

Integrating Data Mining in Program Evaluation of K-12 Online Education

Jui-Long Hung^{*}, Yu-Chang Hsu and Kerry Rice

Department of Educational Technology, Boise State University, Boise, USA // andyhung@boisestate.edu // hsu@boisestate.edu // krice@boisestate.edu

^{*}Corresponding author

ABSTRACT

This study investigated an innovative approach of program evaluation through analyses of student learning logs, demographic data, and end-of-course evaluation surveys in an online K–12 supplemental program. The results support the development of a program evaluation model for decision making on teaching and learning at the K–12 level. A case study was conducted with a total of 7,539 students (whose activities resulted in 23,854,527 learning logs in 883 courses). Clustering analysis was applied to reveal students' shared characteristics, and decision tree analysis was applied to predict student performance and satisfaction levels toward course and instructor. This study demonstrated how data mining can be incorporated into program evaluation in order to generate in-depth information for decision making. In addition, it explored potential EDM applications at the K-12 level that have already been broadly adopted in higher education institutions.

Keywords

Educational data mining, Program evaluation, K–12 virtual school, Pattern discovery, Predictive modeling

Introduction

Traditionally, the majority of online instructors and institutional administrators rely on web-based course evaluation surveys to evaluate online courses (Hoffman, 2003). The data and information are then used to help inform online program effectiveness and generate information for program-level decision-making. While it enjoys wide use, the survey method only provides learners' self-report data, not their actual learning behaviors.

Several studies have found self-reported data were not consistent with actual learning behaviors (Hung & Crooks, 2009; Picciano, 2002). This inconsistency can potentially compound the already problematic lack of direct observation opportunities. Online program administrators need more effective tools to provide customized learning experiences, to track students' online learning activities for overseeing courses (Delavari, Phon-amnuaisuk, & Beikzadeh, 2008), to depict students' general learning characteristics (Wu & Leung, 2002), to identify struggling students (Ueno, 2006), to study trends across courses and/or years (Hung & Crooks, 2009), and to implement institutional strategies (Becker, Ghedini, & Terra, 2000). Each of these needs can be addressed by mining educational data. Nowadays, various educational data are stored in database systems. This is especially true for online programs, wherein student learning behaviors are recorded and stored in Learning Management Systems (LMS). Program administrators can take advantage of emerging knowledge and skills by extracting and interpreting those data. The purpose of this study is to propose a program evaluation framework using Educational data mining.

Program evaluation

Program evaluation is the means by which a program assures itself, its administration, accrediting organizations, and students that it is achieving the goals delineated in its mission statement (Nichols & Nichols, 2000). Evaluation can be done by a variety of means. The most common form of evaluation is through surveying students regarding courses/faculty/programs (e.g., Cheng, 2001; Hoffman, 2003; Spirduso & Reeve, 2011). However, making causal inferences based on a one-time assessment is risky (Astin & Lee, 2003). Nevertheless, perceptual survey data cannot accurately reflect real learning behaviors (Hung & Crooks, 2009; Picciano, 2002). Although various scholars (e.g., Grammatikopoulous, 2012; Vogt & Slish, 2011) have proposed systematic frameworks (e.g., interviews and observation) in order to obtain objective knowledge via multiple means, these methods are difficult to implement in a fully online program.

Educational data mining

Data mining (DM) is a series of data analysis techniques applied to extract hidden knowledge from server log data (Roiger & Geatz, 2003) by performing two major tasks: Pattern discovery and predictive modeling (Panov, Soldatova, & Dzeroski, 2009). Educational data mining (EDM) is a field which adopts data mining algorithms to solve educational issues (Romero & Ventura, 2010). Romero & Ventura (2010) reviewed 306 EDM articles from 1993 to 2009 and proposed desired EDM objectives based on the roles of users. For the purpose of this study, which is designed to inform administrators, the list is limited to objectives for administrators:

- Enhance the decision processes in higher learning institutions
- Streamline efficiency in the decision making process
- Achieve specific objectives
- Suggest certain courses that might be valuable for each class of learners
- Find the most course effective way of improving retention and grades
- Select the most qualified applicants for graduation
- Help to admit students who will do well in higher education settings

Based on the theory of bounded rationality, decision-making is a fully rational process of finding an optimal choice given the information available (Elster, 1983). An ideal program evaluation framework should provide multiple facets of information to decision makers. Therefore, integrating more than one data source and analytic method is essential for an effective program evaluation.

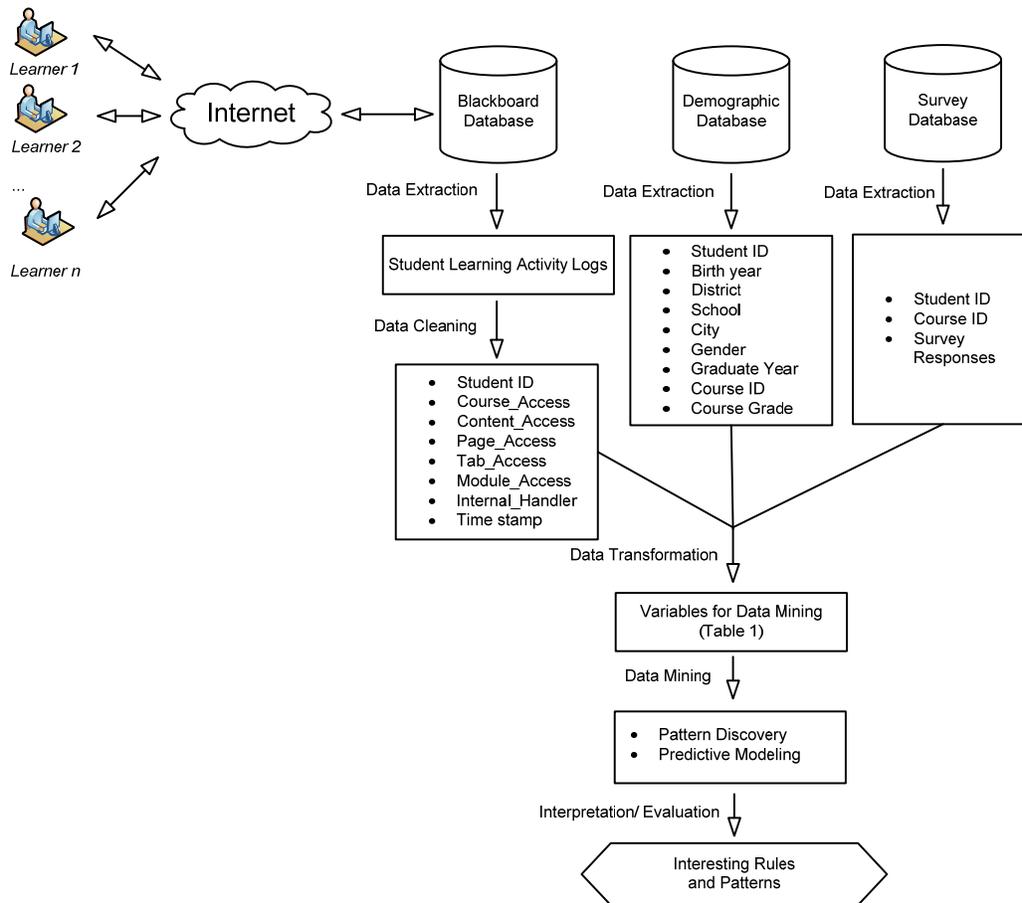


Figure 1. Program evaluation framework

Program evaluation framework

Figure 1 shows the framework of the proposed program evaluation method. The core strategy of this framework is data triangulation (Jick, 1979) which combines multiple data sources (learning logs, course evaluation survey, and demographic data) and multiple methods (pattern discovery and predictive modeling) to generate accurate, in-depth results. Using this framework, the authors conducted a program evaluation case study to evaluate how the proposed program evaluation framework can support administrators' decision making.

Method

Data source

In this case study, data were collected from a statewide K–12 online institution that serves over 16,000 students in a northwestern state in the U.S. The institution provides fully online courses to K–12 students. Courses were designed by subject-matter curriculum designers and subject-matter teachers to standardize course materials. Teachers were required to complete an online orientation prior to teaching courses for the institution. Teachers received the same or similar training for online teaching provided by the institution. Site coordinators are located at each district in the state and regional principals oversee teacher evaluation. The following data were collected for the academic year of 2009–2010 (3,604 students enrolled in Fall 2009 and 3,935 students in Spring 2010): 1) LMS activity logs; 2) student demographic data; and 3) course evaluation survey data. All data tables were stored in the database and interconnected with unique identifiers (e.g., course ID).

LMS activity logs

The LMS activity logs were collected from the Blackboard activity accumulator (Blackboard Inc., 2010) for the Fall 2009 and Spring 2010 academic terms. The following records were removed in data preprocessing: irrelevant fields (e.g., group ID), irrelevant records (e.g., login failure), and data stored in wrong or mismatched fields (about 11.8% of overall activity logs). After data preprocessing, a total of 23,854,527 activity logs were collected from 7,539 students in 883 courses. These students took 1 to 18 courses in the 2009–2010 academic year.

Student demographic data

The following demographic data were collected for data analysis, including age, gender, graduation year, city, school district, number of online course(s) taken, number of online course(s) passed, number of online course(s) failed, and final grade average.

Course evaluation survey

A course evaluation survey investigated students' satisfaction toward their course and instructor. Course satisfaction contained eight questions related to course content, five related to course structure, and eleven related to instructor satisfaction. Records containing any missing values were removed from the analysis. In addition, because student identifiers were not collected during the Fall 2009 survey implementation, which prevents the researchers from associate survey responses with demographic data and LMS activity logs, only Spring 2010 survey data (2618 respondents) were analyzed in this study.

Engagement level

Engagement is considered to be a key variable for enabling and encouraging learners to interact with the material, with the instructor, and with one another, as well as for learning in general. In this study, engagement level was measured by the frequency of various learning interactions that happened within the LMS. Variables under the category "Student Engagement Variable" in Table 1 were applied to measure each student's engagement level, which included:

- Average frequency of logins per course.
- Average frequency of tab accessed per course (if the course was organized using “tabbed” navigation).
- Average frequency of module accessed per course (if the course was organized using “modules”).
- Average frequency of clicks per course.
- Average frequency of course accessed per course (from Blackboard portal to course site).
- Average frequency of page accessed per course (content created using the page tool). The Page tool allows instructors to include files, images, and text as links on the course menu
- Average frequency of course content accessed per course (content created using the content tool). The Content tool allows instructors to create course content within the content area.
- Average number of discussion board entries per course.

Variables

Table 1 lists variables collected from Blackboard, the student demographic database, and the course evaluation survey. Some variables were transformed with calculations in order to generate more meaningful variables for analysis. For example, student’s birth year was transformed to age. The summary of all learning activities was aggregated to a new variable called “frequency of clicks” that represents each student’s total frequency of clicks in the Blackboard LMS. If students took more than one course during the analysis period, variables of learning activities (e.g., frequency of total clicks and frequency of course access), performance (e.g., final grade), and survey (e.g., course satisfaction and instructor satisfaction) were averaged.

Table 1. Variables for data mining

Variables	Descriptions	Category
stuID	Student’s ID	Student Demographic Variable
Age	Student’s age	Student Demographic Variable
City	Student’s residential city	Student Demographic Variable
District	Student’s school district	Student Demographic Variable
Grade_Avg	Average course grade	Student Performance Variable
Click_Avg	Average frequency of clicks/course	Student Engagement Variable
Content_Access_Avg	Average frequency of course content accessed/course	Student Engagement Variable
Course_Access_Avg	Average frequency of course accessed/course	Student Engagement Variable
Page_Access_Avg	Average frequency of page accessed/course	Student Engagement Variable
DB_Entry_Avg	Average number of discussion board entries/course	Student Engagement Variable
Tab_Access_Avg	Average frequency of tab accessed/course	Student Engagement Variable
Login_Avg	Average frequency of logins/course	Student Engagement Variable
Module_Avg	Average frequency of module accessed/course	Student Engagement Variable
Gender	Gender	Student Demographic Variable
HSGradYear	High school graduation year	Student Demographic Variable
School	Student’s school	Student Demographic Variable
No_Course	Number of course taken	Student Demographic Variable
No_Fail	Number of course failed	Student Performance Variable
No_Pass	Number of course passed	Student Performance Variable
Pass rate	Average individual student pass rate for all courses in academic year 2009-2010 (≥ 0 and ≤ 1)	Student Performance Variable
cSatisfaction_Avg	Average course satisfaction including 8 questions related to course content and 5 questions related to course structure	Student Perception Variable
iSatisfaction_Avg	Average instructor satisfaction including 11 questions related to instructor.	Student Perception Variable

Analytic tools

SAS Enterprise Miner 6.1 (SAS Institute Inc., USA) was employed to perform the following data mining tasks in this study: 1) student clustering which describes shared characteristics of students who passed or failed their course; 2) perception and performance predictions which identify key predictors of course satisfaction, instruction satisfaction, and final grade. Because one of the major target audiences of this article is K–12 administrators, the authors utilized methods like decision tree and K-means clustering, which can produce results more intuitive for non-data miners.

Results

Student clustering

Clustering analysis

K-means algorithm (Hartigan & Wong, 1979; Budayan, Dikmen, & Birgonul, 2009) was applied to group students based on their shared characteristics (Internal Standardization = Range; Maximum Number of Clusters = 6). Total clusters were limited to avoid trivially small or exclusive groups, the identification of which was outside the purposes of this case study. A pass rate equal to “1” means a student passed all courses during the period of analysis. A pass rate equal to “0” means a student failed all courses during the period of analysis. A pass rate between “0” and “1” means a student passed some, but not all, courses during the period of analysis. In clustering analysis, pass rate was set up as the standard for classification and six clusters were generated.

Table 2 includes the results of clustering analysis in academic year 2009-2010. The following are shared characteristics of each cluster.

- Cluster 1 (316 students, pass rate = 55.07%, all males): Cluster 1 consists of students who are older than Cluster 3 to 6. They were lower-engaged than Cluster 5 and 6 but higher than Cluster 3 and 4. On average, each student took 2.76 courses and failed about half of them.
- Cluster 2 (320 students, pass rate = 56.11%, all females): Similar to Cluster 1, Cluster 2 consists of students who are older than Clusters 3 to 6. They are lower-engaged than Cluster 5 and 6 but higher than Cluster 3 and 4. On average, each student took 3.03 courses and failed about half of the courses.
- Cluster 3 (594 students, pass rate = 0%, all males): Cluster 3 and 4 includes the lowest-engaged students. Cluster 3 students are all male. On average, each student took 1.43 courses and failed all of them.
- Cluster 4 (601 students, pass rate = 0%, all females): Cluster 4 includes the lowest-engaged female students. On average, each student took 1.39 courses and failed all of them.
- Cluster 5 (2,311 students, pass rate = 100%, all males): Cluster 5 and 6 represent the highest-engaged students. Cluster 5 students are all male. On average, each student took 1.59 courses and passed all of them.
- Cluster 6 (3,397 students, pass rate = 100%, all females): Cluster 6 represents the highest-engaged female students. On average, each student took 1.64 courses and passed all of them.

Table 2. Results of clustering analysis

Variables	CL1	CL2	CL3	CL4	CL5	CL6
Number of Students	316	320	594	601	2311	3397
Pass rate = 0	0	0	594	601	0	0
Pass rate (> 0 and < 1)	316	320	0	0	0	0
Pass rate = 1	0	0	0	0	2311	3397
GenderF	0	320	0	601	0	3397
GenderM	316	0	594	0	2311	0
Age	16.91	17.06	16.69	16.82	16.6	16.59
Grade_Avg	50.11	52.82	22.44	20.85	81.75	85.4
Click_Avg	583.15	549.09	440.17	416.4	892.49	881.69
Content_Access_Avg	112.96	112.2	93.78	89.43	180.34	177.96
Course_Access_Avg	170.26	172.29	133.94	141.52	281.5	284.22
DB_Entry_Avg	4.08	5.28	2.78	4.22	8.28	9.57
Login_Avg	29.4	24.35	23.58	19.18	47.92	46.42
Module_Access_Avg	156.18	145.02	112.7	102.38	249.16	240.79

Page_Access_Avg	99.61	89.39	71.97	62.46	145.43	142.42
Tab_Access_Avg	41.92	37.82	29.98	26.03	62.04	60.72
No_Fail	1.23	1.33	1.43	1.39	0	0
No_Pass	1.52	1.7	0	0	1.59	1.64
No_course	2.76	3.03	1.43	1.39	1.59	1.64

The clusters generated from cluster analysis were associated with two geographical variables: city and school district, in order to examine whether certain types of students were from specific areas. Differences in engagement were found depending on location. Clusters 1 to 6 had similar geographical distributions except for three larger cities (populations larger than 100,000). Cluster 5 (all male, pass rate = 100%) included a larger group of students from one large city. Cluster 6 (all female, pass rate = 100%) included a larger group of students from the other two large cities. There is no notable difference of school district distributions across clusters.

Findings

Findings below were summarized from the clustering analysis.

- 1) Students with higher engagement levels usually had higher performance.
- 2) Younger students (CLs 5 & 6) who lived in larger cities were more successful than those in smaller cities (CLs 3 & 4) and older students (CLs 1 & 2).
- 3) All-failed students who were also low-engaged consisted of approximately 15.9% on average per course.
- 4) All-passed students who were also high-engaged consisted of approximately 75.7% students on average per course.
- 5) Based on Cluster 1 and 2, on average, older students (age > 16.91) tended to take more than two courses with pass rates ranging from 54.09-56.11%.
- 6) On average, high-engaged students demonstrated engagement levels twice that of low-engaged students.
- 7) Frequencies of reading behaviors (such as content access and page access) were much higher than discussion behaviors ($p < 0.001$).
- 8) Female students were more active than male students in online discussions (with higher DB_Entry avg frequency).
- 9) Female students had higher pass rates than male students.

Average clicks per course in different subject areas

Table 3 shows students' average frequencies of total clicks and performances per course in different subject areas. Total clicks were equal to the summarized frequency of overall learning activities. The results show that Math and Science had the highest number of total clicks per course and of total clicks per student per course. However, for those who took Math and/or Science courses, their average final grades (56.70 and 64.41 accordingly) were lower than the overall final grade average (71.11). This indicates students participated actively in courses of these two subject areas, but they failed to achieve expected outcomes (70 or higher). Possible reasons for this outcome could be related to course design and/or best practice in teaching strategies for Math and Science courses. On the other hand, English courses received a lower number of clicks combined with less than expected outcomes. Encouraging motivation and engagement in these courses could have a profound effect on future outcomes. Students in Foreign Language and Health not only participated in learning activities actively, but also obtained the highest grades, on average, in each of these two subject areas.

Table 3. Average frequencies of total clicks and performance in different subject areas.

Subject Areas	Total Clicks	Total Clicks/student	No of Students	Final Grade
Drivers Ed	4,808.00	227.97	21.09	78.40
Electives	5,353.63	247.69	21.61	76.59
English	4,807.79	239.98	20.03	62.09
Foreign Language	7,824.63	439.40	17.81	76.54
Health	6,641.80	269.99	24.60	83.58
Math	7,898.35	444.05	17.79	56.70
Science	9,015.16	603.53	14.94	64.41

Social Studies	4,740.92	235.97	20.09	70.58
Average	6,386.29	323.43	19.75	71.11

Findings

- 10) Subjects where the level of activity was effective and consistent with student outcomes included Driver Education, Electives, Foreign Language, Health, and Social Studies.
- 11) Subjects where the level of activity was inconsistent with student outcomes included Math, Science and English. Math and Science courses had high activity levels with less than expected outcomes.
- 12) Subjects where the level of activity was low and consistent with low student outcomes included English.

Subject preferences

Figure 2 shows percentages of female and male students in different subject areas. Because the original female versus male ratio is 1.34, subject preferences for female and male students were revealed by comparing female/male ratio in each subject with the original ratio. Subjects above the dashed line are those with higher female ratios.

Findings

- 13) Female students preferred taking Electives, Foreign Language, and Social Studies.
- 14) Male students preferred taking Drivers Education, Math, and Science.

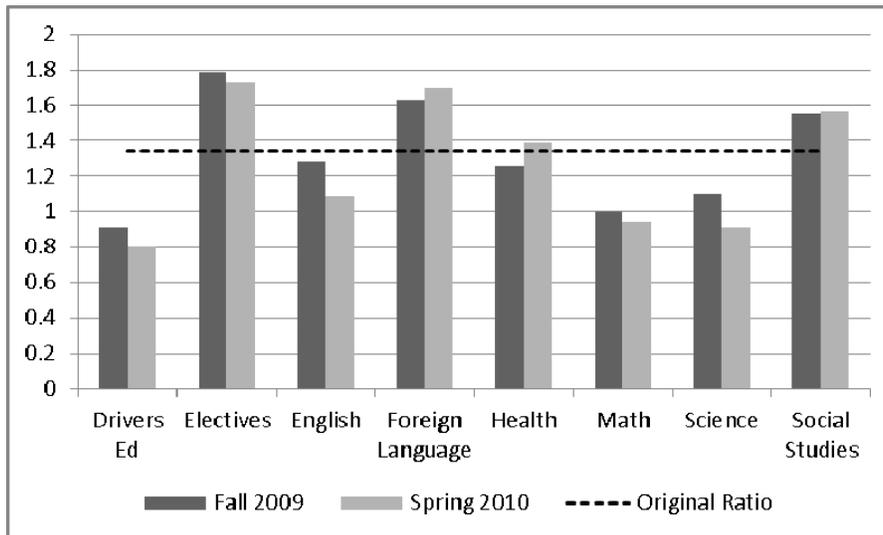


Figure 2. Gender preferences (female students/male students) by subject areas

Pass rate in different subject areas

Table 4 consists of two parts. The first part examines whether pass rates of female and male students in different subjects have significant differences. “F vs. M” compares gender pass rate difference using t-tests. The second part examines pass rate difference between Fall 2009 and Spring 2010 within the same gender. For example, “F vs. F” compares pass rate difference between Fall 2009 and Spring 2010 female students in difference subjects by using t-tests. Numbers marked with asterisks represent differences that have statistical significance.

Table 4. Pass rate comparisons and statistical tests by gender and subject areas

Gender	Fall 2009			Spring 2010			Fall 2009 vs. Spring 2010	
	F	M	F vs. M (p values)	F	M	F vs. M (p values)	F vs. F (p values)	M vs. M (p values)
Drivers Ed	91%	89%	.623	96%	86%	.044*	.156	.779
Electives	85%	79%	.000*	91%	87%	.000*	.025*	.019*

English	69%	57%	.000*	83%	72%	.000*	.000*	.000*
Foreign Language	83%	80%	.088	91%	89%	.121	.006*	.036*
Health	92%	90%	.023*	97%	97%	.058	.067	.084
Math	55%	54%	.769	72%	70%	.761	.000*	.000*
Science	71%	67%	.223	85%	81%	.418	.000*	.000*
Social Studies	79%	74%	.000*	89%	84%	.000*	.000*	.002*

Note: Statistical significance refers to the possible accurate rate of a statement by testing it with statistical methods. “ $p < .05$ ” means the statement is at least 95% accurate (error rate is less than 5%). * $p < .05$, ** $p < .001$

Findings

- 15) Overall, females significantly performed better than male students, especially in the following subject areas: Electives, English, and Social Studies.
- 16) The fail rates during the Fall 2009 term was significantly higher than those during the Spring 2010 term, especially in those subjects with higher fail rates such as English, Math, Science, and Social Studies. After these results were revealed the researchers subsequently learned from the administrators of the program reported in this case study that they had adopted an early alarm system in Spring 2010 to track all communications between instructors and students. The results show those strategies could have improved students’ pass rates in most subject areas.

Student performance and engagement by course number

Due to the previous results indicating that students in Math, Science, and English had lower performance than those in other subject areas, researcher were interested in identifying potential anomalies within this group which might help to explain the reasons for the results. Further analysis was applied to identify which Math, Science, and English courses resulted in the highest performance and which Math, Science, and English courses resulted in the lowest performance.

Researchers divided courses into three conditions: (a) high-engaged, high-performance, (b) high-engaged, low performance, and (c) low-engaged, low-performance based on student behaviors within the course. Courses categorized as high-engaged and high-performance might represent courses with both effective design and effective implementation because students were highly engaged and achieved expected outcomes. Those categorized as high-engaged and low-performance might represent courses with less effective course design because students were unable to achieve expected outcomes despite what appears to be effective implementation. Finally, courses categorized as low-engaged and low performance might represent courses with less effective course design and less effective course implementation. Our analysis revealed that regardless of the content area, most high-engaged, low performance, or low-engaged, low performance courses were entry-level courses. Most high-engaged, high performance courses were advanced level courses.

Students’ responses to the survey question asking students to indicate their reasons for taking an online course were then incorporated to help further interpret the results. The majority of responses from young students enrolled in courses categorized as high-engaged and high-performing were “the course was not available in my school.” The majority of older student responses in courses that were categorized as low-engaged and low performance were “I was making up a class I had failed.”

Findings

- 17) Regardless of Math, Science, or English subject-matter, entry level courses tended to have lower performance whether students were categorized as low-engaged or high-engaged. This may speak more to course structure, design, and support than to the effectiveness of instruction.
- 18) The reasons students enrolled in a course may influence their engagement level and performance. Student survey responses indicated that students who retook courses they have previously failed, tended to demonstrate lower engagement and lower performance. If students took courses which were not available in their schools, these students were usually high-engaged and high performing.

Predictive analysis

CRT Decision Tree analysis (Breiman, Friedman, Olshen, & Stone, 1984) was applied to construct predictive models combining course related data and survey results (Splitting Criterion: Gini; Leaf Size: 60; Maximum Depth: 10; Assessment Measure: Average Squared Error). These settings allow for a larger sequence of sub-trees in order to enrich the study's findings. Decision Trees classifies instances by sorting them down the tree from the root to the leaf nodes. In the tree structures, leaf nodes represent classifications, and branches represent conjunctions of features that lead to different target values. The following three variables were adopted as dependent variables in the Decision Tree analysis: 1) Average course grade; 2) average course satisfaction; and 3) average instructor satisfaction. Survey questions on course satisfaction and instructor satisfaction can be retrieved from <http://goo.gl/x8j18>

- Average course grade is each student's final course grade (range: 0-100). If a student took more than one course, average course grade is the average of multiple courses.
- Average course satisfaction was generated by averaging the scores from eight survey questions related to course content and the five survey questions related to course structure (range: 1–7). If a student took more than one course, average course satisfaction is the average of satisfaction scores from multiple courses.
- Average instructor satisfaction was generated by averaging the scores from 11 survey questions related to the instructor satisfaction (range: 1–7). If a student took more than one course, average instructor satisfaction is the average of satisfaction scores from multiple courses.

Final grade prediction

All variables in Table 1 were imported for final grade prediction. Average course grade was used as the dependent variable and the remainders were treated as independent variables. Because the tree results contained too much information, blank nodes were used to represent the results excluded from the data interpretation. Figure 3 shows the decision tree for final grade prediction. In academic year 2009-2010, 75.7% of students passed all courses, 15.9% of students failed all courses, and 8.4% passed some but not all of their courses. The left branch of the decision tree represents students who passed all courses. The results indicate a positive correlation between engagement level and performance (higher engaged => higher performance). The right branch of the decision tree represents students who had failed in one or more courses. The results imply a negative correlation between engagement level and performance (lower engaged => lower performance).

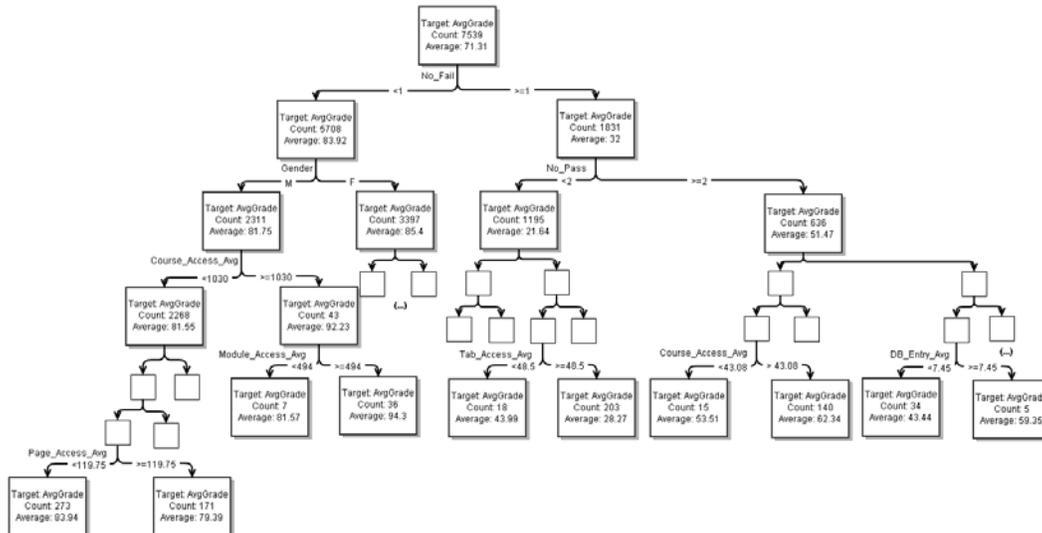


Figure 3. Final grade prediction (complete chart: <http://goo.gl/NIfWu>)

Findings

- 19) Engagement level and gender have stronger effects on student final grades than age, school district, school, and city. For most students, high engaged => high performance.

- 20) Compared with other Blackboard components such as discussion board entries and content access, tab access has negative effects on student performance (higher tab accessed => lower performance).
- 21) Female students performed better than male students.

Final grade prediction (external variables)

Additional decision tree analysis was conducted to investigate how external variables (i.e., non-learning activity variables) influenced student performance. Figure 4 is a portion of the decision tree for academic year 2009-2010.

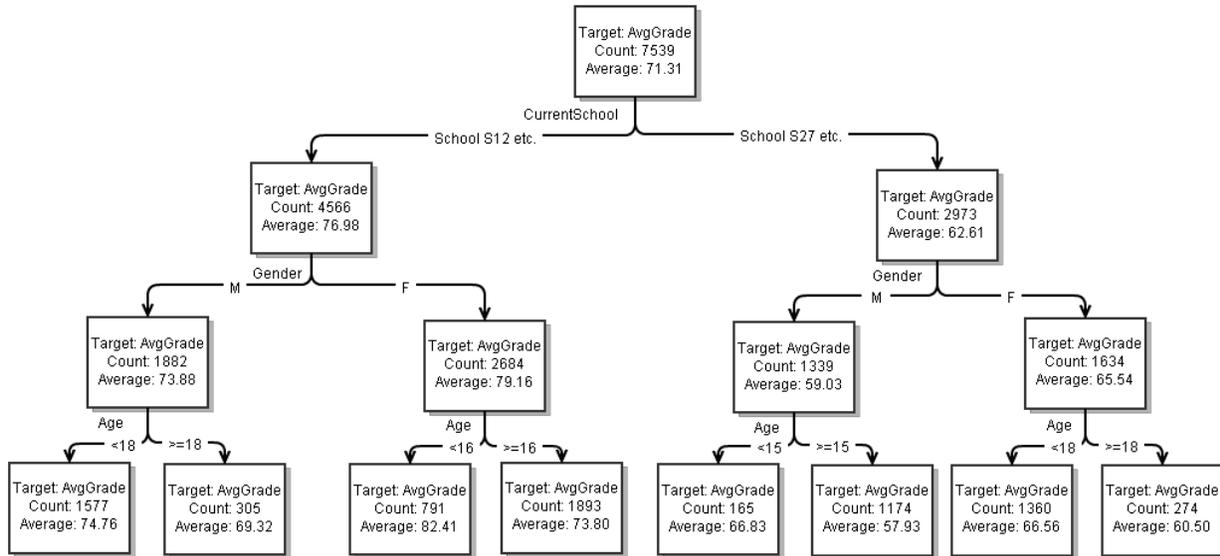


Figure 4. Final grade prediction with external variables only (complete chart: <http://goo.gl/B8AvB>)

Findings

- 22) Based on the predictive model, female students performed better than male students.
- 23) Students who were around 16 years old or younger performed better than those who were 18 years or older.

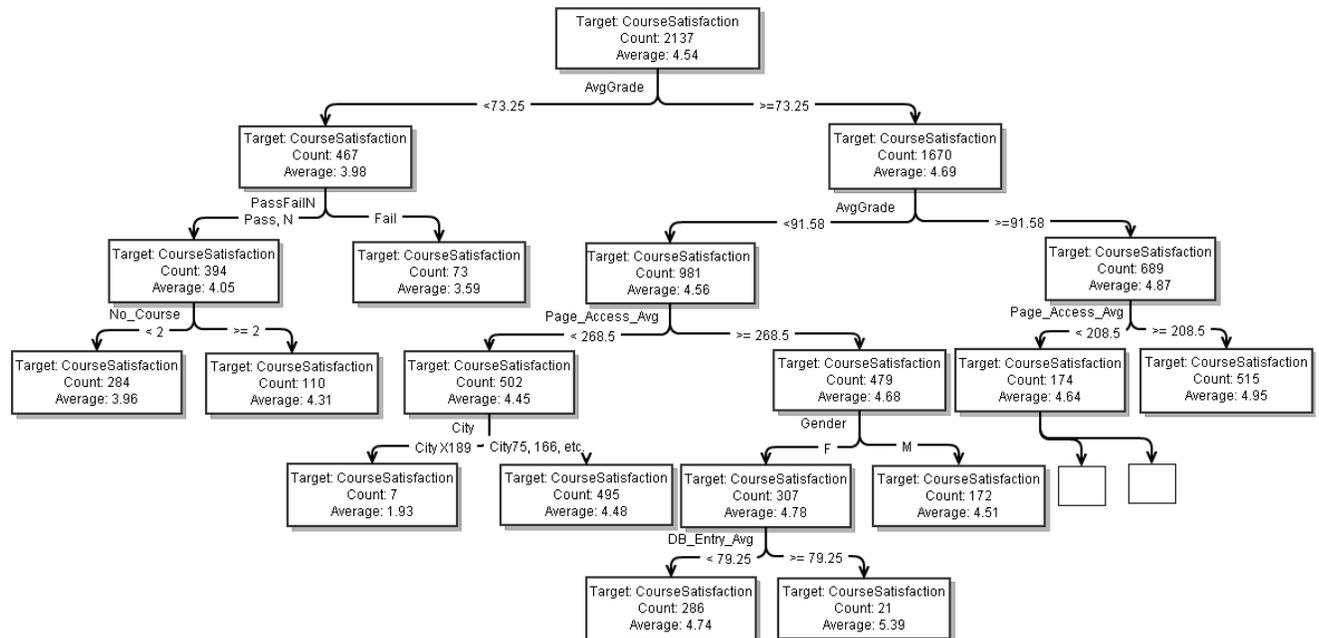


Figure 5. Course satisfaction prediction (complete chart: <http://goo.gl/5NLWI>)

Satisfaction prediction

Decision tree analysis was also conducted to predict students' satisfaction levels toward their course and instructor. Fall 2009 survey data could not be associated with variables in Blackboard, so the following results are limited to Spring 2010 only.

Course satisfaction

All the scores calculated from the responses to survey questions on course satisfaction were averaged into the scores of one course satisfaction variable. The value of "7" for this variable represents highest satisfaction with a course and "1" represents lowest satisfaction with a course. Figure 5 is a portion of the decision tree regarding course satisfaction.

Findings

- 24) Students with higher average final grades (> 73.25 , with a maximum score of 100) had higher course satisfaction.
- 25) Students who passed all courses or passed some of their courses had higher course satisfaction than all-failed students.
- 26) Students who took two or more courses in Spring 2010, whether they passed those courses or not, had higher course satisfaction.
- 27) Female students had higher course satisfaction than male students.
- 28) Online behaviors (i.e., frequency of page accessed and number of discussion board entries) had minor effects on course satisfaction (higher frequency/number \Rightarrow higher course satisfaction).
- 29) Students in different cities showed different course satisfaction levels.

Instructor satisfaction

All the scores calculated from the responses to survey questions on instructor satisfaction were averaged into the scores of one instructor satisfaction variable. The value of "7" for this variable represents highest satisfaction with an instructor and "1" represents lowest satisfaction with an instructor. Figure 6 is a portion of the decision tree regarding instructor satisfaction.

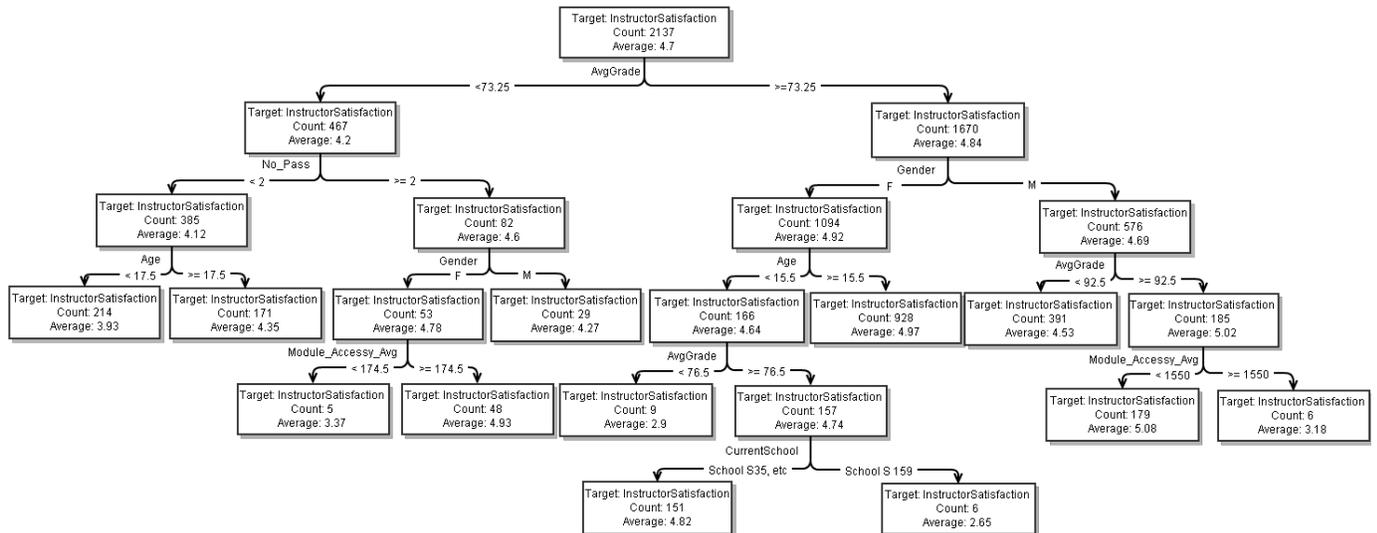


Figure 6. Instructor satisfaction prediction (complete chart: <http://goo.gl/QCdpw>)

Findings

- 30) Students with higher average final grades (> 73.25 , with a maximum score of 100) indicated higher instructor satisfaction.

- 31) Students who took two or more courses in Spring 2010, whether they passed those courses or not, showed higher instructor satisfaction.
- 32) Female students indicated higher instructor satisfaction than male students.
- 33) Online behaviors (frequency of module accessed) had minor effects on instructor satisfaction (higher frequency => higher course satisfaction). However, there were six students indicated low instructor satisfaction, despite extremely high frequency of course access and high final grades.
- 34) Older students taking one course (> 17.5 years old) had higher instructor satisfaction
- 35) Students from different schools showed different satisfaction levels for their online instructors..
- 36) Younger female students (<15.5 years old) with lower average final grade (<76.5) indicated lower instructor satisfaction.

Discussion

This study is a first attempt at program evaluation combining multiple data sources. The goal of this project was to propose a new program evaluation framework in order to generate sufficient information for program-level decision-making. The advantages of this framework are data triangulation and data interpretation. Below are triangulation and interpretation results:

- Female students generally performed better than male students (findings (9), (15), (21), and (22)); however, the findings were limited to the following subject areas: Electives, English, and Social Science (finding (15)).
- High-engaged students generally performed better than low-engaged students (findings (1), (10), (12) and (19)); however, the findings were limited to non-STEM courses (findings (10) and (11)). One possible factor influencing high-engaged students' inability to consistently achieve expected outcomes may have been poor course design.
- Younger students generally performed better than older students (findings (2) and (23)); however, the findings were limited to students in larger cities (findings (2)).
- One possible explanation for older students' generally lower performances may be that older students took more than two courses per semester for credit recovery (findings (5) and (18)). Younger students took fewer courses, but the fact the courses were generally not available in school district may have increased motivation (findings (2) and (18)).

Overall, using multiple forms of data allows for a more meaningful analysis of actual student behaviors, and the identification of potential relationships with demographic data, satisfaction data, and student outcomes. The result is a much richer and deeper analysis of student performance and teaching, as well as of effective course design, than could ever be accomplished with survey data or behavior mining alone.

Demographic and performance

Based on results revealed by the program evaluation framework, some indicators can be applied to identify students more likely to be successful and those more likely to be at-risk. In this study, a student who possessed more of the characteristics listed below was more likely to be successful:

- Female
- Younger than 16.5 years
- Took one or two courses per semester
- Took a Foreign Language or Health course
- Lived in a larger city

A student who possessed more of the characteristics below was more likely to be at risk of failure.

- Male
- Older than 18 years
- Took more than two courses per semester
- Took entry-level courses in Math, Science, or English
- Lived in a smaller city

These indicators can be applied to develop an early warning system (Macfadyen & Dawson, 2010), so administrators and teachers can have a list of successful and at-risk students before each semester starts.

Engagement and performance

Based on data mining analysis, higher-engaged students usually had higher performance. This finding is also supported by previous studies (e.g., Hung & Crooks, 2008; Hung & Zhang, 2009). However, the conclusion may be limited to courses which were well designed and implemented. In this study, entry-level courses tended to have lower performance, regardless as to whether students were categorized as low-engaged or high-engaged. This means high-engaged students might still have lower final grades if they were in a course with course structure, design, and/or support issues.

Lim & Morris (2009), when studying post-secondary students, found junior and senior students had significantly higher survey mean scores in perceived learning, learning application, and learning involvement than freshman and sophomore students. Assuming higher perceived learning, learning application, and learning involvement equates to high motivation, the authors could not explain why older students had significantly higher engagement than young students. Our study, by combining analysis of engagement and performance through data mining with survey responses, revealed why students had different levels of engagement and performance. For example, the majority of responses from the students enrolled in courses categorized as high engagement and high performance were, “The course was not available in my school.” Meanwhile, the majority of responses from students in courses that were categorized as low engagement and low performance were, “I was making up a class I had failed.” The level of engagement in our study may have been influenced by motivation. In addition, Lim & Morris (2009) found older students had a better chance to be successful in online learning at the higher education level. However, our study found older students were more likely to be at-risk students in K-12 online education. Students older than 18 tended to be low engaged and lower performing in their courses.

Satisfaction and performance

There is no confirmed relationship between student performance and satisfaction (positive correlation—Eiszler 2002, Nasser & Hagtvet 2006; No relation—Ladebo, 2003, Walker & Palmer, 2011). In the case study, students with higher final grades usually had higher course and instructor satisfaction. In addition, survey results showed female students reported significantly higher satisfaction with courses and instructors. Similar results were also found in Hermans, Haytko, & Mott-Stenerson’s study (Hermans, Haytko, & Mott-Stenerson, 2009).

Survey results showed that Health courses required the least effort and were considered the least challenging subject. Students also had high satisfaction with Health courses. However, although students had high satisfaction and engagement levels in Science courses, their average final grades were the third lowest. These findings showed high satisfaction and engagement levels could not guarantee high performance. While students obtained the lowest average grades in Math and English courses, they did not show significantly lower satisfaction levels in these two subject areas. Assuming perception data only reflected positive experiences, the picture of students’ experiences in courses could be misrepresented if students’ learning behaviors were not analyzed. These findings illustrate the flaws in solely relying on self-reporting and perception data in program evaluation to inform strategic decisions.

Limitations

It should be noted that this project was not without challenges. We are in the early stages of data mining, and learning management system providers are often not cognizant of the importance of their data collection strategies. For example, the Blackboard activity accumulator grouped wide-ranging learning behaviors into only five categories, which significantly simplified findings that could potentially be revealed by data mining. In addition, issues with missing data, data stored in the wrong fields, or mismatched data fields in the database limited some of the analysis techniques attempted in this evaluation. Blackboard failed to track ever forum reply behavior, and in order to avoid using inaccurate data to generate false results, we discarded students-student and student-instructor interaction analysis. Although multiple data sources provide rich information for data interpretation, some unexpected results

were revealed by survey design and investigation. For example, why did students from different schools show different satisfaction levels for their instructors? Further data interpretation could be undertaken with additional open-ended questions.

Proof of the framework

This study demonstrates the benefits of incorporating data mining into the program evaluation of K-12 online education. It is important to note that demographic data and course evaluation survey data are also indispensable in supporting data-mining-results evaluation and interpretation. In exploring EDM applications at the K-12 level that have already been broadly adopted in higher education institutions we have demonstrated: 1) how data mining can be incorporated into program evaluation in order to support decision-making at the institutional level, 2) how a framework of data triangulation generates high-quality and non-partial results that can be combined with student learning logs, demographic data and course evaluation surveys and that 3) characteristics of successful and at-risk students can be generated by identifying important predictors of student performance, course satisfaction, and instructor satisfaction for K-12 online education.

Future research

Future study should continue in the following directions: 1) Evaluate the usefulness of the information by surveying administrators; and 2) Generalize findings by incorporating more data from various program evaluation projects. For example, many average numbers in this study could be important indicators for retention and performance improvement. The numbers would be more stable for generalization with additional study data; 3) Develop an early warning system using tables and figures in this study to simplify administrators' decision-making processes; 4) Add further possible open-ended questions for data interpretation; 5) Conduct a follow up study. For example, entry-level Science courses were identified as classes that might have course design issues. Follow up research might help identify which parts are difficult for students; and 6) Utilize the framework to validate or generate educational models.

References

- Astin, A. W. & Lee, J. J. (2003). How risky are one-shot cross-sectional assessments of undergraduate students? *Research in Higher Education*, 44(6), 657–672.
- Becker, K., Ghedini, C., & Terra, E. (2000). Using KDD to analyze the impact of curriculum revisions in a Brazilian university. In B. V. Dasarathy (Ed.), *Proceedings of Data Mining and Knowledge Discovery: Theory, Tools, and Technology II*, 4057, 412–419.
- Blackboard Inc. (2010). Blackboard database schema. Retrieved January 5, 2011, from <http://library.blackboard.com/ref/fea8bb86-1720-4b14-bb81-542d973b5338/>
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadworth International Group: Chapman and Hall.
- Budayan, C., Dikmen, I., & Birgonul, M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. *Expert Systems with Applications*, 36(9), 11772-11781.
- Cheng, D. X. (2001). Assessing student collegiate experience: where do we begin? *Assessment and Evaluation in Higher Education*, 26(6), 525–38.
- Delavari, N., Phon-amnuaisuk, S., & Beikzadeh, M. (2008). Data mining application in higher learning institutions. *Informatics in Education*, 7(1), 31–54.
- Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, 43(4), 483–501.
- Elster, J. (1983). *Sour grapes: Studies in the subversion of rationality*. Cambridge, UK: Cambridge University Press.
- Grammatikopoulous, V. (2012). Integrating program theory and systems-based procedures in program evaluation: A dynamic approach to evaluate educational programs. *Educational Research and Evaluation*, 18(1), 53–64.

- Hartigan, J.A., & Wong, M.A. (1979). Algorithm AS136: A k-means clustering algorithm. *Applied Statistics*, 28, 100–108.
- Hermans, C. M., Haytko, D. L., & Mott-Stenerson, B. (2009). Student satisfaction in web-enhanced learning environments. *Journal of Instructional Pedagogies*, 1(1), 1–19.
- Hoffman, K. M. (2003). Online course evaluation and reporting in higher education. *New Directions for Teaching and Learning*, 96, 25–29.
- Hung, J. L., & Crooks, S. (2009). Examining online learning patterns with Data Mining techniques in peer-moderated and teacher-moderated course. *Journal of Educational Computing Research*, 40(2), 183–210.
- Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*, 4(4), 426–437.
- Jick, T. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24, 602–611.
- Ladebo, O. J. (2003). Relationship between Agricultural trainees' performance and satisfaction with academic program. *Journal of International Agricultural and Extension Education*, 11(1), 55–60.
- Lim, D. H., & Morris, M. L. (2009). Learner and instructional factors influencing learning outcomes within a blended learning environment. *Educational Technology & Society*, 12(4), 282–293.
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an early warning system for educators: A proof of concept. *Computers and Education*, 54, 588–599.
- Nasser, F., & Hagtvet, K. A. (2006). Multilevel analysis of the effects of student and instructor course characteristics on student ratings. *Research in Higher Education*, 47(5), 559–90.
- Nichols, J. O. & Nichols, K. W. (2000). *The departmental guide and record book for student outcomes assessment and institutional effectiveness*, 3rd edition. New York: Agathon Press.
- Panov, P., Soldatova, L., & Dzeroski, S. (2009). Towards an ontology of data mining investigations. *Lecture Notes in Artificial Intelligence*, 5808, 257–271.
- Picciano, A. G. (2002). Beyond student perceptions: Issues of interaction, presence, and performance in an online course. *Journal of Asynchronous Learning Networks*, 6(1), pp. 21–40.
- Roiger, R. J., & Geatz, M. W. (2003). *Data mining: a tutorial-based primer*. Boston, MA: Addison Wesley.
- Romero, C., & Ventura, S. (2010). Educational data mining : A review of the state of the art. *IEEE Transactions on Systems Man and Cybernetics Part C: Applications and Reviews*, 40(6), 601–618.
- Spiriduso, W. & Reeve, T. G. (2011). The national academy of Kinesiology 2010 review and evaluation of doctoral programs in Kinesiology. *Quest*, 63(4), 411–440.
- Ueno, M. (2006). Online outlier detection of learners' irregular learning processes. In C. Romero & S. Ventura, (Eds.), *Data Mining in E-learning* (pp. 261–278). Billerica, MA: WitPress.
- Vogt, F. D. & Slis, D. (2011). A model for programmatic evaluation by student assessment. *Journal of Science Education and Technology*, 20, 796–802.
- Walker, D.J. & Palmer, E. (2011). Observing the relationship between student satisfaction and performance in an Australian engineering program. *Assessment & Evaluation in Higher Education*, 36(2), 157-170.
- Wu, K. W., & Leung, C. H. (2002). Evaluating learning behavior of web-based training (WBT) using web log. In B. Werner (Ed), *Proceedings of International Conference on Computers in Education* (pp. 736–737). New Zealand: IEEE Computer Society.