Identifying Relevant Interaction Metrics for Predicting Student Performance in a Generic Learning Content Management System

by

Eric Beerman

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2015 by the
Graduate Supervisory Committee:

Kurt VanLehn, Chair
Ian Gould
Ihan Hsiao

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

The growing use of Learning Management Systems (LMS) in classrooms has enabled a great amount of data to be collected about the study behavior of students. Previously, research has been conducted to interpret the collected LMS usage data in order to find the most effective study habits for students. Professors can then use the interpretations to predict which students will perform well and which student will perform poorly in the rest of the course, allowing the professor to better provide assistance to students in need. However, these research attempts have largely analyzed metrics that are specific to certain graphical interfaces, ways of answering questions, or specific pages on an LMS. As a result, the analysis is only relevant to classrooms that use the specific LMS being analyzed.

For this thesis, behavior metrics obtained by the Organic Practice Environment (OPE) LMS at Arizona State University were compared to student performance in Dr. Ian Gould's Organic Chemistry I course. Each metric gathered was generic enough to be potentially used by any LMS, allowing the results to be relevant to a larger amount of classrooms. By using a combination of bivariate correlation analysis, group mean comparisons, linear regression model generation, and outlier analysis, the metrics that correlate best to exam performance were identified. The results indicate that the total usage of the LMS, amount of cramming done before exams, correctness of the responses submitted, and duration of the responses submitted all demonstrate a strong correlation with exam scores.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

1.0 INTRODUCTION

As an increasing amount of education is now being done online, there have been many open-sourced platforms created to manage the content and structure of an online class ("Babson Study"). Called Learning Management Systems (LMS), these platforms typically provide teachers the ability to create and assign students to classes, create practice problem sets, post ideas in discussion boards, view student progress, and potentially more, though features tend to vary from one to the next. Popular LMS examples include Moodle, Canvas, eFront, Sakai, Blackboard, and ATutor, to name a few.

One of the most important features of an LMS is the ability to view student progress, so that the teacher knows which students have been completing the digital assignments. However, while this information may provide an accurate portrayal of how much time the student has been working on the problem sets, it does not necessarily indicate whether the student will end up being successful in the class. Being able to predict success in the classroom can be extremely useful to teachers, so that they know which students may need more help, instruction, or assistance than others in order to maximize their learning. Furthermore, knowing which study behaviors are more effective when learning in an online course can help the teacher better structure the class, as well as provide effective study advice to the students. Such intervention practices based on performance prediction has already shown to be effective with the Course Signals project at Purdue (Arnold and Pistilli).

To that end, there have been several research projects that have attempted to identify which online course usage metrics correlate with successful results in the classroom. Edwards et al. found that the earlier students began working on problems on the website, the better students performed in the class. Feng, Mingyu, et al. created a student model that predicted standardized math test scores of 392 students who used an intelligent tutoring system called ASSISTment. This study looked more at how the students worked on the problems themselves, such as average number of hints received, time spent per question, number of questions answered, etc.

Additionally, studies have been done on metrics recorded while using specific LMS platforms. Filippidi, Tselios, et al. analyzed usage metrics of 117 students using a Moodle-based website for a one semester technology in education course. In their analysis, they looked at the amount of time students viewed various pages and the site overall, and compared it with the student performance for the course to determine which features correlated most with higher performance. Haig, Falkner, et al. created a plugin for Moodle to attempt to graph such usage data real time in the hopes of identifying students who are at risk of not succeeding in the course. Using a different platform, Macfadyen and Dawson studied metrics obtained from 118 students using Blackboard for an undergraduate biology course. They analyzed various Blackboard pages and student usage of them in relation to how the students performed in the class.

There are issues with the current research in this field, however. First, multiple studies, such as the work mentioned in the previous paragraph on Moodle and Blackboard, focus solely on the system resources the students are using and how that is associated with performance. Such data is not useful outside of the platform upon which

the research is conducted, nor is it useful in platforms that simply do not have a wide variety of resources available (such as platforms which are heavily practice problem based and do not emphasize social interaction). Additionally, of the studies that do analyze problem-based metrics, such as ASSISTment, metrics are often included that are specific to the way questions are presented on that platform. ASSISTment, for example, analyzes the effectiveness of the amount of hints students receive when working out questions, which may or may not be a feature of other LMS systems. Since there are a near infinite amount of ways to potentially present and interact with questions in an LMS, and a near infinite amount of features and pages that can be added to one, this experiment instead looks at metrics that are independent of those factors and thus can be applied to a wider range of platforms.

To conduct this research, an LMS was developed at Arizona State University called Organic Practice Environment (OPE). It has been used in seven organic chemistry classes at ASU and Centre College. The platform is simple and problem-oriented – it has no discussion boards or social features and has merely 3 different pages. However, it provides a high volume of organic chemistry practice problems for the students to work on. Over 2,000 students have used or are currently using the platform, and have submitted roughly 1,000,000 question responses with it.

For this experiment, the vast student metric data collected from the OPE platform was analyzed to determine which student behaviors when using the platform correspond best with performance in the classroom. In doing so, there are three main objectives.

- Find which metrics correlate best with student performance when used individually and not in combination with other metrics. This would allow

professors to quickly look at usage statistics and make an educated guess for how well students will perform.

- Find which *combination* of metrics can be used to create a prediction model to accurately predict student test results, giving the professor the potential to create an accurate forecast for exactly how well the students will perform in the rest of the class.
- Find any outliers from common trends, and determine any patterns in these outliers, to help a professor understand when a prediction may not be valid.

Although not all of the metrics analyzed in this experiment may apply to all platforms, none of them are specific to the OPE platform. Therefore, as a result of this research, any instructor or researcher can know which metrics can be used to predict student performance in their class, regardless of which platform they use.

2.0 EXPERIMENTAL SETUP

In the Fall 2014 semester, 519 students in Dr. Ian Gould's Organic Chemistry I class at ASU used the OPE platform to practice organic chemistry concepts and work on their homework. In the platform, there were a total of 988 questions split among 56 categories. Of those 988 questions, 560 counted as credit towards their class grade, and the other 428 questions were simply for more practice for the student. In this paper, all questions that count towards the student's grade are referred to as "credit questions", while all other questions are referred to as "non-credit questions".

The OPE platform allows for multiple types of questions to be presented to the student. Of those types, some of the more commonly used types include:

- Self-reporting, where the student declares if they got the question correct or incorrect after viewing the answer.
- Multiple-choice, where the student is given multiple options to choose from and the system grades their answer.
- Input, where the student must type an answer and the system grades it.

Of the 988 questions presented to the students, 64 were input questions, 678 were self-reporting, 32 were multiple-choice, and the remaining 214 allowed the student to choose between multiple-choice, input or self-reporting. Regardless of type, all questions allow the students to attempt them as many times as they would like. Students can visually see which questions they have previously worked on, and whether they have gotten it correct or not already. Additionally, all questions have an explanation that is shown to the student upon answering it correctly, or upon request for input questions.

For this course, Dr. Gould provided 3 midterm exams evenly spread throughout the semester, along with a cumulative final exam at the end of the semester. The dates for each exam can be seen in table 1**.** Each midterm throughout the semester tested roughly one month's worth of new material, and the final tested both new and old material from the entire semester. Student grades were based on a combination of performance on exams, performance on weekly online class quizzes, and amount of online homework credit questions attempted. For the sake of this experiment, only the performance on exams was counted towards the student performance for the class.

3.0 METHODS

All data in the OPE platform were stored in a MySQL database, of which the schema can be found in Appendix A. A global database keeps track of all the data shared amongst all classes, and each specific class in the system has its own database that keeps track of the questions and categories within each class, as well as the student responses to the questions. When a response is submitted, it is graded for correctness, and then stored along with various details about the response. Such details include the time it took for the student to work on the question, the time spent viewing the explanation, the timestamp for its submission, the type of response, the type of question, and whether it was correct.

To analyze the behavior metrics collected by the system, the desired metrics to be analyzed were first defined, and then a query was created to extract the desired data from the database. The query used can be found in Appendix B. The following is the complete list of metrics collected:

- **Total Responses –** The total responses submitted by the student, including correct and incorrect responses. In describing these metrics, a "response" is a single attempt to answer a question, whereas a "question attempted" refers to a question that the student has submitted at least one response to. As an example, if a student submits an answer five times to one question, he will have five responses and only one question attempted. Thus, although the OPE system has 988 questions, "Total Responses" could be larger because students may attempt the same question multiple times. "Total Responses" was also broken down into subcategories for credit and non-credit questions.

- **Correct Responses –** The amount of responses submitted by the student that the system graded as correct or that the student self-reported as correct. This was broken down into subcategories for credit and non-credit questions as well. Please note that a "correct" response in this paper simply indicates that the answer provided by the student is equivalent to the answer stored in the database, and does not factor in how the question is presented or interacted with.
- **Percent Correct Responses** – The percent of responses submitted that were correct. This can be thought of as the response success rate.
- **Total Responses Self-Reporting Questions** – The amount of responses to questions that had self-reporting input.
- **Percent Correct Responses Questions** – The percent of the time that the student marked his answer as correct when reporting the correctness of his result.
- **Total Responses Non-Self-Reporting Questions** – The amount of responses to questions that graded the student's response.
- **Percent Correct Responses Non-Self-Reporting Questions** – The percent of the time that students got questions correct when the system (rather than the student) determined the correctness of the student's response.
- **Correct Response Ratio** – The ratio between how frequently the students got questions correct for self-reporting questions versus all other questions. Essentially this is an attempt to measure how honest the student is when self-reporting correctness.
- **Questions Attempted** – Of the 988 questions in the OPE system, this metrics counts the number that the student attempted at least once, regardless of whether

any of the attempts was answered correctly or incorrectly. This was broken down into credit and non-credit questions as well.

- **Percent Questions Attempted** – The percent of question attempted compared to the number of questions in the system (988).
- **Cramming Percentage –** The percent of the questions attempted that did not occur the weekend before the exam, compared to the questions attempted during the entire exam period. Essentially, the higher this number, the less the student crammed for the exam.
- **Average Responses Per Question** – The average number of attempts per question.
- **Questions Completed** – The number of questions that the student answered correctly at least once.
- **Questions Incomplete** – The number of questions that the student attempted at least once but never correctly solved.
- **Average Duration** – The average amount of time the student took to submit each response. The OPE system starts recording this time when the page loads, and stops recording when the student submits the response. Please note that the highest 1% of these values were discarded for being too high above the average, and were thus likely caused by a student leaving their browser window open despite not actually working on the problem. This was also broken down into subcategories of duration for completed and incomplete questions.
- **Average Explanation Duration** – The average amount of time spent viewing the explanation for the question. Each question in the OPE system has an explanation

tied to it which can display if the student requests it (for input questions) or answers the question correctly. This involves the time spent both reading and comprehending the content in the explanation.

- **Study Density Score** – This metric attempts to identify how well the student spread out his studying over the course of the semester. Each exam period throughout the semester was given a weight proportional to the amount of points the exam was worth. The percent of questions attempted during each period was then compared to this weight to see if the student put a desired amount of emphasis on it when studying. The larger the value of this metric, the more the student concentrated his question attempts on specific time periods than spreading them throughout the semester. The formula is as follows:

$$\left(\sum_{x}\left|\frac{P_x}{2P_t}-\frac{x_1}{T}\right|+\left|\frac{P_x}{2P_t}-\frac{x_2}{T}\right|\right)^2$$

  In the formula, x refers to the exam (midterms 1-3 and the final), P_x refers to the total possible points for that midterm, P_t refers to the total possible points for the class as a whole, x_1 refers to the questions attempted in the weeks leading up to the exam, x_2 refers to the questions attempted in the weekend before the exam, and T refers to the student's total questions attempted throughout the semester.

- **Average Questions Per Day** – Measures the average amount of questions that the student attempted per day, for days in which the student attempted at least one question. As an example, if throughout the course of the semester, the student used the site a total of 5 days to attempt at total of 10 questions, he would

have averaged 2 questions per day. Please note that this measures questions attempted, *not* total amount of responses for those questions.

- **Questions Attempted to Daily Average Ratio** – Measures the ratio of the questions attempted compared to the Average Questions Per Day. Essentially measures how spread out a student's studying was for a given period of time.

For this experiment, student exam scores are used to measure performance in the classroom. In order to analyze the effect of various behaviors on scores for each exam, this experiment split the semester into several different periods of time, as defined in table 1. Essentially, the usage metrics for each exam were looked at for the weeks leading up to the exam since the prior exam, for the weekend immediately before the exam, for the total time since the last exam, and for the total time since the start of the semester. The aforementioned query was then run against each of these time periods, and the data was stored in a Microsoft Excel spreadsheet.

Once in the spreadsheet, the data were loaded into a statistical analysis program called IBM SPSS. For each objective of the experiment, various different statistical analysis techniques were performed on the data set to extract the desired information. These techniques are described in more detail in the sections that follow.

Table 1

Time Periods Queried Against for the Semester

Period	Start Date	End Date
Midterm 1 Period 1	August 23, 2014	September 18, 2014
Midterm 1 Period 2	September 19, 2014	September 22, 2014
Midterm 1	August 23, 2014	September 22, 2014 (morning of exam)
Midterm 2 Period 1	September 23, 2014	October 16, 2014
Midterm 2 Period 2	October 17, 2014	October 20, 2014
Midterm 2	September 23, 2014	October 20, 2014 (morning of exam)
Cumulative Midterm 2	August 23, 2014	October 20, 2014
Midterm 3 Period 1	October 21, 2014	November 13, 2014
Midterm 3 Period 2	November 14, 2014	November 17, 2014
Midterm 3	October 21, 2014	November 17, 2014 (morning of exam)
Cumulative Midterm 3	August 23, 2014	November 17, 2014
Final Period 1	November 18, 2014	December 4, 2014
Final Period 2	December 5, 2014	December 9, 2014
Final	November 18, 2014	December 9, 2014 (morning of exam)
Total Semester	August 23, 2014	December 9, 2014

4.0 METRICS ANALYSIS RESULTS

The first objective of the experiment is to find which individual metrics contribute the most to student performance in the classroom. To accomplish this, each exam period (except for the first) was analyzed in two ways. First, exam scores for each exam were compared against the metrics collected since the previous exam, referred to as the "period" metrics. Next, exam scores for each exam were compared against the metrics collected cumulatively throughout the semester up until that exam, referred to as the "cumulative" metrics. As an example, "Midterm 3 Period" statistics would refer to how metrics collected from the day Midterm 2 was taken to the day Midterm 3 was taken affected the Midterm 3 exam scores, while "Midterm 3 Cumulative" would refer to how metrics collected from the start of the semester to the day Midterm 3 was taken affected the Midterm 3 exam scores. This concept is shown in figure 1.

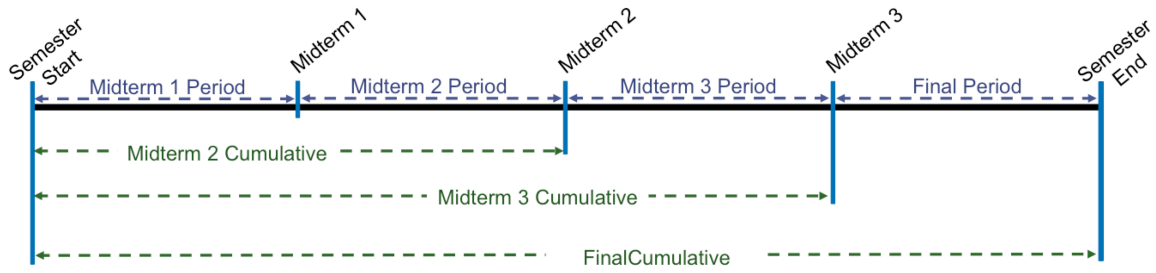

Figure 1

Definition of Collection Periods for Metrics Analysis

There were two statistical analysis techniques used to do this. First, a simple bivariate correlation was performed on various metrics against each exam, to discover how much of a correlation existed between the metric and the exam scores. In SPSS, this was accomplished with the CORRELATIONS function, using a Pearson correlation

coefficient and a two-tailed test of significance. The results of this analysis produce a number from the range of -1.0 (strong negative correlation) to 1.0 (strong positive correlation). In subsequent tables, the columns labeled “Bivariate” report these Pearson correlation coefficients.

Although a bivariate correlation can declare if a correlation exists between a metric and an exam score, it does not give a qualitative sense of how large an effect that metric had on the exam score. To find this, the class was first split into two groups around the median of the metric using the RANK function in SPSS. Then, the mean exam score of each group was calculated, factoring in an ANOVA test to determine the significance of the difference. Henceforth, the group of students whose metric values were above the median shall be called the “High Group” and the group of students whose metrics values were below the median shall be called the “Low Group”. As a point of reference, table 2 shows the maximum possible difference between the High Group and Low Group for each exam, based on how the class performed.

Table 2

Maximum Difference Between High Group and Low Group per Exam

Exam	**High Group**	**Low Group**	**Standard Deviation**
Midterm 1	164	106	38
Midterm 2	160	105	35
Midterm 3	154	89	39
Final	317	203	73
Total	787	537	158

In tables 3-7, the “Bivariate” column represents the bivariate correlation between the metric and the exam scores for that period, the “High Group” column represents the mean exam score for the group of students above the median for that metric, and the “Low Group” column represents the mean exam score for group of students below the median for that metric. Please note that an “N/A” means that the statistic was not applied to the period in question. An “N/S” (Not Significant) in the “High Group” or “Low Group” column, meanwhile, means that the difference between the High Group and Low Group, according to the ANOVA results, had a p-value greater than 0.05 and thus was deemed statistically insignificant. An “N/S” in the “Bivariate” column similarly means that the p-value of the bivariate correlation is greater than 0.05. All other values displayed in the tables had p-values less than 0.05 for their respective tests.

Table 3

Midterm 1 Metric Analysis

Factor	Midterm 1 Period		
	Bivariate	High Group	Low Group
Total Responses	0.398	148	122
Total Responses Non-Credit Questions	0.335	147	124
Percent Correct Responses	0.498	154	117
Total Responses Self-Reporting	0.520	153	118
Percent Correct Responses Self-Reporting	0.450	142	129
Percent Correct Responses Non-Self-Reporting	0.437	150	121
Correct Response Ratio	0.259	N/S	N/S
Questions Attempted	0.561	151	119
Credit Questions Attempted	0.593	153	117
Non-Credit Questions Attempted	0.411	148	123
Cramming Percentage	0.194	141	105
Average Responses Per Question	N/S	126	144
Questions Completed	0.559	152	118
Non-Credit Questions Completed	0.410	148	123
Questions Incomplete	0.089	140	133
Average Explanation Duration	N/S	N/S	N/S
Average Duration	0.351	147	128
Average Duration Correct Responses	0.362	148	128
Average Duration Incorrect Responses	0.224	145	131
Average Questions Per Day	0.191	143	128
Study Density Score	N/A	N/A	N/A
Questions Attempted to Daily Average Ratio	0.440	153	123

Table 4

Midterm 2 Metric Analysis

Factor	Midterm 2 Period			Midterm 2 Cumulative		
	Bivariate	High Group	Low Group	Bivariate	High Group	Low Group
Total Responses	0.407	146	119	0.423	145	120
Total Responses Non-Credit Questions	0.285	141	124	0.317	142	123
Percent Correct Responses	0.342	143	122	0.453	149	116
Total Responses Self-Reporting	0.483	148	117	0.529	148	117
Percent Correct Responses Self-Reporting	0.348	137	127	0.286	136	128
Percent Correct Responses Non-Self-Reporting	0.287	140	124	0.373	144	120
Correct Response Ratio	0.130	136	128	N/S	127	138
Questions Attempted	0.449	147	118	0.535	148	117
Credit Questions Attempted	0.484	148	117	0.573	150	114
Non-Credit Questions Attempted	0.307	142	123	0.374	143	122
Cramming Percentage	0.142	139	120	N/A	N/A	N/A
Average Responses Per Question	0.171	N/S	N/S	-.148	124	141
Questions Completed	0.445	147	118	0.534	148	117
Non-Credit Questions Completed	0.305	142	123	0.375	143	122
Questions Incomplete	N/S	N/S	N/S	N/S	N/S	N/S
Average Explanation Duration	N/S	N/S	N/S	N/S	N/S	N/S
Average Duration	0.280	143	126	0.369	145	120
Average Duration Correct Responses	0.294	142	127	0.382	145	121
Average Duration Incorrect Responses	0.161	140	129	0.242	143	122
Average Questions Per Day	0.144	139	126	N/S	N/S	N/S
Study Density Score	N/A	N/A	N/A	N/A	N/A	N/A
Questions Attempted to Daily Average Ratio	0.415	147	124	0.460	148	119

Table 5

Midterm 3 Metric Analysis

Factor	Midterm 3 Period			Midterm 3 Cumulative		
	Bivariate	High Group	Low Group	Bivariate	High Group	Low Group
Total Responses	0.408	138	104	0.458	136	105
Total Responses Non-Credit Questions	0.284	133	109	0.334	133	109
Percent Correct Responses	0.298	126	116	0.438	137	105
Total Responses Self-Reporting	0.419	137	104	0.513	138	103
Percent Correct Responses Self-Reporting	0.322	124	117	0.277	124	111
Percent Correct Responses Non-Self-Reporting	0.257	129	112	0.338	132	109
Correct Response Ratio	0.194	126	116	N/S	117	125
Questions Attempted	0.442	138	103	0.540	139	102
Credit Questions Attempted	0.448	138	103	0.575	139	103
Non-Credit Questions Attempted	0.296	132	109	0.367	134	108
Cramming Percentage	0.146	129	107	N/A	N/A	N/A
Average Responses Per Question	0.160	126	116	-.164	112	130
Questions Completed	0.434	138	104	0.537	139	103
Non-Credit Questions Completed	0.292	132	109	0.367	134	108
Questions Incomplete	0.180	131	112	N/S	N/S	N/S
Average Explanation Duration	N/S	N/S	N/S	N/S	N/S	N/S
Average Duration	0.226	131	110	0.303	133	109
Average Duration Correct Responses	0.254	131	110	0.319	133	109
Average Duration Incorrect Responses	0.128	131	111	0.174	129	113
Average Questions Per Day	0.091	128	113	N/S	N/S	N/S
Study Density Score	N/A	N/A	N/A	N/A	N/A	N/A
Questions Attempted to Daily Average Ratio	0.402	138	110	0.473	137	105

Table 6

Final Metric Analysis

Factor	Final Period			Final Cumulative		
	Bivariate	High Group	Low Group	Bivariate	High Group	Low Group
Total Responses	N/S	269	250	0.369	281	238
Total Responses Non-Credit Questions	0.117	269	249	0.292	279	240
Percent Correct Responses	0.314	279	240	0.438	287	231
Total Responses Self-Reporting	N/S	269	250	0.425	281	237
Percent Correct Responses Self-Reporting	0.252	N/S	N/S	0.211	N/S	N/S
Percent Correct Responses Non-Self-Reporting	0.396	288	231	0.435	284	234
Correct Response Ratio	N/S	247	272	0.134	274	245
Questions Attempted	N/S	269	250	0.471	284	235
Credit Questions Attempted	N/S	N/S	N/S	0.474	284	235
Non-Credit Questions Attempted	0.139	274	244	0.331	280	238
Cramming Percentage	0.142	275	241	N/A	N/A	N/A
Average Responses Per Question	N/S	N/S	N/S	-.148	247	272
Questions Completed	N/S	269	250	0.469	284	234
Non-Credit Questions Completed	0.135	275	243	0.330	280	239
Questions Incomplete	N/S	266	251	N/S	N/S	N/S
Average Explanation Duration	N/S	N/S	N/S	N/S	N/S	N/S
Average Duration	0.233	278	249	0.312	280	239
Average Duration Correct Responses	0.237	279	249	0.331	280	240
Average Duration Incorrect Responses	0.183	274	253	0.171	275	244
Average Questions Per Day	-.172	N/S	N/S	-.161	N/S	N/S
Study Density Score	N/A	N/A	N/A	N/A	N/A	N/A
Questions Attempted to Daily Average Ratio	0.294	282	246	0.475	289	230

Table 7

Total Semester Metric Analysis

Factor	Bivariate	Total High Group	Low Group
Total Responses	0.361	705	617
Total Responses Non-Credit Questions	0.300	705	617
Percent Correct Responses	0.493	730	592
Total Responses Self-Reporting	0.432	708	614
Percent Correct Responses Self-Reporting	0.224	N/S	N/S
Percent Correct Responses Non-Self-Reporting	0.487	724	598
Correct Response Ratio	-.164	623	700
Questions Attempted	0.477	713	610
Credit Questions Attempted	0.471	712	611
Non-Credit Questions Attempted	0.346	709	613
Cramming Percentage	0.100	670	588
Average Responses Per Question	-.179	628	695
Questions Completed	0.476	716	607
Non-Credit Questions Completed	0.345	709	613
Questions Incomplete	N/S	N/S	N/S
Average Explanation Duration	N/S	N/S	N/S
Average Duration	0.349	709	614
Average Duration Correct Responses	0.364	709	614
Average Duration Incorrect Responses	0.206	699	625
Average Questions Per Day	-.173	649	670
Study Density Score	-.461	608	715
Questions Attempted to Daily Average Ratio	.484	728	595

For this analysis, the collected metrics were divided into similar groups. Each group contains its own section below to analyze and interpret the results for its metrics.

4.1 TOTAL USAGE METRICS

Table 8

List of Total Usage Metrics

Total Responses
Total Responses Non-Credit Questions
Total Responses Self-Reporting
Questions Attempted
Credit Questions Attempted
Non-Credit Questions Attempted
Questions Completed
Non-Credit Questions Completed
Average Responses Per Question

Before running the experiment, it was hypothesized that the more a student used the class website built on the OPE platform, the better the student would do on the exams. For this hypothesis, the metrics involving number of responses, questions attempted, and questions completed were analyzed.

With the exception of the Final Period, every exam period shown in tables 3-7 displayed a positive correlation between exam score and questions attempted, questions completed, and total responses. Thus, each High Group scored higher on average than their corresponding Low Group. Overall, students who were in the High Group for Total Responses, Total Responses Non-Credit Questions, Questions Attempted, Credit Questions Attempted, Non-Credit Questions Attempted, Questions Completed, and Non-Credit Questions Completed scored at least 8% higher on all exams combined than their

counterparts in the Low Group. Of those, Questions Completed and Questions Attempted had the highest difference at 20%. In other words, students who used the site more scored on average nearly an entire letter grade higher than students who used the site less, which indicates that total usage of the site does correlate with better exam performance. In this course, exam scores larger than 90% warranted an A grade, exam scores between 90% and 80% earned a B, and so on at 10% intervals.

As noted above, however, one period that this does not hold true for is the Final Period. Of the aforementioned metrics, only Non-Credit Questions Attempted, Non-Credit Questions Completed, and Total Responses Non-Credit Questions have any significant correlation on the Final Period. Final Cumulative, on the other hand, has strong correlations between its metrics and the final exam score. The difference between the final exam and the midterms is that the midterms cover mostly new material, while the final covers both old and new material. Therefore, students who had studied early and frequently throughout the semester had already mastered more material on the final than the students who had studied little, and were less likely to use the site as much during that period. Thus, although this period breaks the site usage trend, it is less a reflection of the validity of those metrics and more a reflection of an inappropriate scenario to analyze them under.

By looking closer at the Total Usage metrics, additional trends can be seen to shed light on the validity of certain metrics. In all time periods, the Questions Attempted had a higher correlation to exam grades than Questions Completed, and the average scores between the High Group and Low Group were nearly identical. This suggests that while it is undoubtedly important to look at the amount of questions the student has

eventually solved correctly, it is even more important to look at the amount of questions the student has simply worked on. One reason this could be the case is because students can still learn from questions answered incorrectly, since the explanation is still shown to them in the OPE system. If the explanation were not shown, this correlation may be different.

Not only is it important to look at the total amount of questions attempted, but it is also important to look at the types of questions attempted as well. When breaking down the questions into credit and non-credit types, the results show that students who do more non-credit questions do on average 10%-15% better on every exam than students who do less of them. This is not surprising, since the students who are doing more non-credit questions are most likely more motivated than the other students and are also simply putting in more effort in general. However, what is interesting is that Credit Questions Attempted and Credit Questions Completed have a higher correlation and mean difference than their non-credit counterparts for all periods. This may be due to the fact that the 560 credit questions provided enough material for most students to master the subject, and that the top students didn't require the additional questions. It is likely that for systems that have credit and non-credit questions and significantly fewer credit questions to work on, that these correlations may be different.

Additionally, it can be seen from the data that the sheer amount of responses submitted is not as important to analyze as the amount of questions worked on. (Recall that "question" refers to one of the 988 questions in the OPE system, and that students can respond to the same question more than once). Although all periods (except for the Final Period) showed a positive correlation between Total Responses and exam scores,

and featured roughly a 10% average difference between the High Groups and Low Groups, it had less correlation and less impact on averages across the board than Questions Attempted. This encourages the notion that it is more important for students to attempt a greater number of problems, and thus see a potentially greater variety of material when studying, than the amount of time they spent on the site.

Further encouraging this notion is the data for the Average Responses Per Question metric. Before running the experiment, it was hypothesized that students who attempted questions multiple times would do better on the exams since they show a greater determination for getting the question correct. However, this was shown not to be the case, as this metric had very little correlation or effect on the average for all periods except for the Final Cumulative and Total periods. Even for those two periods, students who had a lower average, not higher, did better on the exams. This could be caused by the fact that students who knew the material well did not need to submit multiple responses to the questions, versus students who did not know the material well and thus needed to. In any case, due to the weak correlation and inconsistent behavior of the metric, Average Responses Per Question does not appear to be a strong candidate for assessing student performance.

Key Takeaways

- The more responses students submitted and questions students answered, the better they performed on the exams.
- For exams with cumulative material, the total site metrics during that period have less of a correlation with exam scores.

- It is more important to look at total questions attempted than total questions completed.
- There was stronger correlation between students who worked on more mandatory questions than students who worked on more optional questions.
- The amount of questions attempted is more important than the amount of responses submitted.
- The average amount of responses the student has per question does not correlate well with exam scores.

4.2 SELF-REPORTING METRICS

Table 9

List of Self-Reporting Metrics

Total Responses Self-Reporting
Percent Correct Responses Self-Reporting
Percent Correct Responses Non-Self-Reporting
Correct Response Ratio

Another hypothesis of this experiment was that students are not truthful about self-reporting correctness on question attempts, and thus the percentage of correct responses for such questions would not be a viable metric for determining student performance. To this end, the metrics listed in table 9 were analyzed. From these metrics, a couple trends can be seen.

First, the Percent Correct Responses Self-Reporting metric had only a low correlation with exam scores for all periods, and had no statistically significant difference in the averages for five of the eight periods. Additionally, 95% of all self-reported responses for the semester were marked as correct, whereas only 59% of graded responses were marked as correct. Although it is a possibility that the self-reporting questions were simply that much easier than the rest of the questions, it is unlikely given the degree of difference between the two percentages. Therefore, this supports the original hypothesis that this is indeed not a good metric for performance prediction.

One self-reporting metric that does strongly correlate with exam performance is the Percent Correct Responses Non-Self-Reporting metric. It featured greater than a 0.4 correlation for three of the eight periods, and it's lowest was 0.257. Furthermore, students in the High Group for this metric scored on average 13.7% higher in the Total period than their Low Group counterparts. This indicates that when students are not assessing themselves, their percent of correct responses in a system is a good indicator for performance on exams. In a system that lacks self-reporting questions, this would simply be the percent of correct responses for all responses.

A better question in relation to self-reporting metrics is, can the honesty of a student indicate better success in a classroom? To address this question, the Correct Response Ratio metric was devised, which essentially determines how often students self-report responses as correct compared to how often they actually get responses correct. The closer the metric is to 1, and thus the lower the metric is, the more honest the student is. By simply looking at the Total period, this may appear to be the case, since

students in the Low Group outscored their peers in the High Group by an average of 8.4% on the final exam.

Further analysis of the Correct Response Ratio, however, indicates that it is not a viable statistic to indicate student performance. First, it had a maximum absolute correlation of 0.259, and less than 0.20 for the rest of the periods. Second, the difference in exam averages was less than 5% for five of the eight periods between the High Group and Low Group. Instead, what is likely occurring for the Total period is that since students report such a high amount of responses as correct across the board, students with a lower Correct Response Ratio simply have a higher percent of correct non-self-reporting responses. As shown above, the Percent Correct Responses Non-Self-Reporting metric has a high correlation with exam performance, and thus the Correct Response Ratio metric is not likely to be a valid performance metric.

Key Takeaways

- The amount of times students report their answers as correct has no correlation with exam scores.
- The percent of correct responses students have for questions where the system grades the answer has a significant correlation with exam performance.
- The ratio of self-reported correct answer frequency to system-graded correct answer frequency has no correlation with exam performance.

4.3 RESPONSE CORRECTNESS METRICS

Table 10

List of Response Correctness Metrics

Percent Correct Responses
Questions Incomplete

In regards to correctness of student responses, one hypothesis of the experiment was that the percent of correct student responses would be a viable metric towards predicting student performance in the classroom. The reasoning is simple – as long as the exam questions are of the same material and somewhat as difficult as the questions on the website, students who get more responses correct on the website will likely get more correct responses on the test as well.

Upon analyzing the data, this hypothesis has also been supported. In every period, the Percent Correct Responses metric had a correlation of at least .298, and greater than .430 for five of the eight periods. Additionally, it had high differences in the means between the High Groups and Low Groups. For the Total period, students in the High Group outscored the Low Group by 14% on average, and the difference was even as high as 20% for Midterm 1 Period.

An interesting observation about this metric, however, is that while Midterm 2 Period, Midterm 3 Period, and Final Period all showed significant correlations and differences in the means between the two groups, Midterm 2 Cumulative, Midterm 3 Cumulative, and Final Period all had stronger correlations and larger mean differences.

This could likely be explained by the fact that the information in the course builds upon itself as the semester goes on. Each exam in Dr. Gould's class requires students to be able to apply information from prior exams in order to succeed. Therefore it is necessary to look at not only how correct the student has been answering questions from new material, but how correct the student has been answering questions from previous material as well. If the percent of correct responses is to be analyzed by a course, and the course includes information that builds upon itself throughout, then the metric should be analyzed from a cumulative standpoint and not from a specific exam period.

Another hypothesis of the experiment was that the more questions a student would leave incomplete (as in, attempt the question at least once but never get it correct), the worse the student would do on the exams. However, this metric only had a statistically significant correlation in two of the eight periods, and did not have a significant difference between the means in six of the eight periods. Upon looking closer at the data, the average student attempted 513 questions, and yet only left an average of 13 questions incomplete. Therefore, this suggests that this metric is not a strong candidate to assess student performance.

Key Takeaways

- The percent of correct responses for the student has a high correlation with exam grades.
- The percent of correct responses should be analyzed from a cumulative standpoint for courses with content that builds upon itself.

- The amount of questions attempted but never solved by a student does not have a correlation with exam grades.

4.4 TIMING METRICS

Table 11

List of Timing Metrics

Cramming Percentage
Average Questions Per Day
Study Density Score
Questions Attempted to Daily Average Ratio

One feature that the OPE platform provides to professors is the ability to see how often the site has been used over the course of the semester. By the end of the first semester, it became clear that there was a definitive trend of student usage – starting about four or five days before an exam, usage would start ramping up, to the point where the weekend before the exam would see usage many times higher than average. This can be seen clearly in figure 2, which features a screen shot taken from the OPE platform. Each of the peaks in that graph corresponds to the weekend before an exam was taken.

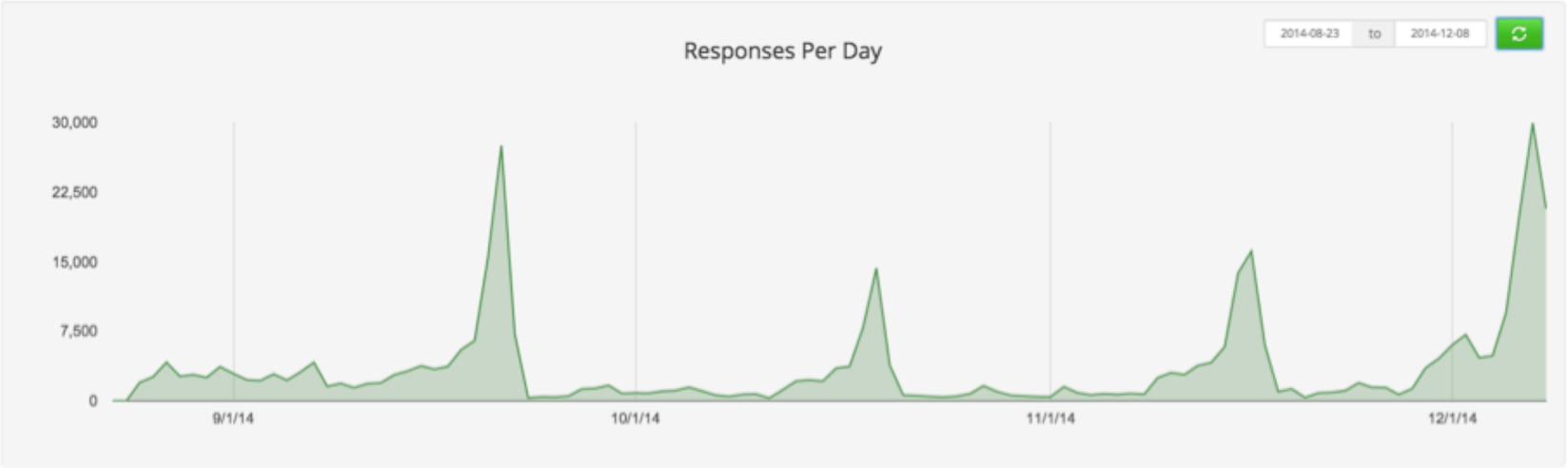

Figure 2

Graph of Student Responses per Day

It was hypothesized that this cramming behavior would not lead to good test results, since too much information is being processed for the student in too short of a span of time, and therefore any metrics that could capture a cramming behavior would be worth knowing to predict student performance in the classroom.

To this end, the Cramming Percentage metric was calculated, which contains the percent of the questions attempted that did not occur the weekend before the exam, compared to the questions attempted during the entire exam period. As an example, a Cramming Percentage of 90% means that the student attempted 90% of the questions for an exam period in the weeks leading up to the exam, and 10% of the questions the weekend before the exam. This was the only statistic that the High Group and Low Group were not split upon the median value. For this metric, the High Group represents students who had a Cramming Percentage greater than 10%. That is, the Low Group students did at least 90% of the questions the weekend before the exam and thus did the most cramming.

Analyzing the results of the Cramming Percentage metric support the original hypothesis. Although the correlation is low for all periods (the highest being 0.194 in Midterm 1 Period), the difference in means between the High Group and Low Group is significant (the lowest being 8% for the Total Period, and the highest being 20% for Midterm 1 Period). This indicates that it is not a linear correlation, yet still has a significant effect on exam scores. Therefore, the data suggests that this is a strong metric to look at when predicting student performance in a class.

A resulting hypothesis from this data was that students that had a more highly concentrated use of the site would do worse on exams than students that spread out their

work. The Average Questions Per Day metric was thus calculated to show how many questions students attempted per day, on days that at least one question was attempted. The results for this metric however were less than conclusive. For all periods, the absolute correlation was less than 0.2, and the correlation even switched from being positive in two periods to negative in three others. Furthermore, the difference in means between the High Group and Low Group was insignificant for two periods and as low as 3.9% for the Total Period. The reason for this is likely due to contradicting factors that affect this metric. On one hand, students who cram will have a higher value for this metric and will be expected to do worse, but on the other hand, students who do more questions will have a higher value but will be expected to do better. As a result, the Average Questions Per Day metric should not be used to predict student performance.

A similar approach towards testing this hypothesis is to instead look at how many days the student spent using the site. However, one weakness with that metric is that students could theoretically do the same questions every day for the entire semester and have a very high value for the metric, but would not be expected to perform well because their questions attempted would be low. To account for that, the Questions Attempted to Daily Average Ratio metric was calculated. Essentially, the metric takes the total questions attempted for that period and divides it by the Average Questions Per Day. In the case where a student is doing the same questions every day, this would produce a low value. However, if the student only works on every question on exactly one day, this metric would be equivalent to the amount of days spent using the site.

The data collected shows that the Questions Attempted to Daily Average Ratio metric consistently has one of the strongest correlations and difference of means

throughout all periods. Only for Final Period did it have less than a 0.4 correlation, and had at least a 15% difference in means between the High Group and Low Group for six of the eight periods. This supports the hypothesis that a student that spreads out studying more does better on the exams, and suggests that the Questions Attempted to Daily Average Ratio metric is significant to look at when predicting student performance. In fact, due to its higher correlation and larger average difference between group means, it is more significant to consider than the Cramming Percentage metric.

One potential weakness with the Questions Attempted to Daily Average Ratio metric when applied to the Total Period is that it does not account for the possibility of a student doing all studying for the semester in just one or two exam periods, and not using the site for the rest. As an example, let's say the average student in the class works a total of 30 days on the website in the semester. A student may work 20 days each for the first and second midterm, but not use the website at all for the third midterm and final. The student thus worked 40 days in the semester, which is above average, but yet he is likely to not do well in the class because he likely did badly on the third midterm and final due to lack of preparation.

To see the effect that a consistent usage of the website throughout the semester had on performance, the Study Density Score metric was created. Essentially, this metric measures how spread out the student's site usage was over all exam periods, with lower values indicating the usage was ideally spread out, and higher values indicating the student likely concentrated usage on specific periods or exams. The results show that there was a -0.461 correlation between total score and this metric, and the difference in means between the High Group and Low Group was 8.3%. This suggests that this is a

good metric for predicting performance. Knowing this, the potential weakness of the Questions Attempted to Daily Average Ratio metric is also validated. For future research, that metric's calculation should be updated to account for this scenario.

Key Takeaways

- Students who do 90% or more of their studying the weekend before an exam do on average one to two letter grades worse than students who do not.
- The average amount of questions a student works on per day does not correlate to exam performance.
- The ratio of questions attempted to the average questions attempted per day yields a strong correlation to exam performance.
- The degree of which site usage was spread out over the semester has a high correlation to exam performance.

4.5 DURATION METRICS

Table 12

List of Duration Metrics

Average Explanation Duration
Average Duration
Average Duration Correct Responses
Average Duration Incorrect Responses

Another feature of the OPE platform is the ability to see how long students spent working on each question and viewing the explanation. For this experiment, it was hypothesized that the longer the student spends working on questions and viewing the explanations, the more effort he is putting into the question, and thus the more likely he is to score better on the exams.

To test this, the Average Duration metric was first calculated and analyzed. According to the data, Average Duration had a correlation of at least 0.3 for five of the eight periods, going as high as 0.351 for Midterm 1 Period. It also featured a difference of at least 10% in the means of the High Group and Low Group for each period. This therefore reinforces the hypothesis and suggests that this is a good predictor for student performance.

Next, the Average Duration metric was broken down into two subsets – Average Duration Correct Responses, and Average Duration Incorrect Responses, to measure the difference in validity between how long the students spent working on responses that they got correct versus incorrect. The results show that the Average Duration Correct Responses had a higher correlation with exam scores for every period than both Average Duration Incorrect Responses and Average Duration. The difference in means for the High Groups and Low Groups were nearly identical for the three metrics across all periods, however.

The reason behind the higher correlation with Average Duration Correct Responses is likely due to the self-reporting questions that the course offered. For any type of question, if a student is trying to game the system and simply answer to get the credit, he can answer in little to no time, by either clicking on a random multiple choice

option, typing in random strings into an input, instantly self-reporting as correct, and so on. However, the difference between self-reporting questions and all others is that the student is more than likely going to self-report the response as correct in this scenario, whereas in a scenario with actual input the student is more than likely going to record an incorrect response. Therefore, in systems relying heavily on self-reporting questions, Average Duration Correct Responses likely will have a higher correlation with exam scores than systems that rely heavily on student input questions. Further research should be done on this topic to validate that claim.

In addition to the Average Duration metric, the Average Explanation Duration metric was also calculated, to show how long a student spends on average viewing the explanation (feedback) for a question. Not a single period showed this metric to provide a statistically significant correlation or difference between the means. However, this metric could only be calculated for a total of 23 students, which indicates that there was likely a bug in the OPE system that was preventing this metric from being recorded. More research should be done on this to determine its validity in predicting student performance.

Key Takeaways

- The average duration the student spends working on a problem correlates strongly to exam performance.
- The average duration the student spends working on responses that he gets correct has a stronger correlation with performance than the duration for incorrect responses.

4.6 METRICS ANALYSIS SUMMARY

Upon analyzing the correlation between individual metrics and exam scores, it is clear that there is a strong positive correlation between site usage and exam scores. Of all simple usage metrics, the amount of questions the student has attempted is the most important to consider when predicting exam performance. For systems that give students the ability to self-report the correctness of their answers, there is no benefit to analyzing the percent of correct responses. For systems that grade student answers, the percent of correct responses has a very significant positive correlation on exam scores. Meanwhile, the data suggests that there is a highly significant negative correlation between students who do more cramming before exams and students that do not. Similarly, the more spread out a student's studying has been, the more likely they are to perform better on the exams. Finally, the longer the student works on each response, the better the student tends to do on exams.

5.0 PREDICTION MODEL RESULTS

The second objective of the experiment is to find which combination of metrics can be used to create a prediction model to accurately predict student test results. To do so, a linear model was first generated using the Automatic Linear Modeling tool in SPSS. In this tool, a forward stepwise regression using Akaike information criterion is used to determine which metrics should belong in the model. From there, a 10-fold cross validation was performed on the model to determine its effectiveness. This was accomplished using a combination of the REGRESSION function in SPSS to generate the predicted scores, and Microsoft Excel to generate the statistics for each test.

For this objective, there were a few questions this experiment sought to answer:

1. Which metrics can be used in combination to best predict student performance, and how accurately can they predict the total score in the class for students?
2. Could the same combination of metrics be used to predict every exam?
3. What is the impact of introducing the students' prior exam scores into the models?
4. Are the models more accurate at predicting individual exam scores or cumulative scores across multiple exams?
5. How accurate are the models that only take into account total usage metrics and no others?
6. Which combinations of metrics appear in the most models?

To answer these questions, models were generated and validated against each exam in multiple ways, which will be explained further in this section. Each of the following factors appear in the analysis for every linear model generated:

- **Metrics in Model** – The metrics that compose the model.
- **Points Possible** – The total possible points that could be earned on the exam(s) the models are predicting against.
- **Correct Grade Predictions** – The percent of students the model correctly predicted the grade (A, B, C, D, F) for.
- **R^2** – The R^2 value calculated based on the predicted exam results and actual results.
- **Root Mean Squared Error –** The Root Mean Squared Error calculated based on the predicted exam results and the actual results.
- **Mean Absolute Error –** The Mean Absolute Error calculated based on the predicted exam results and the actual results.
- **Mean Absolute Error Percentage –** The percent of the Points Possible that is accounted for by the Mean Absolute Error. Essentially describes what percent of the exam score the predicted model was off by.

Please note that in the tables to follow, if multiple metrics appear in the model, they are listed in the order of greatest significance as determined by the stepwise regression.

5.1 TOTAL SCORE PREDICTION

To determine the optimal combination of metrics to predict exam scores, a stepwise linear regression was first run against all metrics to predict the total score for all students that completed the course. This analysis contains an additional statistic – the "Mean Absolute Error [A|B|C|D|F] Students", which refers to the Mean Absolute Error for students that the model predicted would get an A, B, C, D, or F on the exam in question. Please note that in this case, an A is a score of 90% or higher, a B 80% - 90%, and so forth. The results can be seen in table 13.

Table 13

Results of Prediction Model to Predict Total Class Score for Students

Exams	Total
Metrics in Model	• Percent Correct Responses Non-Self-Reporting Questions • Total Responses Self-Reporting Questions • Average Responses Per Question • Average Questions Per Day • Total Responses Non-Credit Questions • Correct Responses Non-Self-Reporting Questions • Correct Response Ratio
Points Possible	920
Correct Grade Predictions	41.8%
R^2	0.48
Root Mean Squared Error	111.5
Mean Absolute Error	87.4
Mean Absolute Error Percentage	9.5%
Mean Absolute Error A Students	50.6
Mean Absolute Error B Students	52.7
Mean Absolute Error C Students	81.9
Mean Absolute Error D Students	93.5
Mean Absolute Error F Students	161.1

According to the results of the cross validation in table 13, the model was able to predict the total exam scores for students within 9.5% of what they actually scored on average, using nothing other than metrics collected from the website. In addition, it was able to predict the final grade for about 42% of the students in the class, with an R^2 of 0.48. When considering each of these statistics, it is clear that the metrics in the model do indeed do well in predicting student performance. Additionally, since the Root Mean Squared Error is only about 127% that of the Mean Absolute Error, this indicates that

there are not many extreme outliers in the data and that most students conform to this model well.

However, one observation made from the results is that the mean absolute error increases nearly consistently as the model predicts lower exam scores – that is, when the model predicts a student will earn an A or B, it is more accurate than for students it predicts will earn a C, D, or F. Statistically, the students that the model predicts will earn an A or B are the largest positive outliers from the average in terms of the metrics included in the model. Since those students are the most accurately predicted, this further suggests that students who more strongly reflect the behaviors accounted for in this model will have a higher chance of succeeding in the class. Finally, the data in reinforces this notion, as it shows that students will on average score significantly higher in the class than their counterparts who were predicted to do worse.

Table 14

Average Actual Total Scores for Students with Each Predicted Grade

Predicted Grade	**Average Actual Total Score**
A	839
B	793
C	689
D	600
F	457

Although the aforementioned observation helps reinforce certain trends in the data, it also raises a new question. Just as students predicted to get an A are the most positive outliers from the average in the set of behaviors included in the model, students

predicted to get an F are the most negative outliers. Of all grades, the model does the worst at predicting these students. Is that because those students demonstrate the same set of behaviors but in a more inconsistent manner? Or do those students demonstrate a different set of behaviors altogether? To help answer this, the class was split into two groups around the median total score, and stepwise linear regressions were run against each. The results can be seen in table 15.

As can be seen in the results, the model for higher performing students has a significantly higher R^2, lower Root Mean Squared Error, and lower Absolute Mean Error than the model for lower performing students. This clearly indicates that the model for higher performing students performed better than its counterpart. Upon looking at the metrics that compose each model, both models consist of similar metrics. In both models, Average Questions Per Day and Percent Correct Responses Non-Self-Reporting Questions make an appearance, and although the other two metrics in the High Performing Student model are not in the other, the Low Performing Student model has similar metrics based around self-reporting questions and credit questions. Since the metrics in the two models are similar and yet perform so much differently, this suggests that the lower performing students exhibit similar behaviors as higher performing students overall, but less consistently.

Table 15

Prediction Model Results for High Performing and Low Performing Students for the Total Semester

Group	High Performing Students	Low Performing Students
Exams	Total	Total
Metrics in Model	• Percent Correct Responses Non-Self-Reporting Questions • Average Questions Per Day • Percent Total Responses for Self-Reporting Questions • Credit Questions Attempted	• Percent Correct Responses Credit Questions • Cramming Percentage • Average Questions Per Day • Average Responses per Credit Question • Percent Correct Responses Non-Self-Reporting Questions • Total Responses Credit Questions • Percent Correct Responses Self-Reporting Questions
Points Possible	920	920
Correct Grade Predictions	55.5%	46.9%
R^2	0.26	0.14
Root Mean Squared Error	45.1	107.2
Mean Absolute Error	37.0	88.5
Mean Absolute Error Percentage	4.0%	9.6%

One notable difference between the two aforementioned models is that the Low Performing Students model contains the Cramming Percentage metric whereas the other does not. This either suggests that the amount of cramming done by lower performing students may have more of an effect on their exam grades than for higher performing students, or that there is more of a disparity in amount of cramming between lower

performing students. Given the high difference in the means analysis for the Cramming Percentage, this is likely due to the latter of the two explanations.

5.2 INDIVIDUAL EXAM PREDICTION

Although it was shown that the Total Score Prediction model was fairly accurate in predicting student performance, it does not indicate how accurate models are for individual exams. After all, if a professor were to create and use a prediction model for predicting exam scores in his own classroom, he would likely be doing so at a point during the semester itself in order to predict an upcoming exam, not at the end of the semester when the grade has been determined. To determine this, models were next created and validated to predict each individual exam. All metrics used in the following models were cumulative – that is, they represent the behavior from the beginning of the semester up until the exam in question. This was done because the exams represent content that builds upon each other, and thus study behaviors from all prior exams should be incorporated into a prediction for any one exam. Also note that each model only accounts for students that took the exam being predicted for. The resulting data can be seen in tables 16 and 17.

According to the results, the exact combination of metrics for the total exam model did not appear in any of the models for each individual exam, and each exam featured a unique combination from each other. This indicates that even within the same class, different behaviors can have more or less significance in predicting exam scores for various exams. That being said, the models were still similar to each other. Four of the seven metrics in the total model were in the models of at least two of the individual

exams, and there were four metrics (Questions Per Day, Credit Questions Attempted, Average Responses Per Credit Questions, Percent Correct Responses Non-Self-Reporting Questions) that appeared in at least three of the individual exam models. Therefore, even though the same combination of metrics might not yield the most optimal prediction results for each exam, it will likely still perform comparably.

It is also worth noting that each individual exam model performed worse than the total exam model, with the R^2 of the midterm 1 model being the only field of any of the individual models performing better than the fields of the total model. This may indicate that the models are simply better at predicting cumulative score totals than individual exams given a period of metrics. This concept will be explored more in Section 5.4.

Table 16

Prediction Model Results for Individual Exams

	Midterm 1	Midterm 2	Midterm 3	Final
Metrics in Model	• Credit Questions Attempted • Percent Correct Responses Non-Credit Questions • Average Responses Per Credit Question • Cramming Percentage • Correct Response Ratio • Percent Correct Responses Self-Reporting Questions • Questions Incomplete	• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Responses Per Credit Questions • Correct Response Credit Questions • Total Responses-Non-Self-Reporting Questions • Average Questions Per Day • Average Responses Per Non-Credit Questions • Percent Total Responses Self-Reporting Questions	• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Responses Per Credit Questions • Average Questions Per Day • Average Responses Per Question	• Percent Correct Responses Non-Self-Reporting Questions • Total Responses Credit Questions • Average Responses Per Question • Average Questions Per Day • Correct Response Ratio
Points Possible	180	180	180	380
Correct Grade Predictions	40.9%	35.7%	37.5%	38.6%
R^2	0.52	0.40	0.41	0.42
Root Mean Squared Error	25.87	25.9	29.7	55.3
Mean Absolute Error	19.5	20.4	23.9	42.9
Mean Absolute Error Percentage	10.9%	11.3%	13.2%	11.3%

5.3 PREDICTION WITH PRIOR EXAM SCORES

In the models for individual exams, there were essentially two pieces of information that were represented by the cumulative metrics – how much knowledge the student concluded the previous exam with (based on the metrics gathered for all prior exams), and how much effort the student spent on the site for the current exam (based on the metrics gathered for the current exam period). However, a more definitive measure of how much knowledge the student concluded the previous exam with is their actual prior exam scores. It was thus hypothesized that a model that incorporates the previous exam scores along with site metrics would outperform individual exam models using site metrics alone.

To test this hypothesis, linear models were generated for each exam using previous exam scores along with metrics collected specifically during that exam period (i.e., no cumulative). Note that this was not done for the first exam because there were no prior exams to incorporate into the model. The results can be seen in table 17.

Table 17

Prediction Model Results for Models Factoring in Previous Exams

	Midterm 1	Midterm 2	Midterm 3	Final
Metrics in Model	• Credit Questions Attempted • Percent Correct Responses Non-Credit Questions • Average Responses Per Credit Question • Cramming Percentage • Correct Response Ratio • Percent Correct Responses Self-Reporting Questions • Questions Incomplete	• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Responses Per Credit Questions • Correct Response Credit Questions • Total Responses-Non-Self-Reporting Questions • Average Questions Per Day • Average Responses Per Non-Credit Questions • Percent Total Responses Self-Reporting Questions	• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Responses Per Credit Questions • Average Questions Per Day • Average Responses Per Question	• Percent Correct Responses Non-Self-Reporting Questions • Total Responses Credit Questions • Average Responses Per Question • Average Questions Per Day • Correct Response Ratio
Points Possible	180	180	180	380
Correct Grade Predictions	40.9%	35.7%	37.5%	38.6%
R^2	0.52	0.40	0.41	0.42
Root Mean Squared Error	25.87	25.9	29.7	55.3
Mean Absolute Error	19.5	20.4	23.9	42.9

As can be seen in the results, the models incorporating prior exam scores outperformed their metrics-only counterparts in every statistic. The Midterm 2 model correctly predicted the grade for 5% more of the students, while the difference for midterm 3 was 10% and for the final 17%. In addition, the R^2 value was roughly 0.2 higher for midterm 1 and 2, and 0.3 higher for the final. This supports the hypothesis that the original exam scores would improve the model performance.

Due to the degree of which the models were improved by introducing exam scores, it was next questioned whether the site metrics were needed at all for prediction models, and if models that incorporated nothing but prior exam scores would perform as well or better than models that incorporated both. To this end, models involving only prior exam scores were run through 10-fold cross validation, with the results shown in table 18.

Table 18

Performance of Prediction Models Using Only Prior Exam Scores

	Midterm 2	Midterm 3	Final
Metrics in Model	• Midterm 1 Score	• Midterm 1 + 2 Score	• Midterm 1 + 2 + 3 Score
Points Possible	180	180	380
Correct Grade Predictions	34.9%	43.9%	54.7%
R^2	0.53	0.58	0.69
Root Mean Squared Error	23.0	25.1	35.7
Mean Absolute Error	18.6	19.7	27.9
Mean Absolute Error Percentage	10.3%	10.9%	7.3%

According to the results, each of the models using just prior exam scores performed worse than the models that incorporated exam scores and site metrics, yet performed better than models that just included site metrics. This holds true for every statistic related to the model performance. These results thus suggest that previous exam scores are clearly the most important factor when predicting future exam performance. If absolutely no other metrics are available to a professor, creating a linear regression equation solely based on past exam performance will reasonably accurately predict how a student will do on an upcoming exam. However, if a professor wants the prediction to be as accurate as possible, he should use a combination of site metrics and exam scores, which further reinforces the notion that student behavior metrics do correlate with exam performance.

Another observation from the results is that the accuracy of the final model was considerably higher than that of the other models. As noted previously, one of the major differences between the final and the rest of the exams is that although the final does contain new material, it also tests students again on content that appeared in previous exams. Therefore, the result of previous exams likely has more of an impact on how well a student will perform on the final, and thus likely explains this observation. As noted above, a professor should include prior exam scores into the model for predicting individual exam results, but this especially holds true for any exams that involve testing concepts previously tested.

5.4 CUMULATIVE EXAM PREDICTION

As noted in section 5.2, each individual exam model performed worse than the total exam model. It was thus hypothesized that student behavior prediction models are better at predicting cumulative exam results than individual exam results. This would make sense conceptually as well. Consider the scenario where a student did not use the site for the first midterm and scored an F, but did use the site extensively for the second midterm and scored an A. Since his usage of the site is only half that of the students that have used the site for both midterms, the individual exam model would most likely predict roughly a C for midterm 2, which is not accurate. However, a model that predicts his midterm 1 and midterm 2 score combined would also likely predict a C, which is accurate to what his actual midterm 1 and midterm 2 average would be.

To test this hypothesis, two new models were created and validated – one which would take the site metrics from the start of the semester until midterm 2 to predict the cumulative score of midterm 1 and 2, and one which would take metrics until midterm 3 to predict the cumulative score of all three midterms. Note that this was not done for the final, since conceptually this would be equivalent to the total model. Nor was it done for midterm 1, since it would be equivalent to the midterm 1 individual model. Also note that past exam scores were not incorporated into these models. Results can be seen in table 19.

Table 19

Performance of Prediction Models to Predict Cumulative Exam Scores Using Only Site Metrics

	Midterm 2	Midterm 3
Metrics in Model	• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Responses Per Question • Average Questions Per Day • Correct Responses Credit Questions • Percent Correct Responses Self-Reporting Questions • Questions Incomplete	• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Responses Per Credit Questions • Average Questions Per Day • Correct Responses Credit Questions • Percent Correct Responses • Percent Correct Responses Self-Reporting Questions • Percent Total Responses Self-Reporting • Percent Correct Responses Non-Self-Reporting Questions
Points Possible	360	540
Correct Grade Predictions	40.0%	39.2%
R^2	0.49	0.49
Root Mean Squared Error	46.9	65.0
Mean Absolute Error	34.9	50.5
Mean Absolute Error Percentage	9.7%	9.3%

The results from table 19 show that the cumulative models performed better in every category than the equivalent individual exam models. Furthermore, the metrics involved in both types of models were nearly identical. For midterm 2, four metrics were included in both models, while three metrics were in both models for midterm 3. Of the metrics not included in both models, the majority was still highly related. As an example,

the midterm 2 individual model contained Average Responses Per Credit Questions and Average Responses Per Non-Credit Questions, whereas the cumulative midterm 2 model contained simply Average Responses Per Question. Since the metrics in both sets of models are highly related and yet the predictive performance for the cumulative models was significantly better than that of the individual exam models, the hypothesis has been supported. Therefore, if a professor were to create such prediction models for his class, he should use it to predict cumulative exam scores instead of individual ones.

Additionally, the cumulative exam models performed comparatively to the individual exam models with exam scores. Although the R^2 was still significantly higher for the models with exam scores, the midterm 3 cumulative model had a better mean absolute error percentage, while the midterm 2 cumulative model had a mean absolute error percentage just 0.3% off its counterpart. This further indicates the viability of using only site metrics to predict performance in the classroom.

5.5 TOTAL USAGE METRIC PREDICTION

In section 4, it was shown that some of the strongest individual metrics that could be used to predict student performance were centered on the total usage of the website. Clearly, metrics such as Total Responses and Questions Attempted alone have a large impact on how a student performs in the classroom. As a result, a professor may be inclined to simply make a regression equation based solely on one of these metrics, which would be easier to calculate than models involving several metrics. However, it is unclear how well such an equation would perform. To find out the answer, three models were created and validated using exactly one metric each, using three of the most

individually significant total usage metrics. Each model was used to predict the total score in the class. The results can be seen in Table 20.

Table 20

Performance of Prediction Model Using Only One Metric

Exam	Total	Total	Total
Metrics in Model	• Total Responses	• Questions Attempted	• Questions Completed
Points Possible	920	920	920
Correct Grade Predictions	21.8%	25.5%	25.3%
R^2	0.10	0.22	0.19
Root Mean Squared Error	147.4	138.1	139.1
Mean Absolute Error	119.0	111.5	111.6
Mean Absolute Error Percentage	12.9%	12.1%	12.1%

The results show that when using nothing else but Questions Attempted and Questions Completed, the prediction model was able to predict within 12.1% of the students' final grades on average. Total Responses did the worst of all three, which further reinforces the claims in section 4 that attempting more questions is more important for students than submitting more responses. However, all three models did significantly worse in every category than the total model outlined in section 5.1. Therefore, this suggests that while using a single significant metric to predict exam scores will still produce a roughly accurate prediction, it will not predict as accurately as a model that incorporates a variety of different metrics.

5.6 PREDICTION MODEL METRICS COMBINATIONS

Of the aforementioned models described throughout this section, there were a total of 10 created that incorporated the whole class and contained multiple different metrics. However, each of these models contained different combinations of metrics, which leaves it unclear as to which combinations should actually be used if a prediction equation were to be applied to a class. Therefore, analysis was run on each combination of metrics in each model, to determine which combinations and individual metrics appear in the most models. Tables 21 and 22 show the results below.

Table 21

Combination of Metrics Appearing in the Most Prediction Models

Metrics	Models
• Percent Correct Responses Non-Self-Reporting Questions • Average Questions Per Day	7
• Credit Questions Attempted • Percent Correct Responses Non-Self-Reporting Questions • Average Questions Per Day	4
• Credit Questions Attempted • Average Responses Per Credit Question	4
• Correct Responses Credit Questions • Average Questions Per Day	4
• Percent Correct Responses Non-Self-Reporting Questions • Average Questions Per Day • Average Responses Per Question	4
• Credit Questions Attempted • Correct Responses Credit Questions • Percent Correct Responses Non-Self-Reporting Questions • Average Questions Per Day	3
• Credit Questions Attempted • Percent Correct Responses Self-Reporting Questions	3
• Credit Questions Attempted • Average Responses Per Credit Question • Percent Correct Responses Non-Self-Reporting Questions • Average Questions Per Day	3
• Average Questions Per Day • Percent Total Responses Self-Reporting Questions	3

Table 22

Single Metrics that Appear Most Frequently in Prediction Models

Metric	Models
Average Questions Per Day	9
Percent Correct Responses Non-Self-Reporting Questions	7
Credit Questions Attempted	5
Average Responses Per Credit Question	4
Correct Responses Credit Questions	4
Average Responses Per Question	4
Percent Correct Responses Self-Reporting Questions	3
Percent Total Responses Self-Reporting Questions	3
Percent Correct Responses Non-Credit Questions	3
Correct Response Ratio	3

Based on the results shown, the combination of the Percent Correct Responses Non-Self-Reporting Questions metric and Average Questions Per Day metric appeared in the most models. In fact, the Percent Correct Responses Non-Self-Reporting Questions only appeared in models when accompanied by the Average Questions Per Day. This may suggest that a relationship exists between these two variables. Furthermore, since that combination appears in models with Credit Questions Attempted and Average Responses Per Question four times each, this suggests that the potential relationship is strengthened when adding those factors in. Alternatively, it could mean that those variables are simply the most independent from other variables, and thus appear in so many models because no other variables can account for what they offer. In any case, due to the prevalence of that duo of metrics, and the fact that Credit Questions Attempted has a more significant effect on exam scores than Average Responses Per Question, it is

advised that the combination of Percent Correct Responses Non-Self-Reporting Questions, Average Questions Per Day, and Credit Questions Attempted be used when generating prediction models for a class.

5.7 PREDICTION MODEL SUMMARY

Upon creating, validating, and analyzing various student performance prediction models, it is reasonable to expect to be able to predict student test scores within one letter grade of what students will actually earn, by using nothing other than statistics collected from the website. This accuracy will improve when prior exam scores are added to the models, especially if the exam incorporates material previously tested. When creating a prediction model using site statistics, it is advised to start with a model including the percent of responses graded correctly by the system, amount of credit questions attempted (or just questions attempted in a system with no concept of credit versus non-credit), and the average amount of questions worked on per day. There are a couple trends to keep in mind when doing so. Students that the model predicts to score well will have a more accurate prediction than students that the model predicts to score poorly. Additionally, when only using site metrics, prediction models should be used to predict cumulative test scores for a combination of exams, instead of predicting the results of a single exam.

6.0 OUTLIER ANALYSIS

The final objective of this experiment is to find outliers in the data and determine if any common trends exist between them. To accomplish this, the prediction model used to predict the total exam grades was run against each of the students, and the resulting predicted scores were compared to the actual scores. The top ten outliers of which the system over-predicted the score were then gathered and analyzed, as were the top ten outliers of which the system under-predicted the score. In the data that follows, student names have been omitted. Please note that only students who completed the course were included in this analysis. As a reminder, the prediction model used for total exam grades contained the following metrics:

- Average Responses Per Question
- Percent Correct Responses Non-Self-Reporting Questions
- Average Questions Per Day
- Total Responses Non-Credit Questions
- Correct Responses Non-Self-Reporting Questions
- Total Responses Self-Reporting Questions
- Correct Response Ratio

6.1 UNDER-PREDICTED OUTLIERS

Table 23

Observations Regarding Largest Under-Predicted Outliers from Data

Student	Predicted	Actual	Observations
A	2	404	Had a very high average responses to questions.
B	539	877	Not many questions attempted, but averaged many attempts per question.
C	337	617	Used the site exactly twice throughout the semester.
D	541	782	Didn't use the site for the final, only answered about half the questions provided.
E	541	767	Had a high average number of responses per Non-Self-Reporting questions.
F	497	722	Only used the site for the final and crammed. Did poorly on the final but well on other exams.
G	536	756	Only used the site for the final, worked on only credit questions, and crammed.
H	596	797	Had low number of responses and questions attempted. The performance on each exam was similar to site usage for each exam.
I	627	827	Had low number of responses and questions attempted. Didn't use the site for the final and performed poorly on it.
J	505	705	Worked almost exclusively on Non-Self-Reporting questions. Had a low correct response percentage.

A couple trends can be seen in the data in table 23. First, there were three cases (students A, B, and E) that had an abnormally high Average Responses Per Question metric value. As discussed in section 4, this metric has inconsistent correlation behaviors, which this further demonstrates. On the whole with total exam scores, this metric holds a negative correlation, but for these three students it may have had a positive correlation. For students B and E, who both earned an A in the course, the high amount of responses may instead have been an indicator for their determination to figure out questions after initially solving them incorrectly, which is a positive trait for students to have. Thus, if a professor were to use this metric in a prediction model, he should be wary of the

predictive performance of the model for students that display extreme values of this metric.

Another trend that can be seen is that the majority of students in table 23 simply did not use the site very often throughout the semester. Unfortunately, this is an unavoidable aspect of analyzing such metrics – some students simply do not need as much practice as other students to master the material. Furthermore, Dr. Gould provides additional material for practice outside of the website, such as old exams, quizzes, and more. These students may have just been concentrating their studying on those materials instead of the website, which is impossible to measure using just the data available in this experiment. For these students, there is not much that can be done from a statistical analysis perspective to predict their exam performance.

Finally, student J shows another interesting set of behaviors. This student solved primarily Non-Self-Reporting questions (710 total Non-Self-Reporting question responses versus 183 total self-reporting responses), and appeared to be honest about his performance on the self-reporting responses, as he reported only 44% of his responses as correct (compared to the class average of 91%). Since students had a tendency to do more self-reporting questions than Non-Self-Reporting questions, and mark them correct almost every time, this student had a significantly low percentage of correct responses. Typically, this indicates that the student will do worse on exams, but in his case he still earned a B in the class. Despite the data shown in section 4, which indicated that Percent Correct Responses and Percent Correct Responses Non-Self-Reporting Questions performed roughly equally in terms of predicting performance, this suggests that the

latter metric should be analyzed instead of the former metric for systems that provide self-reporting questions.

6.2 OVER-PREDICTED OUTLIERS

Table 24

Observations Regarding Largest Over-Predicted Outliers from Data

Student	Predicted	Actual	Observations
K	593	219	Had an average amount of questions attempted and responses, but left a high amount of questions incomplete.
L	657	304	Only used the site immediately before the final.
M	493	176	Used the site exclusively over the course of a couple days.
N	723	409	Used the site only for midterm 1 and the final.
O	624	326	Used the site only for midterm 1 and the final.
P	690	403	Only used the site for the final the last two days before the exam.
Q	457	196	Didn't use the site for the final.
R	513	255	Didn't use the site the weekend before the third or final exam.
S	616	363	Didn't begin using the site until the final.
T	490	252	Didn't use the site for midterm 2 or midterm 3.

Of the students listed in table 24, the clear trend is that the model over-predicts students who severely concentrate their usage of the site for one or two exams, and completely neglect the site for the rest of the semester. The reasoning behind this is simple – the students are cramming their site usage so densely into specific exam periods that as a whole, their usage is comparable to students who have evenly spread out their attempts throughout all exam periods. This causes the prediction model to classify them as heavy users of the system and thus over-predicts their performance.

Although the Average Questions Per Day appears in the total prediction model, this may not account for this use case as well as the Study Density Score, which did not make it in the model. Of the ten students in table 24, five of them had a Study Density Score more than double the class average for students who attempted at least 200 questions. Therefore, this suggests that it is important to consider outliers of the Study Density Score when looking for potentially over-predicted students.

6.3 OUTLIER SUMMARY

When looking at the ten most over-predicted and under-predicted students for the total semester, a couple lessons can be learned. If the Average Responses Per Question metric is used in the prediction model, a professor should be wary of students that show too high of a value outside the average, for the model may under-predict their performance. Additionally, although the Percent Correct Responses and Percent Correct Responses Non-Self-Reporting Questions metrics are nearly equivalent in terms of their correlation to student performance, an example of one of the outliers demonstrates the value of using the latter metric to the former when predicting student performance. Finally, outliers to the average Study Density Score should be considered as well, for that could indicate cases where the model may over-predict student performance.

7.0 CONCLUSION

When a professor wants to predict how well a student will do in his class based on the behavior metrics of the LMS being used, there are several metrics and approaches to consider. Without any other data, the single most effective predictor is simply the amount of questions that the student has worked on, regardless of whether or not the question was solved correctly. The more questions students attempt, the better they score on exams on average. In fact, students that attempted more questions than the median for the class averaged more than an entire letter grade higher than students below the median. Still, nearly any metric that measures the amount of time the student has spent on the LMS has a strong positive correlation to exam scores. This includes the number of questions solved correctly, total question responses submitted, number of questions attempted per question type if there are multiple, and so forth. There are additional strong positive correlations between the exam scores and the amount of time students spend working on questions, as well as the percent of responses that the student has gotten correct as graded by the LMS.

A slightly more complicated, but equally powerful, approach towards predicting performance is to look at the timing of the students' use of the LMS throughout the course. The amount of questions that the student attempts the weekend before an exam compared to the weeks leading up to the exam has a strong negative correlation to how the students will perform on the exam. It is not enough to simply look at the average amount of questions worked on per day, however, as that by itself has a low correlation. Instead, an effective approach has been to divide the course into equally important periods of time, and determine the relative amount of LMS usage in each period. Students

that spread out question attempts evenly throughout the course perform better than students who do all studying in shorts bursts.

In addition to the average questions attempted per day, there are a few metrics that should not be used to predict exam performance without any other context. The amount of questions that students attempt but never solve correctly was believed to have a negative correlation with exam scores, but no significant correlation appears to exist. Similarly, no significant correlation seems to exist between the average amount of responses per question and exam scores. Finally, although the percent of responses graded as correct has a strong correlation with exam scores, the percent of responses graded as correct by the student has very little correlation. For LMS platforms that allow students to grade their own work, any collected statistic regarding the correctness of their responses should be ignored.

When creating a prediction model to predict exactly what the student will earn in the course, it is advised to use a model that contains the amount of questions attempted, the percent of responses the LMS has graded to be correct, and the average questions worked on per day. Prediction models that included this combination of metrics were shown to predict within about 9% on average of what the student actually earned in the course. When predicting the result of an upcoming exam, it is advised to include the previous exam scores in the model as well, as that is shown to improve accuracy further. This especially holds true for exams that test previously covered material, such as a cumulative final exam. Models based solely on collected LMS metrics perform best when predicting the cumulative score of multiple exams, than the score of one individual exam. It is also important to keep in mind that students which the model predicts will perform

well will likely have actual exam results closer to what the model predicts, than students that the model predicts will do poorly.

There are certain cases that should be looked for when determining if a student will perform as predicted by the prediction model. Although the average amount of attempts per question may positively impact the accuracy of a model, extreme deviations in the average from this metric tend to produce predictions significantly lower than reality. Conversely, students that deviate too much from the average amount of cramming may have predictions significantly higher than reality, and thus this should also be kept in mind when evaluating student predictions.

For future work on this topic, the effect of the average time spent viewing question explanations should be looked at, as this data was improperly collected in this experiment. This analysis should also be repeated for other classrooms of differing subjects, to remove the effect that organic chemistry potentially has on the results, and should incorporate more thorough factor covariance analytics when creating prediction models. Additionally, for questions that have multiple ways of inputting answers (multiple-choice and self-evaluation, for example), the way students choose to input their answer should be analyzed for possible correlation with exam scores. Finally, the questions that students choose to work on should be analyzed in more detail, to determine if specific subsets or question types have correlation with exam scores.

REFERENCES

Arnold, Kimberly E., and Matthew D. Pistilli. "Course signals at Purdue: using learning analytics to increase student success." *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. ACM, 2012.

"Babson Study: Distance Education Enrollment Growth Continues, But at Slowest Rate Ever". *Online Learning Consortium.* Online Learning Consortium, 5 Feb. 2015. Web. 8 Aug. 2015.

Edwards, Stephen H., et al. "Comparing effective and ineffective behaviors of student programmers." *Proceedings of the fifth international workshop on Computing education research workshop*. ACM, 2009.

Feng, Mingyu, et al. "Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test?." *Educational Data Mining 2008*. 2008.

Filippidi, Andromahi, Nikolaos Tselios, and Vassilis Komis. "Impact of Moodle usage practices on students' performance in the context of a blended learning environment." *Proceedings of Social Applications for Life Long Learning* (2010): 2-7.

Haig, Thomas, Katrina Falkner, and Nickolas Falkner. "Visualisation of learning management system usage for detecting student behaviour patterns."*Proceedings of the Fifteenth Australasian Computing Education Conference-Volume 136*. Australian Computer Society, Inc., 2013.

Macfadyen, Leah P., and Shane Dawson. "Mining LMS data to develop an "early warning system" for educators: A proof of concept." *Computers & Education* 54.2 (2010): 588-599.

APPENDIX A

RELEVANT DATABASE SCHEMAS

Student Response Table

```
CREATE TABLE `student_response` (
  `id` int(11) unsigned NOT NULL auto_increment,
  `user_id` int(11) unsigned NOT NULL,
  `correct` tinyint(4) NOT NULL,
  `class_question_id` int(11) unsigned default NULL,
  `duration` int(11) NOT NULL default '0',
  `explanation_duration` int(11) default NULL,
  `date` datetime default NULL,
  `question_type` varchar(64) NOT NULL default '',
  `response_type` varchar(64) NOT NULL default '',
  `source` varchar(255) default NULL,
  `data` blob,
  PRIMARY KEY  (`id`),
  KEY `student_response_user` (`user_id`),
  KEY `student_response_class_question` (`class_question_id`),
  KEY `student_response_source` (`source`),
  KEY `student_response_date` (`date`),
  CONSTRAINT `student_response_class_question` FOREIGN KEY
(`class_question_id`) REFERENCES `class_question` (`id`) ON DELETE SET NULL
ON UPDATE CASCADE
) ENGINE=InnoDB AUTO_INCREMENT=399475 DEFAULT CHARSET=latin1;
```

Class Question Table

```
CREATE TABLE `class_question` (
  `id` int(11) unsigned NOT NULL auto_increment,
  `question_id` int(11) unsigned NOT NULL,
  `category_id` int(11) unsigned NOT NULL,
  `order` int(3) default NULL,
  `for_credit` tinyint(1) NOT NULL default '0',
  `mc_enabled` tinyint(1) NOT NULL default '1',
  `input_enabled` tinyint(1) NOT NULL default '1',
  `guided_enabled` tinyint(1) NOT NULL default '1',
  `user_id` int(11) unsigned default NULL,
  PRIMARY KEY  (`id`),
  KEY `class_questions_category` (`category_id`),
  KEY `class_questions_question` (`question_id`),
  CONSTRAINT `class_questions_category2` FOREIGN KEY (`category_id`)
REFERENCES `category` (`id`) ON DELETE CASCADE ON UPDATE CASCADE
) ENGINE=InnoDB AUTO_INCREMENT=1482 DEFAULT CHARSET=latin1;
```

User Table

```
CREATE TABLE `user` (
  `id` int(11) unsigned NOT NULL auto_increment,
  `name` varchar(50) NOT NULL default '',
  `password` varchar(512) NOT NULL default '',
  `email` varchar(50) default NULL,
  `username` varchar(50) NOT NULL default '',
  `external_id` varchar(50) default NULL,
  `type` enum('user','professor','admin') NOT NULL default 'user',
  PRIMARY KEY  (`id`),
  KEY `login` (`username`,`password`),
  KEY `user_type` (`type`)
) ENGINE=InnoDB AUTO_INCREMENT=430 DEFAULT CHARSET=latin1;
```

APPENDIX B

DATABASE QUERY

```
set @start_date='2014-08-23';
set @end_date='2014-12-09 09:20:00';
drop table if exists response_density;
drop table if exists stats;
drop table if exists student_response_2;
create table if not exists student_response_2 as (select sr.*, u.external_id as external_id
from student_response sr, ope_query_global.user u where u.id=sr.user_id AND
sr.date<=@end_date AND sr.date>=@start_date);
alter table student_response_2 add index ext_id (external_id, `date`, class_question_id);
create temporary table if not exists response_density as (select sr.external_id as
external_id, MONTH(sr.date) as month, DAY(sr.date) as day,
count(distinct(sr.class_question_id)) as count from student_response_2 sr where
sr.date<=@end_date AND sr.date >=@start_date group by sr.external_id, month(sr.date),
day(sr.date));
create temporary table if not exists stats as (
select CONCAT(""A"", u.external_id) as external_id,
(SELECT COUNT(1) FROM student_response_2 sr WHERE
sr.external_id=u.external_id and sr.date<=@end_date AND sr.date>=@start_date) as tr,
(SELECT COUNT(1) FROM student_response_2 sr, class_question cq WHERE
cq.id=sr.class_question_id AND sr.external_id=u.external_id AND sr.date<=@end_date
AND sr.date>=@start_date AND cq.for_credit=1) as trcq,
(SELECT COUNT(1) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.correct=1) as cr,
(SELECT COUNT(1) FROM student_response_2 sr, class_question cq WHERE
cq.id=sr.class_question_id AND sr.external_id=u.external_id AND sr.date<=@end_date
AND sr.date>=@start_date AND cq.for_credit=1 AND sr.correct=1) as crcq,
(SELECT COUNT(1) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.question_type=""generic"" AND sr.response_type=""input"") as trg,
(SELECT COUNT(1) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.question_type=""generic"" AND sr.response_type=""input"" AND sr.correct=1) as
crg,
(SELECT COUNT(distinct(sr.class_question_id)) FROM student_response_2 sr,
class_question cq WHERE cq.id=sr.class_question_id AND sr.external_id=u.external_id
AND sr.date<=@end_date AND sr.date>=@start_date AND cq.order <> -1) as qa,
(SELECT COUNT(distinct(sr.class_question_id)) FROM student_response_2 sr,
class_question cq WHERE cq.id=sr.class_question_id AND sr.external_id=u.external_id
AND sr.date<=@end_date AND sr.date>=@start_date AND cq.order <> -1 AND
cq.for_credit=1) as cqa,
(SELECT COUNT(distinct(sr.class_question_id)) FROM student_response_2 sr,
class_question cq WHERE cq.id=sr.class_question_id AND sr.external_id=u.external_id
AND sr.date<=@end_date AND sr.date>=@start_date AND cq.order <> -1 AND
sr.correct=1) as qc,
```

```
(SELECT COUNT(distinct(sr.class_question_id)) FROM student_response_2 sr,
class_question cq WHERE cq.id=sr.class_question_id AND sr.external_id=u.external_id
AND sr.date<=@end_date AND sr.date>=@start_date AND cq.order <> -1 AND
sr.correct=1 AND cq.for_credit=1) as cqc,
(SELECT AVG(sr.duration) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.duration < 798952) as ad,
(SELECT AVG(sr.duration) FROM student_response_2 sr, class_question cq WHERE
cq.id=sr.class_question_id AND sr.external_id=u.external_id AND sr.date<=@end_date
AND sr.date>=@start_date AND cq.for_credit=1 AND sr.duration < 798952) as acd,
(SELECT AVG(sr.duration) FROM student_response_2 sr, class_question cq WHERE
cq.id=sr.class_question_id AND sr.external_id=u.external_id AND sr.date<=@end_date
AND sr.date>=@start_date AND cq.for_credit=0 AND sr.duration < 798952) as ancd,
(SELECT AVG(sr.explanation_duration) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.explanation_duration < 826817) as ae,
(SELECT AVG(sr.explanation_duration) FROM student_response_2 sr, class_question
cq WHERE cq.id=sr.class_question_id AND sr.external_id=u.external_id AND
sr.date<=@end_date AND sr.date>=@start_date AND cq.for_credit=1 AND
sr.explanation_duration < 826817) as ace,
(SELECT AVG(sr.explanation_duration) FROM student_response_2 sr, class_question
cq WHERE cq.id=sr.class_question_id AND sr.external_id=u.external_id AND
sr.date<=@end_date AND sr.date>=@start_date AND cq.for_credit=0 AND
sr.explanation_duration < 826817) as ance,
(SELECT AVG(sr.duration) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.duration < 798952 AND sr.correct=1) as adc,
(SELECT AVG(sr.duration) FROM student_response_2 sr WHERE
sr.external_id=u.external_id AND sr.date<=@end_date AND sr.date>=@start_date AND
sr.duration < 798952 AND sr.correct=0) as adi,
(SELECT AVG(rd.count) FROM response_density rd WHERE
rd.external_id=u.external_id) as da
from ope_query_global.user u
where u.id > 2 and u.external_id <> ""NULL"" group by external_id);
select * from stats where tr > 0;
```