Abstract — Online homework systems are being increasingly used for auto-graded, instant feedback homework and practice for students in math, science and engineering. Students may use these systems, which often allow multiple or unlimited tries, in ways that are different from completing traditional paper-based homework, however research relating online homework system patterns of usage and learning outcomes is limited. This study explores online homework submission patterns and their links to student learning outcomes (weighted individual grades) by analyzing the submission patterns of two second-year engineering courses (~130 students each) from our institution over the 2017-2018 academic year using WeBWorK, an open online homework platform.

Students in each of the two courses were clustered into three groups using a K-means algorithm based on when during the homework period they tended to submit attempts. Clusters were used to approximately represent a submission pattern, meaning groups of students that submit attempts mostly early, mostly late, or more evenly over the period. Conducting one-way ANOVAs for each course, we found that there is a significant difference between clusters (submission patterns) in terms of mean individual weighted grades on tests and exams (p < 1.07e-08, p < 2.68e-5). Post-hoc analyses revealed that the best performing cluster (students who submit attempts mostly early) had a mean tests/exams grades that were about 10% higher than worst performing cluster (students who submit attempts mostly late) (p < 2.6e-06, p < 9.9e-05).

Keywords: Online homework, WeBWorK, homework submission patterns, problem attempts, unlimited attempts, learning outcomes.

1. INTRODUCTION

Engineering courses have growing incentives to adopt online homework systems (OHW systems) due to benefits such as reduced marking time and rapid feedback. OHW systems typically allow students to reattempt problems and immediately correct errors in understanding. As a result, students use OHW systems in a different way from pencil and paper-based homework (PPH). Instructors, however, use OHW systems in a very similar way: assigning problems and collecting grades after the due date. This paper investigates how students use OHW to potentially increase learning by capitalizing on the capabilities of OHW systems.

In many studies, when compared to PPH or other/no alternatives, OHW increases student engagement [1]–[3], interest in course topics [4], understanding of subject matter [1], [5], [6], and motivation [6]–[9]. Research finds that OHW increased student retention rates in courses where it was used [9]–[11].

Students are generally reported to have positive reactions to OHW [11]–[13] and that this positive reaction is shared between different student populations regardless of what OHW system is used [14]. Students who prefer OHW over PPH tend to cite instant feedback [8], [9], [13], having the option to correct answers [9], and/or having multiple attempts [8] as the reason/s for their preference. Students have also said that, in subjects where OHW was used, they have perceived themselves to have improved due to OHW [4] and additionally feel more satisfaction in answering OHW where feedback is present [2], [3]. Students feel that OHW is an effective method for studying course material [15] and that doing OHW is worth the effort [11].

Researchers have disagreed on whether OHW has any statistical significance on mean exam marks, with many claiming it does not [8], [13], [16]–[18], some claiming that it does [19], [20], and others claiming that at least some positive correlation is present [21], [22]. Regardless of the effects of OHW on exam grades, most researchers do seem to agree that it is as good as PPH or is a suitable replacement for PPH when it comes to learning [1], [12], [17], [23].

While explorations of student satisfaction and comparisons to PPH have been positive, few studies have looked at how OHW can be configured to increase learning. Some have considered the impact of offering multiple attempts [24]–[26], with mixed recommendations. Huang [27] and Jayhooni [28] examined the impact of different online homework feedback types and styles. Kortemeyer [29] examined the time between consecutive submissions, in an effort to infer whether students were making serious subsequent attempts.
Specific research linking student usage patterns in OHW with measures of academic performance is still limited. The unique advantage of OHW systems is that student interaction with the system is recorded, allowing detailed analysis of usage, and the creation of testable, evidence-informed strategies to improve learning. Our goal for this pilot study is to see if we can identify usage patterns in OHW associated with test and exam success (as a proxy for learning). This may inform approaches educators may take to structure/layout OHW, better incorporate OHW into pedagogy, and/or to help students improve their study skills (and change their OHW usage pattern).

2. DATA COLLECTION

2.1. Course and Student Details

We collected WeBWorK usage data from MECH 221 and MECH 222 during the 2017-2018 academic year. MECH 221 and MECH 222 are two consecutive mechanical engineering courses that enroll the entire cohort of about 130 students in the first (MECH 221) and second (MECH 222) terms of second year. MECH 221 is effectively a 16-credit course that includes dynamics, electrical circuits, materials, solid mechanics, and differential equations. MECH 222 is effectively a 10-credit course that includes fluid mechanics, thermodynamics, and multivariable calculus. About 80% of the students in these courses used WeBWorK during their first-year mathematics courses at UBC.

2.2. Sources of Data

2.2.1. Student individual test/exam grades: We collected individual student test and final exam grade information for the two courses. Ethics research board approval was obtained for this study. MECH 221 has nine weekly tests and three final exams, and MECH 222 has six weekly tests and two final exams. We used the combined (weighted) tests and exam marks for each student as a proxy for learning.

2.2.2. WeBWorK student answer log: WeBWorK automatically generates a log file (answer_log) of student attempt data containing several hundred thousand rows of text, with each row representing one attempt an individual student made for a given problem. The information in each row consists of:

- The submission timestamp, accurate to one second.
- A student’s WeBWorK-assigned ID.
- The specific problem set and problem attempt.
- A binary string indicating a correct (‘1’) or incorrect (‘0’) attempt for each answer field in the problem. For example, if a question asks for position, velocity and acceleration values and the student answers only velocity correctly, the binary string in the answer_log file will be ‘010’.
- The exact answer string a student submitted. This can be a numerical answer (e.g. 32.51), an expression (e.g. 4/3*pi*6^3), a multiple-choice option (e.g. B), etc.

2.2.3. WeBWorK set definition files: Set definition files contain data that WeBWorK uses to import and export problem sets between courses. It also holds metadata related to a specific problem set and each of the individual problems within the set. WeBWorK generates one set definition file per problem set, from which the following pieces of information were used:

- The time and date a problem set opens (after which problems become visible and attempts are allowed).
- The time and date a problem set closes (after which no further attempts are allowed).
- The problems contained in a specific set and associated data: each problem’s weight, order in the set, filename, and number of attempts. In this study, all problems have equal weights in their respective problem sets.

Data from 558 problems across 39 problem sets were collected for analysis.

2.3. Data Cleaning

Using Python scripts, data from the three sources were cleaned up and written into sets of .csv files for statistical analysis. The data cleaning process filtered out the following:

- Students who had fewer than 100 submissions in a course. These would be outliers, as the number of questions and average attempts per student were typically much higher than 100 in both courses (see Table 1).
- Problems where the maximum number of attempts were limited (as opposed to unlimited attempts). Typically, these were multiple-choice problems. 186 problems were removed from MECH 221 and 23 were removed from MECH 222. A total of 8 problem sets that were purely composed of multiple-choice problems were filtered out as a result.

Our final data set summary details are shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Final data set summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MECH 221</strong></td>
</tr>
<tr>
<td>Students</td>
</tr>
<tr>
<td>Problem sets</td>
</tr>
<tr>
<td>Included problems</td>
</tr>
<tr>
<td>Mean (SD) attempts per student</td>
</tr>
</tbody>
</table>
3. STATISTICAL METHODS

We completed several stages of statistical analysis in order to identify patterns in student submissions, as well as possible associations between WeBWorK submission patterns and course outcomes. First, we aggregated all submission time data and defined a WeBWorK attempt distribution function (WADF), which is used in this study to characterize the WeBWorK submission time patterns. After that, we used the K-means algorithm for clustering and assigned students into three clusters. Each cluster represents one of three different patterns of submission times: early submissions, late submissions, and a third pattern (a more constant rate of submissions). We then analyzed the relationship between submission patterns and students’ final grades by applying a one-way ANOVA test. The following sections provide details about each stage of analysis.

3.1. WeBWorK Attempts Distribution Function

In order to find patterns in submissions, we needed to determine an objective way to characterize the distribution of submitted attempts as a function of time to be able to define thresholds for our cluster. We considered time to start when the problem set was released. Instead of considering different assignments separately, we integrated the attempts for all assignments together and scaled the available time into the normalized range [0, 1]. The reason for this scaling was to create a uniform measure for time, since the allowed time differed between assignments.

We then defined a WeBWorK Attempts Distribution Function (WADF) for each student. The WADF is denoted by \( F(x) = y \), where \( x \) is the time taken to complete the assignments normalized by the time available, and \( y \) is the percentage of the overall attempts completed within normalized time \( x \). For instance, \( F(0.7) = 0.22 \) means that 22% of the attempts from a student had been submitted within the first 70% of the time allowed for all assignments. We then produced a graph of the WADF for each student (Figure 1) that characterizes their submission pattern and can be viewed as a cumulative distribution function.

3.2. Wasserstein Distance

The Wasserstein distance [30] is a way to measure how “different” two distributions are. In our case, the “distance” between two WADFs represents the difference between the corresponding students’ WeBWorK submission patterns.

Figure 1: Sample WeBWorK Attempt Distribution Function (WADF) showing results for one student in MECH 221. “Percentage of available time” can be also read as “percentage of time since the release of the assignment”.

The L2-Wasserstein distance between two cumulative distribution functions, \( F \) and \( G \), is as follows [31]:

\[
W_2(F, G) = \left( \int_0^1 |F^{-1}(z) - G^{-1}(z)|^2 \right)^{1/2}
\]

This method accounts for the distance between the quantiles of two distributions, while other distance measures, such as the L2 distance, do not take this into consideration. We used the above expression to compute the Wasserstein distance between students’ WADFs.

3.3. Clustering

The K-means algorithm [32] is an unsupervised machine learning algorithm used for clustering that groups data sets into \( k \) clusters. In our analysis, the object to be clustered is the WeBWorK attempts distribution function for each student, where we used the Wasserstein distance to measure the “difference” between objects.

We first needed to determine the number of clusters, \( k \). To choose a reasonable number of clusters for this data set, we used a visual approach and interpreted the graph shown in Figure 2. (This paper only shows the graph for MECH 221, since MECH 222 produced a similar result.)
The x-axis of Figure 2 shows the number of clusters, and the y-axis shows the sum of square distances between elements within the same cluster. The values on the y-axis thus indicate the variation inside a cluster, or the “within-cluster sum of square”. A lower within-cluster sum of square means that the objects in the same cluster are more similar to one other. Naturally, with the increase of clusters, the within-cluster sum of square decreases.

We aimed to pick the number of clusters that produces a lower within-cluster sum of squares without introducing too many clusters. In our case, we picked the “elbow point” [33] at three clusters. At this point, increasing the number of clusters to four would not result in a significant decrease in the within cluster sum of squares. However, a decrease of the number of clusters to two would lead to a proportionally large increase in the within cluster sum of squares. Three clusters were thus chosen as the optimal number.

After applying K-means clustering with three clusters to the data set, differences between clusters in their submission patterns were observed, and an ANOVA was applied to determine differences in mean test/exam marks for students in each cluster.

4. RESULTS

Initial observations of WADFs visualized as colour-coded by weighted tests and exams marks indicated that students who had higher tests/exams marks tended to start their homework earlier, and students who had lower tests/exams marks tended to start their homework later than their peers. This was observed for both MECH 221 and MECH 222 (Figure 3 and Figure 4).
We were able to cluster students into three distinct groups based on submission patterns. Figure 5 illustrates the clustering results for MECH 221, showing 10 randomly-chosen students from each cluster, with each student corresponding to a WADF. The three clusters identified appear to correspond to three different submission patterns.

Figure 5 shows that students in cluster 1 tend to submit more attempts early, while those in cluster 3 are more likely to submit more attempts closer to the deadline.; students in cluster 2 do not show a tendency towards either behaviour. The clustering result and interpretation for MECH 222 is similar.

After our clustering process, we labeled each student with their cluster number and compared clusters based on average weighted tests and exams marks. Differences between clusters can be visualized in box plots, showing the cluster with the early submission pattern has higher average tests and exams marks (Figure 6, Figure 7). We can see that students who submit most attempts earlier (cluster 1) tend to achieve better grades, while those who submit most of their attempts later (cluster 3) are more likely to get lower marks; this is true for both courses.

Summary data (Table 2 and Table 3) shows that, for a problem set open for about a week, the best performing cluster would, on average, submit 30%, 60%, and 90% of their attempts within about two days, three days, and five days, respectively. In contrast, the worst performing cluster grades-wise would only complete an average of 30% of their attempts within about five days and 90% well past the sixth day (right before the deadline). The actual number of attempts was fairly similar among the clusters, although cluster 3 in both courses had slightly lower attempts.

To check the association between WeBWorK attempts patterns and test/exam marks, we applied the ANOVA test to the data from each course (Table 4 and Table 5).
We found p-values less than 2.9e-5 for both courses, which indicates that we have sufficient statistical evidence to conclude that the WebWork submission patterns we studied are significantly associated with students’ tests and exams marks.

Finally, we considered the difference between each pair of clusters to find out which kind of pattern has the highest impact on weighted tests and exams marks. We conducted a pairwise t-test with the Bonferroni p-value adjustment ($\alpha = 0.05$, Table 6 and Table 7).

### Table 2: Summary of means and standard deviations of the student’s final grades, number of attempts, and attempt pattern in MECH 221 per cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mean (SD) Grades (%)</th>
<th>Mean (SD) Number of Attempts</th>
<th>Mean (SD) of # of days that pass before x% of attempts are completed*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74 (9)</td>
<td>588 (164)</td>
<td>1.7 (0.8) 3.2 (1.1) 5.1 (1.0)</td>
</tr>
<tr>
<td>2</td>
<td>69 (6)</td>
<td>645 (169)</td>
<td>3.4 (0.7) 4.9 (0.5) 6.2 (0.4)</td>
</tr>
<tr>
<td>3</td>
<td>64 (10)</td>
<td>578 (165)</td>
<td>5.1 (0.7) 6.0 (0.5) 6.6 (0.1)</td>
</tr>
</tbody>
</table>

*all problem sets are normalized to a week in duration

### Table 3: Summary of means and standard deviations of the student’s final grades, number of attempts, and attempt pattern in MECH 222 per cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mean (SD) Grades (%)</th>
<th>Mean (SD) Number of Attempts</th>
<th>Mean (SD) of # of days that pass before x% of attempts are completed*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79 (8)</td>
<td>678 (212)</td>
<td>1.8 (0.9) 3.2 (0.9) 4.9 (1.2)</td>
</tr>
<tr>
<td>2</td>
<td>73 (9)</td>
<td>624 (162)</td>
<td>3.5 (0.7) 5.1 (0.6) 6.3 (0.4)</td>
</tr>
<tr>
<td>3</td>
<td>68 (10)</td>
<td>539 (196)</td>
<td>5.4 (0.7) 6.3 (0.5) 6.6 (0.2)</td>
</tr>
</tbody>
</table>

*all problem sets are normalized to a week in duration

### Table 4: ANOVA results for MECH 221

<table>
<thead>
<tr>
<th>DoF</th>
<th>SS</th>
<th>MS</th>
<th>F Score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>2</td>
<td>2690</td>
<td>1345</td>
<td>21.21</td>
</tr>
<tr>
<td>Residuals</td>
<td>130</td>
<td>9528</td>
<td>63</td>
<td></td>
</tr>
</tbody>
</table>

### Table 5: ANOVA results for MECH 222

<table>
<thead>
<tr>
<th>DoF</th>
<th>SS</th>
<th>MS</th>
<th>F Score</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>2</td>
<td>1766</td>
<td>883</td>
<td>11.46</td>
</tr>
<tr>
<td>Residuals</td>
<td>125</td>
<td>9628</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6: Post-hoc pairwise t-tests for MECH 221 showing comparisons between clusters. *Statistically significant difference.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1 and 2</td>
<td>0.0022*</td>
</tr>
<tr>
<td>Cluster 1 and 3</td>
<td>2.6e-06*</td>
</tr>
<tr>
<td>Cluster 2 and 3</td>
<td>0.035*</td>
</tr>
</tbody>
</table>

### Table 7: Post-hoc pairwise t-tests for MECH 222 showing comparisons between clusters. *Statistically significant difference.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1 and 2</td>
<td>0.0089*</td>
</tr>
<tr>
<td>Cluster 1 and 3</td>
<td>9.9e-05*</td>
</tr>
<tr>
<td>Cluster 2 and 3</td>
<td>0.1444</td>
</tr>
</tbody>
</table>
We observed statistically significant p-values when comparing the tests/exams marks between clusters 1 and 2, and between clusters 1 and 3. A small but statistically significant difference exists between clusters 2 and 3 for MECH 221, but not for MECH 222. These results suggest that the early submission pattern specifically, which corresponds to cluster 1, is associated with the highest tests and exams marks.

5. DISCUSSION AND CONCLUSIONS

We were able to characterize patterns of problem attempts in online homework for students in two large second-year mechanical engineering classes, and relate these patterns to individual test/exam grades. We found that students who are submitting attempts earlier are also scoring higher grades on their tests and exams.

The differences in mean grades between the clusters in both courses is striking – about 10% between the best and worst performing clusters (p < 2.6e-06 in MECH 221 and p < 9.9e-05 in MECH 222). This is a surprisingly large difference observed based on homework attempt timing. Additionally, this behavior, and corresponding level of performance in class per cluster, is similar for MECH 221 and MECH 222. However, it is unlikely that the relationship we observed is a causal one – that is, attempting online homework early probably does not directly lead to better grades. It may be possible that those that submit homework attempts earlier are already familiar with the material or have better study habits (which consequently lead to higher final grades). It is also likely that attempting homework earlier allows more opportunities to consult external resources, instructors, teaching assistants, and peers in order to further understanding of the course material.

The strengths of this study include using actual student attempt data from the WebWorK system rather than self-reports on homework system usage. We evaluated over 30 problem sets between the two courses, for almost 4100 student-problem set combinations. Both the number of clusters and the cluster patterns were determined from the data itself, rather than looking for pre-determined patterns.

Limitations of this study include that, while we were also able to show similar patterns in two courses, the majority of students in the first course also took the second course. Other limitations include not separating attempts between successful and unsuccessful attempts. Examining the success of each attempt could possibly help explain the WADF patterns, as once students solve a problem there will be no more attempts on that problem. If they solve several problems early in the week, the proportion of total attempts later in the week will be lower. However, preliminary evidence does not suggest that students in cluster 1 were solving problems with substantially fewer attempts overall (i.e. had greater command of the material at the time of completing the problem set), as the mean total number of attempts per cluster is fairly similar between the early and late submitters (Table 2, Table 3). It is also not clear if these patterns are limited to online homework, or if they might be true in PPH as well.

The learning impact of attempting homework early may be tested in future through modifications to the online homework system. For example, an OHW system could assess part of the mark for a particular homework set based on how many attempts were completed within the first three days with a correct-order-of-magnitude submitted answer; this may encourage students to meaningfully attempt the problems earlier. It may also be possible to present this evidence to students to encourage them to meaningfully alter their study habits to allow earlier work on their homework sets.

Acknowledgements

This work was completed on the traditional, ancestral, and unceded territory of the Musqueam people.

We would like to acknowledge the STAT 450/550 instructors who supervised the statistical analysis for this project during the course, and Stella Song and Qifan Tang who worked on this project as part of their coursework. We also thank the Faculty of Applied Science Centre for Instructional Support for providing facilities and support for data collection.

We gratefully acknowledge the financial support for this project provided by UBC Vancouver students via the Teaching and Learning Enhancement Fund.

References


L. Wang, “Teaching differential equations using WeBWorK as an online tool,” in *E-Learn*, 2018, pp. 1089–1092.


T. (Orchid) Huang, “The role of task-specific adapted knowledge of response feedback in algebra problem solving online homework in a college remedial course,” University of Southern California, 2008.


