameworks are scarce. Another limitation of existing frameworks is the missing link of learner characteristics (e.g., prior learning), learning behavior (e.g., access of materials), and curricular requirements (e.g., competences, sequencing of learning).

Therefore, Fig. 1 illustrates a holistic view of a LA framework linking various types of educational information in a meaningful way (Ifenthaler in press).

The LA framework combines data directly linked to (1) individual stakeholders, their interaction with the (2) social web and the (5) online learning environment, as well as (4) curricular requirements. Additionally, data from (3) outside of the educational system is integrated. The (6) processing and analysis of the combined data is carried out in a multilayer data warehouse and (7, 8) returned to the stakeholders, e.g., (10) governance or (9) institution, in a meaningful way.

Characteristics of (1) individual stakeholders include socio-demographic information, personal preferences and interests, responses to standardized inventories (e.g., learning strategies, achievement motivation, personality), demonstrated skills and competencies (e.g., computer literacy), acquired prior knowledge and proven academic performance, as well as institutional transcript data (e.g., pass rates, enrolment, dropout, special needs).

Associated interactions with the (2) social web include preferences of social media tools (e.g., Twitter, Facebook, LinkedIn) and social network activities (e.g., linked resources, friendships, peer groups, web identity).

Data from (3) outside the educational system is collected through various systems, for example through a library system (i.e., university library, public library). Other physical data may include sensor and location data from mobile devices (e.g., study location and time), or affective states collected through reactive tests (e.g., motivation, emotion, health, stress, commitments).

The (5) online learning environment (i.e., learning management system, personal learning environment, learning blog) provides rich data of stakeholder activities which are mostly numeric, for example logging on/off, viewing and/or posting discussions, results on assessment tasks, or responses to ratings and surveys. These data can be aggregated to produce data trails, such as navigation patterns or learning preferences and pathways. More importantly, rich semantic and context specific information are available from discussion forums as well as from complex learning tasks, for example from written essays, Wikis, or blog posts. Additionally, interactions of various stakeholders (e.g., student–student; student–teacher; tutor–teacher) are tracked.

Closely linked to the content and activities available from the online learning environment is the (4) curricular information which includes meta data of all features of the online learning environment. This meta data reflects the learning design (e.g., sequencing of materials, tasks and assessments) and expected learning outcomes (e.g., specific competencies). Ratings of materials, activities, and assessments as well as formative and summative evaluation data are directly linked to specific curricula and stakeholders.

Structured and unstructured data from all systems are combined and processed in a multilayer data warehouse using adaptive algorithms, referred to as the (6) LA engine. The results of the data mining process are validated before further analyses are computed. Data analytics approaches include supervised and unsupervised machine learning methods as well as linear and nonlinear modeling methods. Such approaches include Support Vector Machines, Bayesian networks, neural networks, natural language processing, survival analysis, and hierarchical linear modeling which need to be closely linked to the underpinnings of applied pedagogical theories (see Sect. 2.3 for detailed information of these approaches).

The (7) reporting engine uses the results of the LA engine and presents them in forms of interactive dashboards, heat maps, statistics and graphs, as well as automated reports. These automated reports are utilized for specific stakeholders such as the (10) governance level (e.g., for cross-institutional comparisons), a (9) single institutions (e.g., for internal comparisons, optimization of sequence of operations), as well as the (4) curriculum level including insights and reports for learning designers and facilitators for analyzing instructional processes, processing of learning materials, and students' pathways.

The (8) personalization and adaption engine uses the results of the LA engine for informative real-time feedback and scaffolds in the (5) online learning environment. Interactive elements include simple learning prompts and recommendations (e.g., reminder of deadlines, links to further learning materials, social interaction), rich personalized visualizations (e.g., current and forecast of learning paths), as well as informative scaffolds for specific learning activities and assessment tasks.

The implementation of the LA framework requires access to a real-time data collection and adaptive processing of available data. This allows all stakeholders to personalize the LA process in order to meet their individual requirements.

2.2 Profiles

Based on the above described LA framework, three profiles have been identified for implementation: (1) student profile, (2) learning profile, and (3) curriculum profile (see Fig. 2).

The student profile includes static and dynamic parameters. Static parameters do not change quickly over time and include gender, age, education level and history, work experience, current employment status, etc. Dynamic parameters are changing over time

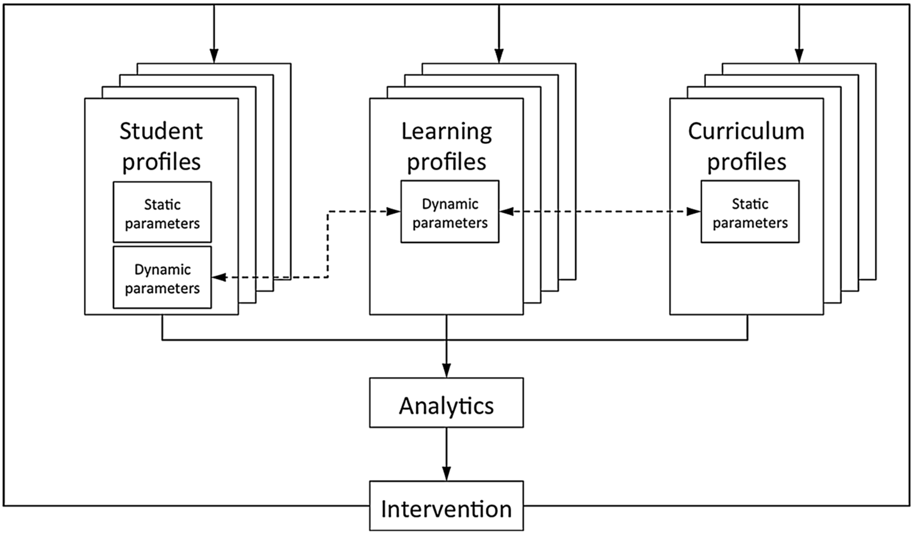

Fig. 2 Connectedness of student, learning, and curriculum profiles of the LA framework

and include interest, motivation, response to reactive inventories (e.g., learning strategies, achievement motivation, emotions), computer and social media skills, enrolments, drop outs, pass-fail rate, average performance rate, etc.

The learning profile includes variables reflecting the current performance within the learning environment (e.g., learning management system). Dynamic parameters include time specific information such as time spent on learning environment, time per session, time on task, time on assessment. Other parameters of the learning profile include login frequency, task completion rate, assessment activity, assessment outcome, learning material activity (upload/download), discussion activity, support access, ratings of learning material, assessment, support, effort, etc.

The curriculum profile includes parameters reflecting the expected and required performance defined by the learning designer and course creator. Static parameters include course information such as facilitator, title, level of study, and prerequisites. Individual learning outcomes are defined including information about knowledge type (e.g., content, procedural, causal, meta cognitive), sequencing of materials and assessments, as well as required and expected learning activities.

The available data from all profiles are analyzed using pre-defined analytic models allowing summative, real-time, and predictive comparisons. The results of the comparisons are used for specifically designed interventions that are returned to the corresponding profiles. The automated interventions include reports, dashboards, prompts, and scaffolds. Additionally, stakeholders receive customized messages for following up with critical incidents (e.g., students at risk, assessments not passed, satisfaction not acceptable, learning materials not used, etc.).

2.3 Support Vector Machines

The relative new field of LA and big data in education does not provide standardized analytical strategies for informing LA frameworks and related profiles. Currently, major analytical strategies for LA involve variations of regression analysis, such as, linear regression models, logistic regression models, hierarchical linear models (da Silva et al. 2013). Other stochastic approaches include Bayesian networks and neural networks which enable adjustments to the applied algorithms based on previous results (Bartholomew 1967). However, to identify highly non-linear and complex parameter relationships, the above-mentioned analytical strategies have obvious limitations.

Besides random forest (Breiman 2001) and decision tree (Quinlan 1986) approaches, support vector machines (SVM) is a promising alternative data analytic approach for educational data and LA. SVM is a binary classification technique based on supervised machine learning in the broad area of artificial intelligence (Drucker et al. 1997). Major applications include pattern recognition, classification, and regression modeling (Christmann and Steinwart 2008). The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier (Cortes and Vapnik 1995). Given a set of training examples, each marked as belonging to one of two categories; an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The advantages of SVM can be summarized as follows (Williams 2011; Cleophas and Zwinderman 2013):

- SVM offer flexibility in modeling non-linear educational data.
- SVM has short training times to create new models and offer very fast testing speeds when new samples are classified. These capabilities satisfy the demands of a real-time LA system.
- SVM are flexible with regard to interactions between educational parameters from different sources and hardly effected by the correlated parameters unlike most other regression techniques.
- SVM does not rely on priori-knowledge on event probabilities that are often unavailable and unreliable in education data.
- SVM can process imperfect educational data by providing a better sensitivity for modeling dependent variables. SVMs are inherently robust against parse data and outliers.

2.4 The Case Studies

Not all educational data is relevant and equivalent (Macfadyen and Dawson 2012). Therefore, the theoretical and empirical validity of underlying profiles and the accuracy of algorithms as well as its reliable analyses are critical for generating useful summative, real-time, and predictive insights from LA. This initial investigation of the above-presented LA framework includes two case studies focussing on the (1) student profile and (2) learning profile through the application of SVM as an alternative data analytic approach.

The purpose of the *first case study* is to validate the above described student profile. Well accepted empirical investigations identified variables directly linked to the student profile (e.g., age, gender, education background, work hours, etc.) as critical factors for study success (Tinto 1999; James et al. 2010; Thomas 2011; Crosling et al. 2009; Tinto

1982). As part of assessing the validity of the proposed student profile, we adhere to the question which specific factors of the student profile best explain study unit outcomes?

Hypotheses 1a It is hypothesized that student profile factors can be identified which explain at least 40 % of variance of study unit outcomes.

Further, a major benefit expected from the underlying student profile of the LA framework is providing early personalized interventions for students as well as facilitating their on-going learning progression towards successful study unit outcomes (Aflalo and Gabay 2012; Fenwick and Cooper 2012; Allen et al. 1988; Perumallaa et al. 2010; Lockyer et al. 2013; Greller and Drachsler 2012; Macfadyen and Dawson 2010). Therefore, the algorithms of the applied SVM model for the student profile require a high accuracy for suggesting interventions for successful study unit outcomes (Williams 2011). This leads to our second research question: Do the algorithms of the SVM model for the student profile contain sufficient information for providing recommendations for personalized interventions for predicting study unit outcomes with acceptable accuracy?

Hypothesis 1b It is hypothesized that the student profile can predict study unit outcomes with at least 80 % accuracy.

Another challenge for establishing a LA framework is the interpretation of results against the educational setting and its contextual idiosyncrasies (Coates 2010). Consequently, the interpretation of analysis results depends on the context in which the educational data were collected (Lockyer et al. 2013). In other words, variables and indicators can carry different meanings and can therefore have different implications. Further, these variables and indicators may be underpinned by different data from different contexts such as distinct area of studies or various institutions (Coates 2009, 2010; Bauer 1966). Therefore, this case study investigates the student profile in the light of the idiosyncrasies of higher education institutions and area of studies.

Hypothesis 1c It is hypothesized that the explained variance of the student profile differs across higher education institutions.

Hypothesis 1d It is hypothesized that the explained variance of the student profile differs across area of studies.

The *second case study* seeks to investigate the validity of the above-described learning profile. More specifically, the case study investigates which specific factors of the learning profile explain study unit outcomes.

Hypotheses 2a It is hypothesized that learning profile factors can be identified which explain at least 80 % of variance of study unit outcomes.

Similar to the first case study, the algorithms of the applied SVM model for the learning profile require a high accuracy for suggesting interventions towards successful study unit outcomes. Therefore, we adhere to the question whether the algorithms of the SVM model for the learning profile contain information for predicting study unit outcomes with acceptable accuracy?

Hypothesis 2b It is hypothesized that the learning profile can predict study unit outcomes with at least 80 % accuracy.

Finally, the purpose of a higher education institution's course is that students should attain a higher level of competence through a constant evolving and changing of individual dispositions as a result of learning experiences (Brabrand and Dahl 2009; Ifenthaler and Seel 2011; Robinson 2004). The learning profile has the potential to track the individual learning experiences (through reactive and non-reactive measures) and provide meaningful interventions towards successful study unit outcomes. Hence, a final focus of the second case study is the change of explained variance and accuracy of the learning profile during a specific study period.

Hypothesis 2c It is hypothesized that the explained variance of the learning profile increases over the course of the study period.

Hypothesis 2d It is hypothesized that the predictive accuracy of the learning profile increases over the course of the study period.

3 Case Study 1: Student Profile

This case study intended to use large existing datasets from multiple higher education institutions and area of studies in order to validate the student profile of the LA framework.

3.1 Method

3.1.1 Participants

The sample consisted of $N = 146{,}001$ students (54,073 male; 91,928 female) enrolled in 1,509 unique study units (1,030,778 total enrolments) with major higher education institutions in Australia. Their mean age was 33.06 years ($SD = 9.90$). 85 % of the participants reported that they completed secondary school. 5 % of the students reported having a disability. 94 % studied at undergraduate levels and 6 % at postgraduate levels.

3.1.2 Data Models

Table 1 shows the data models that were implemented for the student profile. The first model includes variables referring to the students' background and demographic data. Variables of student background include first language spoken, country of residence, and citizenship. Variables of demographic data include gender, age, socio-economic status, and disability. The second model includes the variables of model 1 plus parameters referring to the student's and family's historical education background such as completion of secondary school, highest education level of the student, and highest education level of the parents. The third model includes the variables of model 2 plus variables referring to information related to the study unit. Variables of study unit include undergraduate and postgraduate level study, study area, enrolment mode, delivery method, and study support utilized. The forth model includes the variables of model 3 plus student's historical education record with the institution such as time since last unit, study load, dropped and swapped study units. The fifth model includes the variables of model 4 plus the historical study performance of the student, i.e., average grade. The sixth and final model includes the essential variables identified from previous models. It is important to note that the current work-in-progress study does not include all variables of the above presented student profile (see Sect. 2.2). As the project is progressing, more variables will be included which will be added in future analysis.

3.2 Results

3.2.1 Explained Variance and Predictive Accuracy of the Student Profile Models

For each model of the student profile (see Table 1), we conducted a linear regression analysis and a SVM analysis to determine whether the student profile variables were significant predictors as well as showed acceptable accuracy for the study unit outcomes.

Table 2 shows the results of linear regression and SVM analysis for the six student profile models. The explained variance for predicting the study outcome increases from model 1 (R^2–SVR = .059) to Model 6 (R^2–SVR = .451). The findings suggest that variables included in the final student profile model 6 explain more than 40 % of variance. The most important variables of the final student profile model associated with study unit outcomes were the students' historical grade (43 % relative importance), historical cumulative fails (18 %), and highest level of prior education (10 %). Accordingly, the results support Hypothesis 1a.

As a next step, for each of the six models, a training set was randomly chosen to train the SVM classifier (Koggalage and Halgamuge 2004). Each classifier was trained with variables from the models shown in Table 1. We used fivefold cross validation to analyze prediction performance of the SVM classifier models. The predictive accuracy of each SVM classifier model is reported in Table 2. Classifier with variables from model 1 predicted the correct study unit outcome with an accuracy of 59 %. The classifier created with variables from model 6 which were determined as the most significant for the SVM regression models predicted the correct study unit outcome with an accuracy of 80 %. The training data contained students with no historical record with the institution. Since the historical record is a significant factor, large portion of the misclassifications were first time students. A classifier identical to model 6 and trained with data from students that have taken more than one study unit showed a final prediction accuracy of study unit outcome of 85 %. To sum up, the findings suggest that variables included in the final student profile model 6 account for 80 % accuracy for predicting study unit outcome. Accordingly, the results support Hypothesis 1b.

3.2.2 Idiosyncrasies of Student Profile Models

Table 3 shows the results of linear regression and SVM analysis for the student profile model 6 separated by eight higher education institutions. The explained variance for predicting the study outcome varies among the higher education institutions: Lowest R^2–SVR = .353 (UniR) and highest R^2–SVR = .489 (UniC), *SD* = .126. Accordingly, the results support Hypothesis 1c.

Similar results are shown in Table 4 that presents the linear regression and SVM analysis for the student profile model 6 separated by area of studies. Given the overall standard deviation of *SD* = .129, the lowest R^2–SVR = .359 was found for IT and highest R^2–SVR = .517 was found for Law and Justice. Accordingly, the results support Hypothesis 1d.

4 Case Study 2: Learning Profile

This case study intended to use interaction data of the learning management system from two study units of a higher education institution in order to validate the learning profile of the LA framework.

Table 1 Model descriptions for student profile

Model 1	Student background and demographic data
Model 2	Student background and demographic data
	Student's and parent's historical education background
Model 3	Student background and demographic data
	Student's and parent's historical education background
	Study unit related information
Model 4	Student background and demographic data
	Student's and parent's historical education background
	Study unit related information
	Historical education record with institution
Model 5	Student background and demographic data
	Student's and parent's historical education background
	Study unit related information
	Historical education record with institution
	Average historical grade within institution
Model 6	Most important parameters identified from previous models

Table 2 Student profile model performance comparison

	R^2	Adjusted R^2	R^2-SVR	Predictive accuracy (SVM) (%)
Model 1	.057	.057***	.059	58.63
Model 2	.128	.128***	.130	63.80
Model 3	.187	.187***	.192	67.50
Model 4	.361	.361***	.424	79.52
Model 5	.441	.446***	.438	79.69
Model 6	.444	.435***	.451	80.03

*** $p < .001$; *SVR* support vector regression, *SVM* support vector machines

Table 3 Student profile model performance comparison for higher education institutions

Higher Education Institution	R^2	Adjusted R^2	R^2-SVR	Predictive accuracy (SVM)
UniC	.464	.463***	.489	81.69 %
UniG	.453	.453***	.460	79.65 %
UniS	.431	.431***	.460	79.64 %
UniA	.372	.372***	.381	76.57 %
UniM	.438	.437***	.443	80.71 %
UniR	.364	.364***	.353	76.31 %
UniO	.434	.433***	.460	80.28 %
UniU	.372	.371***	.356	78.25 %
SD	.096	.096	.126	.024

*** $p < .001$; *SVR* support vector regression, *SVM* support vector machines

4.1 Method

4.1.1 Participants

A total of 12,686 enrolments of a major higher education institution in Australia were considered. Due to institutional regulations, detailed information on the participants (e.g., age, gender, socio-economic status, etc.) was not available. After cleaning the dataset, the final sample consisted of $N = 12{,}002$ students enrolled in two unique study units ($N = 4{,}978$ in unit A and $N = 7{,}024$ in unit B).

4.1.2 Data Models

Table 5 shows the data models that were implemented for the learning profile. The first model includes variables referring to the students' interaction with the online learning environment. Variables include access of learning materials, time spent, forum activities, and self-assessment attempts. The second model includes the variables of model 1 plus final assessment results.

4.2 Results

4.2.1 Explained Variance and Predictive Accuracy of the Learning Profile Model

For the learning profile model, data from the learning management system were analyzed for two study units from the Science and Engineering programs of a higher education institution. Table 6 shows the results of the linear regression and SVM analyses for the learning profile model 1. The explained variance for predicting the study outcome varied between the study units: Study unit A (R^2–SVR = .906) and Study unit B (R^2–SVR = .896). Despite the variability of the results, the findings suggest that variables included in the learning profile model 1 explain more than 80 % of variance. The most important variables of the learning profile model 1 associated with study unit outcomes were the assessment attempts (65 % relative importance), learning materials accessed (26 %), and self-assessments (7 %). Accordingly, the results support Hypothesis 2a. Additionally, the SVM regression models of the learning profile model 1 predicted the study unit outcome with an accuracy of over 90 % (see Table 6). Accordingly, the results support Hypothesis 2b.

4.2.2 Learning-Dependent Change of the Learning Profile Model

In order to investigate the change of the overall performance of the learning profile model 1, the study period was divided into four equal time periods. Table 7 shows the results of the linear regression and SVM analyses for the learning profile model 1 over the four time periods. The explained variance for predicting the study unit outcome increased from the initial interaction with the learning environment (R^2–SVR = .528) to the final interaction with the learning environment (R^2–SVR = .878). The findings suggest that as the learning profile model performance gains with the increased interaction over the study period. Accordingly, the results support Hypothesis 2c.

Additionally, the predictive accuracy of the learning profile increases over time with an accuracy of over 90 % (see Table 7). Accordingly, the results support Hypothesis 2d.

The results of a post hoc analysis including the final assessment outcomes (learning profile model 2) are shown in Table 8. The additional information included in the learning

Table 4 Student profile model performance comparison for area of studies

Area of studies	R^2	Adjusted R^2	R^2-SVR	Predictive accuracy (SVM)
Arts and Humanities	.430	.430***	.450	79.88 %
Business	.405	.405***	.436	78.02 %
Education	.489	.489***	.505	82.39 %
Law and Justice	.490	.490***	.517	82.69 %
IT	.373	.373***	.359	77.56 %
Science and Engineering	.423	.422***	.423	80.01 %
SD	.107	.107	.129	.027

*** $p < .001$; *SVR* support vector regression, *SVM* support vector machines

Table 5 Model descriptions for learning profile

Model 1	Assessment attempts, access of learning materials, videos seen, video seen repeated, time spent, time taken for assessments, forum discussions started, forum posts, forum replies, forum social polarity, forum positive votes, forum negative votes
Model 2	Assessment attempts, access of learning materials, videos seen, video seen repeated, time spent, time taken for assessments, forum discussions started, forum posts, forum replies, forum social polarity, forum positive votes, forum negative votes, final assessment outcomes

Table 6 Learning profile model performance comparison

	R^2	Adjusted R^2	R^2-SVR	Predictive accuracy (SVM) (%)
Study unit A	.880	.879***	.906	95.18
Study unit B	.845	.843***	.896	94.12
Combined	.854	.854***	.905	95.41

*** $p < .001$

profile model 2 has a major impact on the explained variance and predictive accuracy throughout the study period.

5 General Discussion

LA emphasizes insights and responses to real-time learning processes based on educational information from digital learning environments, administrative systems, and social platforms. However, well-established empirical evidence within the emerging field of LA is lacking. As new frameworks for LA are being developed across the education sector, we argue that they need to be empirically tested with regard to their reliability and validity before they may be implemented at larger scale.

The presented LA framework is a work-in-progress and being further developed and implemented within a major higher education institution in Australia. The two case studies provide empirical evidence for the implementation of the proposed student and learning profile. However, as the theoretical profiles have not been fully implemented yet, the results need to be interpreted as preliminary.

Table 7 Learning profile model 1 performance over the progression of a study period

	Time 1	Time 2	Time 3	Time 4
R^2				
Study unit A	.470	.764	.839	.862
Study unit B	.499	.759	.822	.838
Combined	.489	.762	.824	.845
Adjusted R^2				
Study unit A	.467***	.761***	.837***	.859***
Study unit B	.497***	.757***	.821***	.836***
Combined	.488***	.488***	.827***	.844***
R^2-SVR				
Study unit A	.497	.757	.840	.856
Study unit B	.542	.786	.845	.870
Combined	.528	.784	.860	.878
Predictive accuracy (SVM)				
Study unit A	74.98 %	87.54 %	93.26 %	94.67 %
Study unit B	76.94 %	88.07 %	93.51 %	94.33 %
Combined	76.44 %	88.79 %	93.83 %	94.63 %

*** $p < .001$

Table 8 Learning profile model 2 performance over the progression of a study period

	Time 1	Time 2	Time 3
R^2			
Study unit A	.647	.908	.962
Study unit B	.708	.907	.971
Combined	.702	.903	.966
Adjusted R^2			
Study unit A	.644***	.908***	.961***
Study unit B	.707***	.906***	.970***
Combined	.701***	.902***	.965***
R^2-SVR			
Study unit A	.580	.900	.973
Study unit B	.644	.863	.924
Combined	.632	.868	.925
Predictive accuracy (SVM)			
Study unit A	77.68 %	92.01 %	96.71 %
Study unit B	80.34 %	90.29 %	95.48 %
Combined	79.69 %	90.27 %	95.52 %

*** $p < .001$

The first case study focused on the student profile and identified variables which help to better provide early personalized interventions for students as well as facilitating their on-going learning progression towards successful learning outcomes (Fenwick and Cooper 2012; Lockyer et al. 2013). The most important variables associated with study unit outcomes were the students' historical grades and failures as well as their prior study experience. These results support the requirement for early student interventions in order to help students overcome initial problems and provide opportunities for student engagement when commencing higher education studies (Dobozy and Ifenthaler 2014; Thomas and May 2012). However, the findings also indicate that educational data is context specific and variables and indicators carry different meanings and can have different implications across educational institutions and area of studies (Coates 2010). This is evident through the differences of the performance of an identical analytical model in different institutions (e.g., 8 universities presented in Table 3) and different area of studies (e.g., 6 area of studies presented in Table 4). Therefore, a LA framework needs to be sensitive for idiosyncrasies of the educational institution and its stakeholders. Universal LA solutions using global algorithms may be biased and produce incorrect recommendations as well as inaccurate predictions. In adding the dynamic variables to the student profile (see Sect. 2.2) a more accurate performance of the algorithms is expected.

The second case study focused on the learning profile as it is assumed that learners should attain a higher level of competence through a constant evolving and changing of individual dispositions as a result of learning experiences (Brabrand and Dahl 2009). This learning progression is not to be a single, unique pathway to learning, rather, each learner will experience different learning activities, starting from different prior knowledge, and using individual strategies. Hence, it is important to understand the interaction of the learner with the learning environment in real-time in order to provide appropriate and meaningful interventions towards successful learning outcomes. The findings of the second case study indicate that increased data from the learning environments provides stronger evidence for more accurate predictions of students' pathways. Still, data from the first study period already helps to identify almost 50 % of variance of the learning profile. Hence, the initial days and weeks of a study unit provide important opportunities to address students' needs in order to help them to become more successful learners or to not drop out (Aflalo and Gabay 2012; Tinto 1999; Willging and Johnson 2009).

Additionally, the results of both case studies support the application of SVM for LA applications. The flexibility for modeling non-linear educational data, short training times for more robust models, responsiveness to interactions and changing variables, as well as sensitivity to imperfect data sets are strong arguments for further implementation of SVM in LA frameworks (Williams 2011).

To sum up, the findings of the two case studies provide initial but resilient evidence of the reliability, validity, and predictive accuracy of the student and learning profiles, however, the full strength of the LA framework lies in the combination of the student, learning, and curriculum profiles. Hence, limitations of the two case studies need to be addressed and further empirical research is required to replicate and advance the findings of the reported study.

5.1 Implications

The benefits of the holistic learning analytics framework can be associated with four levels of stakeholders: mega-level (governance), macro-level (institution), meso-level (curriculum, teacher/tutor), and micro-level (learner, OLE). An essential prerequisite for LA benefits, however, is the real-time access, analysis, and modeling of relevant educational information.

The mega-level facilitates cross-institutional analytics by incorporating data from all levels of the learning analytics framework. Such rich datasets enable the identification and validation of patterns within and across institutions and therefore provide valuable insights for informing educational policymaking. The macro-level enables institution-wide analytics for better understanding learner cohorts for optimizing associated processes and allocating critical resources for reducing dropout and increasing retention as well as success rates. The meso-level supports the design of the curriculum and related learning materials as well as provides detailed insights about learning processes for course facilitators (i.e., teachers, tutors). This information can be used for improving the overall quality of courses (e.g., sequencing of learning processes, alignment with higher level outcomes and competencies) as well as enhancing learning materials (e.g., their alignment to anticipated learning outcomes and associated assessments). The micro-level analytics supports the learner through recommendations and help functions implemented in the digital learning environment. This may include personalized and adaptive scaffolds that are expected to be more successful for achieving expected learning outcomes and competencies. Another critical component for improving the benefits of LA is information from the physical environment (e.g., learner's current emotional state) which is not directly linked with the educational data. Accordingly, data may be collected within the digital learning environment through reactive prompts and linked with the available educational information.

Table 9 provides a matrix outlining the benefits of LA for stakeholders including three perspectives (Ifenthaler in press): (1) summative, (2) real-time, and (3) predictive. The summative perspective provides detailed insights after completion of a learning phase (e.g., study period, semester, final degree), often compared against previously defined reference points or benchmarks. The real-time perspective uses ongoing information for improving processes through direct interventions. The predictive perspective is applied for forecasting the probability of outcomes in order to plan for future strategies and immediate actions. The benefits matrix enables decision makers to analyze the requirements of LA within an institution and further guide the implementation of a LA framework and strategy at different levels (Ifenthaler in press).

5.2 Limitations and Future Work

The presented research is a work-in-progress providing initial insights into the conceptual development of a holistic LA framework and its empirical validation. Not all variables from the student and learning profiles have been added to the currently implemented learning analytics application yet. As the project is evolving, more and more variables will be included and data collected accordingly. Additionally, there are limitations to the empirical study, which need to be addressed.

First, while our sample size was large enough to achieve statistically significant results, the explained variance for some of our regression models require careful interpretation. This indicates that besides the tested variables other variables may have influenced the outcomes that were not tested in the reported case study. Second, the development of the holistic LA framework is still in progress. Therefore, we tested the student profile and learning profile separately due to lack of being able to link the available data, future research will include rich combined data from the student, learning, and curriculum profiles which will add substantially towards the explained variance of the proposed models. Third, the predictions are only valid for individual study unit outcomes; however, do not reflect higher education outcomes in general. Accordingly, further studies will be needed to cross-validate the initial results of this study. Forth, the expected explained variance reflected in the hypotheses is based on standards within social sciences as no previous empirical findings are available for learning analytics. Future studies may critically review

Table 9 LA benefits matrix

Stakeholder	Perspective		
	Summative	Real-time	Predictive
Governance	Apply cross-institutional comparisons Develop benchmarks Inform policy making Inform quality assurance processes	Increase productivity Apply rapid response to critical incidents Analyze performance	Model impact of organizational decision-making Plan for change management
Institution	Analyze processes Optimize resource allocation Meet institutional standards Compare units across programs and faculties	Monitor processes Evaluate resources Track enrolments Analyze churn	Forecast processes Project attrition Model retention rates Identify gaps
Instructional design	Analyze pedagogical models Measure impact of interventions Increase quality of curriculum	Compare learning designs Evaluate learning materials Adjust difficulty levels Provide resources required by learners	Identify learning preferences Plan for future interventions Model difficulty levels Model pathways
Facilitator	Compare learners, cohorts and courses Analyze teaching practices Increase quality of teaching	Monitor learning progression Create meaningful interventions Increase interaction Modify content to meet cohorts' needs	Identify learners at risk Forecast learning progression Plan interventions Model success rates
Learner	Understand learning habits Compare learning paths Analyze learning outcomes Track progress towards goals	Receive automated interventions and scaffolds Take assessments including just-in-time feedback	Optimize learning paths Adapt to recommendations Increase engagement Increase success rates

the initially set numbers and provide further evidence for expected explained variance in learning analytics applications. Fifth, it is also important to note that the presentation of linear and non-linear model fit results shall highlight the necessity of alternative data analytics approaches for learning analytics applications. Evidently, a direct comparison of linear and non-linear approaches using regression to fit is different for both approaches.

Future work includes empirical validation of all profiles and a full implementation of the holistic LA framework as a dynamic plug-in for digital learning environments. A further iteration of the LA framework will include a natural language processing (NLP) approach which will be utilized for analyzing discussion forums and providing recommendations of social interaction (Dawson et al. 2011; Macfadyen and Dawson 2010) and rich semantic feedback in near real-time (Ifenthaler and Pirnay-Dummer 2011; Pirnay-Dummer and Ifenthaler 2011a, b).

5.3 Concerns and Challenges

Besides the above-described benefits of LA, serious concerns and challenges are associated with the application of LA:

1. Not all educational data is relevant and equivalent (Macfadyen and Dawson 2012; Thompson et al. in press). Therefore, the validity of data and its analyses is critical for generating useful summative, real-time, and predictive insights. This generates a new interdisciplinary research area for cognitive psychology, educational technology, learning design, psychometrics, data management, artificial intelligence, web development, and statistics. The challenges are to investigate the complex processes within LA frameworks and to understand their immediate and long-term effects on learning and teaching processes.
2. Ethical issues are associated with the use of educational data for LA (Slade and Prinsloo in press). That implies how personal data is collected and stored as well as how it is analyzed and presented to different stakeholders. Hence, procedures regulating access and usage of educational data need to come into operation before LA frameworks are implemented. This will also include transparency of applied algorithms and weighting of educational data for predictive modeling. Storing and processing anonymized personal data is only a small step towards a more comprehensive educational data governance structure for LA.
3. Limited access to educational data generates disadvantages for involved stakeholders. For example, invalid forecasts may lead to inefficient decisions and unforeseen problems. A misalignment of prior knowledge, learning pathways, and learning outcomes could increase churn and the late identification of learners at risk may create dropouts. A definition of threshold standards for LA could prevent vast gaps between educational institutions and provide equal opportunities for all stakeholders.
4. The preparation of stakeholders for applying insights from LA in a meaningful way is vital. Professional development for stakeholders ensures that issues are identified and benefits are transformed into meaningful action. Hence, the increased application of LA requires a new generation of experts with unique interdisciplinary competences. This will also require new infrastructures for administration and research in order to accelerate the understanding of LA.
5. Information from distributed networks and unstructured data cannot be directly linked to educational data collected within an institution's environment. An aggregation of such data and uncontrolled relations to existing educational data increases the chance of critical biases as well as invalid analysis, predictions, and decisions. The challenge is to develop mechanisms to filter biased information and warn stakeholders accordingly.
6. An optimal sequence of data collection and economic response times (seconds, minutes, hours, days, weeks) of LA have yet to be determined. This includes the minimum requirements for making valid predictions and creating meaningful interventions. Missing data is a critical challenge for future LA algorithms.
7. Besides the analysis of numerical data (e.g., click streams), a qualitative analysis of semantic rich data (e.g., content of discussion forums, responses to open-ended assessments) enables a better understanding of learners' knowledge and needs. An obvious requirement is the development of automated natural language processing (NLP) capabilities. The major challenge besides the development of real-time NLP is the validation of such algorithms and the link to quantitative educational data.

6 Conclusions

More educational data does not always make better educational data (Greller and Drachsler 2012; Macfadyen and Dawson 2012). Hence, LA has its obvious limitations and data collected from various educational sources can have multiple meanings. Empirically validating LA frameworks and corresponding profiles such as presented in the two case studies may provide evidence for the implementation of intelligent systems which have the capabilities to facilitate learning of individual students, improve instructional practice of teachers, and advance the quality of higher education offerings of individual intuitions and across the education sector.

References

Aflalo, E., & Gabay, E. (2012). An information system for dropout prevention. *Education and Information Technologies, 17*(2), 233–250. doi:10.1007/s10639-011-9156-x.

Allen, C. B., Higgs, Z. R., & Holloway, J. R. (1988). Identifying students at risk for academic difficulty. *Journal of Professional Nursing, 4*(2), 113–118. doi:10.1016/S8755-7223(88)80033-4.

Ashby, F. G. (Ed.). (1992). *Multidimensional models of perception and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science, 33*(5–6), 381–412.

Bartholomew, D. J. (1967). *Stochastic models for social processes*. New York: Wiley.

Bauer, R. (1966). *Social indicators*. Cambridge, MA: MIT Press.

Brabrand, C., & Dahl, B. (2009). Using the SOLO taxonomy to analyze competence progression of university science curricula. *Higher Education, 58*(4), 531–549. doi:10.1007/s10734-009-9210-4.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi:10.1023/A:1010933404324.

Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Educational Technology and Society, 15*(3), 3–26.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. (2010). Academic analytics: A new tool for a new era. *EDUCAUSE Review, 42*(4), 40–57.

Christmann, A., & Steinwart, I. (2008). *Support vector machines*. New York: Springer.

Cleophas, T. J., & Zwinderman, A. H. (2013). Support vector machines. *Machine learning in medicine* (pp. 155–161). Amsterdam: Springer.

Coates, H. (2009). What's the difference? A model for measuring the value added by higher education in Australia. *Higher Education Management and Policy, 21*(1), 69–88.

Coates, H. (2010). Defining and monitoring standards in Australian higher education. *Higher Education Management and Policy, 22*(1), 41–58.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. doi:10.1007/bf00994018.

Crosling, G., Heagney, M., & Thomas, L. (2009). Improving student retention in higher education. *Australian Universities' Review, 51*(2), 9–18.

d'Aquin, M., Dietze, S., Herder, E., Drachsler, H., & Taibi, D. (2014). Using linked data in learning analytics. *E-Learning Papers, 36*, 1–9.

da Silva, J. L., Caeiro, F., Natário, I., & Braumann, C. A. (2013). *Advances in regression, survival analysis, extreme values, markov processes and other statistical applications*. Berlin: Springer.

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.

Dobozy, E., & Ifenthaler, D. (2014). Initial teacher education by open and distance modes: A snapshot of e-competency experiences in Australia. *eLearning Papers, 38*, 43–54.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems 9* (pp. 155–161). Cambridge, MA: MIT Press.

Fenwick, L., & Cooper, M. (2012). Prevailing pedagogies for classes in low SES contexts and the implications for standards-based reform in Australia. *The Australian Educational Researcher, 39*(3), 349–361. doi:10.1007/s13384-012-0066-8.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317. doi:10.1504/IJTEL.2012.051816.

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

Ifenthaler, D. (in press). Learning analytics. In J. M. Spector (Ed.), *Encyclopedia of educational technology*. Thousand Oaks, CA: Sage.

Ifenthaler, D., & Pirnay-Dummer, P. (2011). States and processes of learning communities: Engaging students in meaningful reflection and elaboration. In B. White, I. King, & P. Tsang (Eds.), *Social media tools and platforms in learning environments: Present and future* (pp. 81–94). New York: Springer.

Ifenthaler, D., Pirnay-Dummer, P., & Seel, N. M. (Eds.). (2010). *Computer-based diagnostics and systematic analysis of knowledge*. New York: Springer.

Ifenthaler, D., & Seel, N. M. (2011). A longitudinal perspective on inductive reasoning tasks. Illuminating the probability of change. *Learning and Instruction, 21*(4), 538–549. doi:10.1016/j.learninstruc.2010.08.004.

Ifenthaler, D., & Seel, N. M. (2013). Model-based reasoning. *Computers and Education, 64*, 131–142. doi:10.1016/j.compedu.2012.11.014.

James, R., Krause, K.-L., & Jennings, C. (2010). *The first-year experience in Australian universities: Findings from 1994 to 2009*. Melbourne, VIC: Centre for the Study of Higher Education.

Johnson, L., Adams Becker, S., Cummins, M., Freeman, A., Ifenthaler, D., & Vardaxis, N. (2013). *Technology outlook for Australian tertiary education 2013–2018: An NMC horizon project regional analysis*. Austin, TX: The New Media Consortium.

Kalyuga, S. (2006). Assessment of learners' organised knowledge structures in adaptive learning environments. *Applied Cognitive Psychology, 20*, 333–342.

Koggalage, R., & Halgamuge, S. (2004). Reducing the number of training samples for fast support vector machine classification. *Neural Information Processing-Letters and Reviews, 2*(3), 57–65.

Lin, C. F., Yeh, Y.-C., Hung, Y. H., & Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers and Education, 68*, 199–210. doi:10.1016/j.compedu.2013.05.009.

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist, 57*(10), 1439–1459. doi:10.1177/0002764213479367.

Long, P. D., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 31–40.

Macfadyen, L., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education, 54*(2), 588–599.

Macfadyen, L., & Dawson, S. (2012). Numbers are not enough. Why e-Learning analytics failed to inform an institutional strategic plan. *Educational Technology and Society, 15*(3), 149–163.

Perumallaa, C., Maka, J., Keea, N., & Matthewsa, S. (2010). Integrating web applications to provide an effective distance online learning environment for students. *Procedia Computer Science, 3*, 770–784. doi:10.1016/j.procs.2010.12.127.

Pirnay-Dummer, P., & Ifenthaler, D. (2011a). Reading guided by automated graphical representations: How model-based text visualizations facilitate learning in reading comprehension tasks. *Instructional Science, 39*(6), 901–919. doi:10.1007/s11251-010-9153-2.

Pirnay-Dummer, P., & Ifenthaler, D. (2011b). Text-guided automated self assessment: A graph-based approach to help learners with ongoing writing. In D. Ifenthaler, P. I. Kinshuk, D. G. Sampson, & J. M. Spector (Eds.), *Multiple perspectives on problem solving and learning in the digital age* (pp. 217–225). New York: Springer.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. doi:10.1023/A:1022643204877.

Robinson, R. (2004). Pathways to completion: Patterns of progression through a university degree. *Higher Education, 47*(1), 1–20. doi:10.1023/B:HIGH.0000009803.70418.9c.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. D. (Eds.). (2011). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.

Schreurs, B., de Laat, M., Teplovs, C., & Voogd, S. (2014). Social learning analytics applied in a MOOC-environment. *e-Learning Papers, 26*, 45–48.

Slade, S., & Prinsloo, P. (in press). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, doi:10.1177/0002764213479366.

Thomas, L. (2011). Engaging students to improve retention and success. In L. Thomas, & M. Tight (Eds.), *Institutional transformation to engage a diverse student body* (Vol. 6, pp. 41–55, International perspectives on higher education research). Bingley: Emerald Group Publishing Limited.

Thompson, K., Ashe, D., Carvalho, L., Goodyear, P., Kelly, N., & Parisio, M. (in press). Processing and visualizing data in complex learning environments. *American Behavioral Scientist,* doi:10.1177/0002764213479368.

Tinto, V. (1982). Limits of theory and practice I student attrition. *The Journal of Higher Education, 53*(6), 687–700.

Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. *NACADA Journal, 19*(2), 5–9. doi:10.12930/0271-9517-19.2.5.

Willging, P. A., & Johnson, S. D. (2009). Factors that influence students' decision to dropout of online courses. *Journal of Asynchronous Learning Networks, 13*(3), 115–127.

Williams, G. (2011). Support vector machines. In Data mining with Rattle and R (pp. 293-304, Use R). New York: Springer.