The Relationship Between Student Mental Wellness and the Availability of Pre-Submission Unit Testing: An Exploratory Study

by

Braden Gibson

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Computer Science

Guelph, Ontario, Canada

© Braden Gibson, May, 2019
Abstract

THE RELATIONSHIP BETWEEN STUDENT MENTAL WELLNESS AND THE AVAILABILITY OF PRE-SUBMISSION UNIT TESTING: AN EXPLORATORY STUDY

Braden Gibson
University of Guelph, 2019
Advisor:
Dr. Judi McCuaig
Dr. Daniel Gillis

Students who are learning to program can face difficulty when encountering unknown errors. Current literature indicates that providing immediate, formative feedback can be one of the largest positive factors in reinforcing student knowledge and learning. As such, using test-driven development as a form of immediate feedback in early programming courses has been shown to have a positive impact on code quality and maintainability, while also indicating that students’ confidence and motivation may show improvements. While prior research focused primarily on measuring assignment quality and academic outcomes, our study explores the relationship between providing unit tests to students for their assignments and measures of student mental well-being, such as grit, self-efficacy, and engagement. Although the correlations between mental well-being metrics and unit test engagement metrics were generally weak, a number of individual survey questions were identified that could form the basis for a more representative survey in future research. Students were also classified by their usage types and most frequent activity periods, with differences observed that may inform further work.
Acknowledgements

I would like to express my gratitude to the following people for their support in the completion of this thesis:

Dr. Judi McCuaig
Dr. Daniel Gillis
Lynn & Kelvin Gibson
Peter Wedel
# Contents

Abstract ................................................................. ii
Acknowledgements .................................................... iii
List of Tables ........................................................ vi
List of Figures ........................................................ vii
1 Introduction .......................................................... 1

2 Literature Review .................................................... 3
  2.1 Student Achievement ............................................ 3
  2.1.1 Grit .......................................................... 5
  2.1.2 Self-Efficacy .................................................. 6
  2.1.3 Engagement ................................................... 8
  2.2 Immediate Feedback ............................................ 10
  2.3 Test-Driven Development ....................................... 12
  2.4 Test-Driven Development and Student Learning ............... 13
  2.5 Educational Data Mining ...................................... 16
  2.6 Literature Review Summary .................................... 17

3 Methodology .......................................................... 18
  3.1 Research Questions ............................................. 18
  3.2 Data ............................................................. 18
  3.3 Surveys .......................................................... 19
  3.4 Test-Driven Development ....................................... 19
  3.5 Metrics .......................................................... 21
  3.6 Data Summary ................................................... 23

4 Results ............................................................... 27
  4.1 Initial Analysis .................................................. 31
  4.2 Data Clustering and User Profiles .............................. 35
  4.3 Profile Analysis ................................................ 40
  4.4 Temporal Analysis .............................................. 47
5 Discussion

5.1 Research Questions

5.1.1 Are there any notable correlations between measurements of mental wellness and measurements of engagement with test-driven development activities?

5.1.2 Are there any notable correlations between measurements of mental wellness and measurements of success within test-driven development activities?

5.2 Future Work

5.2.1 Limitations

5.2.2 Further Research

5.3 Conclusion

Bibliography

A Grit Survey

B Self-Efficacy Survey

C Engagement Survey

D Correlational Scatter Plots

E Survey Questions by User Profile
List of Tables

3.1 Description of survey metrics. ............................ 22
3.2 Description of engagement with TDD metrics. .............. 22
3.3 Description of success with TDD metrics. .................. 23
3.4 Description of the conditions under which a given metric was missing a measurement. ......................... 24
3.5 Summary statistics of mental wellness surveys. ............... 24
3.6 Summary statistics of test-driven development activity. ....... 26

4.1 Cross tabulation of the number of Grit, Self-Efficacy, and Engagement score responses, separated by High (greater than or equal to 3) and Low (less than 3) scores. The “No Other Survey” column/row represents the number of students who completed the indicated survey and neither of the other two surveys. The cell where the “No Other Survey” column intersects with the “No Other Survey” row represents the number of students that interacted with the VPL activities but completed no surveys. ......................... 31
4.2 Student counts of the defined user profiles. .................. 39
4.3 Modes of responses to each survey question identified by the chi-squared tests as having differing distributions. ................ 47
4.4 Student counts of the defined activity periods. ............... 49
List of Figures

3.1 Sample of a VPL execution file, with a setUp method to be run before each assertion, and two sample unit tests to be run on the students’ submitted programs. Students were required to provide the Java file defining the Student class in this example. .......................... 20
3.2 Example of the VPL editor’s toolbar. ................................. 21
3.3 Sample Skill Acquisition Speed measurements. .......................... 25
4.1 Histogram of the Grit Score measurements, as a score from 1 to 5. The mean of 3.15 is represented by the solid red line and the mode of 3.35 is represented by the dashed blue line. .......................... 28
4.2 Histogram of the Self-Efficacy Score measurements, as a score from 1 to 5. The mean of 3.48 is represented by the solid red line and the mode of 3.60 is represented by the dashed blue line. .......................... 28
4.3 Histogram of the Engagement Score measurements, as a score from 1 to 5. The mean of 3.76 is represented by the solid red line and the mode of 3.72 is represented by the dashed blue line. .......................... 29
4.4 Histogram of the Total Submissions measurements. .......................... 29
4.5 Histogram of the Number of Sessions measurements. .......................... 30
4.6 Histogram of the Average Submission Percentage measurements. .......................... 30
4.7 Initial correlogram of Pearson correlations. ................................. 32
4.8 Initial correlogram of Spearman correlations. ................................. 33
4.9 Scatter plot of Total Submissions and Average Submission Percentage. .......................... 36
4.10 k-means clustering of Total Submissions and Average Submission Percentage (k = 2). Clusters are indicated by data point colour and shape. .......................... 37
4.11 DBSCAN clustering of Total Submissions and Average Submission Percentage (ε = 5 and MinPts = 3). Clusters are indicated by data point colour and shape, where black circles represent noise. .......................... 38
4.12 OPTICS clustering of Total Submissions and Average Submission Percentage (ε = 5, MinPts = 3, and εc = 4.2). Clusters are indicated by data point colour and shape, where black circles represent noise. .......................... 39
4.13 Boxplot of Grit Scores, separated by user profile. .......................... 40
4.14 Boxplot of Self-Efficacy Scores, separated by user profile. .......................... 41
4.15 Boxplot of Engagement Scores, separated by user profile. .......................... 41
4.16 Boxplot of responses to “Setbacks don’t discourage me. I don’t give up easily,” separated by user profile. 42
4.17 Boxplot of responses to “It is easy for me to stick to my aims and accomplish my goals,” separated by user profile. 43
4.18 Boxplot of responses to “Thinking about the course between classes,” separated by user profile. 44
4.19 Proportional bar plots of Likert scale responses to individual survey questions, separated by user profile. 46
4.20 Distribution of VPL Assignment 1 activity submissions by submission time. Each bin spans a 6-hour period, and the vertical red line indicates the Assignment 1 due date of Oct. 25, 2018 at 8:30am. 48
4.21 Distribution of VPL Assignment 2 activity submissions by submission time. Each bin spans a 6-hour period, and the vertical red line indicates the Assignment 2 due date of Nov. 26, 2018 at 8:30am. 49
4.22 Boxplot of Grit Scores, separated by activity period. 50
4.23 Boxplot of Self-Efficacy Scores, separated by activity period. 51
4.24 Boxplot of Total Number of Sessions, separated by activity period. 51
4.25 Boxplot of Average Session Length, separated by activity period. 52
4.26 Boxplot of Total Submissions, separated by activity period. 53
4.27 Bar chart of activity period students, separated by user profile. 53

D.1 Initial scatter plots of Pearson correlations. 73
D.2 Initial scatter plots of Spearman correlations. 74
Chapter 1

Introduction

An appreciable difficulty that students learning to program encounter is uncertainty about how to proceed when they face an unknown error or piece of functionality. When a student encounters an unexpected compilation or runtime error for the first time, they may find it difficult to know what steps to take to resolve the issue without input from an instructor or peer. As such, feedback is a crucial element in the student learning process. Unfortunately, with the high student-to-instructor ratio in post-secondary education it can be a challenge for instructors and teaching assistants to provide enough meaningful feedback to computing students. A solution to this issue that has been gaining traction in recent years is the use of computer applications to provide immediate, formative feedback to students on their work.

Another problem in computer science education is introducing software testing as an assessment tool, in which the accuracy of a program is verified. Unit testing is one of the core levels of software testing, and provides the most granular information on the tested program code. In order to encourage students to produce high-quality code, a proposed solution has been to introduce test-driven development at an introductory level. Test-driven development has been shown in industry to have a positive impact on code quality (Bhadauria, 2009; Dogša & Batič, 2011; Nanthaamornphong & Carver, 2017; Xu & Li, 2009; Yenduri & Perkins, 2006), code maintainability (Dogša & Batič, 2011; Yenduri & Perkins, 2006), and specification definitions (Jones, 2006; Paul, 2016). These benefits have also been noted when test-driven development has been included in introductory computing curricula (Buffardi & Edwards, 2012; Desai et al., 2009; Edwards, 2004; Fidge et al., 2013; Janzen & Saiedian, 2008; Kauffman & Janzen, 2003; Proulx, 2009; S. M. Rahman, 2007; Yenduri & Perkins, 2006).

An important aspect of student learning and experience is student mental wellness. It is essential for instructors seeking to keep their students engaged with the course material to ensure that the students’ levels of motivation and self-confidence are elevated and maintained. Current literature indicates that
students who are more engaged and have higher levels of self-esteem and self-efficacy are more likely to succeed academically (Lee & Reeve, 2012; Multon & Brown, 1991; Pintrich & De Groot, 1990; Skinner et al., 2017).

This thesis examines the intersection of test-driven development in introductory computing courses and student mental wellness, a combination that is relatively unexplored. Current research into test-driven development in higher education has focused more on academic performance and code quality outcomes. This exploratory study takes a preliminary look at potential relationships with the third major factor in learning, student mental wellness. To delve into this topic, data produced during a second-year computing course was analyzed and investigated.

**Thesis statement:** We believe that student use of test-driven development as a method of assignment production is associated with higher levels of measurements of mental wellness.

To evaluate the thesis, it is important to consider the metrics used to indicate levels of test-driven development usage and student mental wellness. The manner in which these metrics are measured will be a key element in this investigation, and, as such, will have a substantial focus in this thesis.

In order to assess the thesis, the following research questions are discussed and evaluated:

1. Are there any notable correlations between measurements of mental wellness and measurements of engagement with test-driven development activities?
2. Are there any notable correlations between measurements of mental wellness and measurements of success within test-driven development activities?

Chapter two presents a review of literature to situate this study in the current field. Metrics of student motivation and engagement are discussed, along with the rationale behind combining test-driven development with student learning and achievement. This chapter concludes with a review of the goals and stakeholders of educational data mining. Chapter three describes the learning environment which produced the data being examined and establishes the goals and metrics to be used. This chapter concludes with a summary of the measurements collected. Chapter four explores the relationships between metrics, and features an evaluation of the research questions in light of the data. This chapter also proposes potential areas of future research, and experiment design for subsequent studies. Finally, chapter five concludes by discussing any promising relationships for future analysis.
Chapter 2

Literature Review

This exploratory research study examines the intersection of test-driven development with student achievement and mental wellness. This chapter introduces a review of relevant literature to situate the study in the current literature, with regards to the best practices and accepted findings.

First, student learning and achievement are examined along with current, established metrics (Section 2.1). This review then introduces immediate feedback (Section 2.2), and explains its relevance to test-driven development (Section 2.3). Next, current research regarding the connection between test-driven development and student learning is explored, establishing which intersections of metrics have been investigated and which have not (Section 2.4). Finally, this chapter covers educational data mining as a suitable analysis method (Section 2.5).

2.1 Student Achievement

Student learning is a field with many different goals and opportunities for investigation. This includes, but is not limited to, answering questions such as:

- How do students learn?
- How can instructors assess how well students are learning both in and out of the classroom?
- How can instructors motivate students?
- How do students develop their skills?

By measuring and quantifying student learning, researchers can evaluate different pedagogical methods to determine which ones reach students the best, or which ones reach the greatest number of students. Evaluated pedagogical methods can be broad (such as discussions, presentations, modelling, etc.) or specific (such as the impact of deadlines earlier versus later in the week). Continuous implementation and evaluation of pedagogical methods allows for greater improvement in the manner in which students learn.
One of the more difficult aspects of evaluating different pedagogical methods is determining an effective metric. Various studies have used both qualitative and quantitative metric to see if there is an improvement in student learning.

A common metric of student achievement is grades, such as assignment grades throughout the course (Desai et al., 2009; Handelsman et al., 2005; Janzen & Saiedian, 2008), midterm and/or final exam grades (Chingos, 2016; Handelsman et al., 2005; Wei et al., 2015), or the students’ overall grades in the course (de Kleijn et al., 2012; Sancho-Vinuesa et al., 2013, 2017). In some cases grades are used in conjunction with other metrics (de Kleijn et al., 2012; Desai et al., 2009; Janzen & Saiedian, 2008; Sancho-Vinuesa et al., 2013; Wei et al., 2015), and in other studies the grades form the entire basis for analysis (Chingos, 2016; Handelsman et al., 2005; Sancho-Vinuesa et al., 2017). Often, grades are used in these circumstances as a dependent variable to evaluate the effectiveness of a pedagogical method. For example, Chingos (2016) used a variety of grade metrics and indicators, such as whether each student took the final exam, whether the student passed with a C+ or higher, and the students’ final exam scores, to represent student learning in order to examine its relationship with instructional quality. Additionally, Handelsman et al. (2005) used student grades on homework assignments, midterms, and final examinations to examine the validity of their student engagement factor. There are benefits to using grades as a metric, since they can be easier to collect than other metric and, as noted by Chingos (2016), grades based on right/wrong questions should be fairly consistent across different evaluators. However, there are a number of concerns with using grades as an analogue for student learning, especially when used as the sole factor. Beleche et al. (2012) asserted that since final grades can be a combination of multiple factors, such as initial student ability, evaluator grading preferences, or class times, they may not accurately represent learning. This can make it difficult to draw conclusions about the effects of the pedagogical method in question when using grades to represent student learning and achievement.

Another metric that can be used in evaluating student learning methods is pre-method and post-method tests (Melguizo & Wainer, 2015). This method allows researchers to establish a baseline at the beginning of an evaluation period and compare it to the students’ scores at the end to identify the difference as student knowledge gains. This can be more difficult to obtain than other metrics since it requires more of a framework to be in place. Melguizo & Wainer (2015) were able to obtain this information in their study on student learning outcomes because they used data from the Exame Nacional de Desempenho dos Estudantes, an exit exam that Brazil has also been administering to incoming college students. This provided them with two data points to estimate student learning gains with, giving them solid quantitative measurements with which to compare pedagogical methods. However, this metric has similar downsides to using grades and, indeed, most metrics of student learning. Due to the fact that
the metrics are obtained relatively far apart from one another, other factors can influence the results of the post-method test such as student interests and preferences (Beleche et al., 2012), or students’ life events. This means that it can be challenging to ascertain what impact the pedagogical method had on the results versus other contributing factors.

An additional relevant factor in students’ grades or pre-method and post-method testing results can be the students’ mental health. Current literature indicates that poor mental health is correlated with lower grades and reduced educational achievement (Bania et al., 2016; Brännlund et al., 2017). As such, it is worth considering the impact on or from mental health when reviewing studies that use grades or educational attainment as a metric.

Research into student learning can also use qualitative metrics (such as grit, self-efficacy, and engagement) to assess learning methods. The remainder of this section explores the various approaches utilizing these metrics.

2.1.1 Grit

Grit is a metric first proposed in a student learning context by Duckworth et al. (2007), initially defined as “perseverance and passion for long-term goals”. Grit can be described as a combination of metrics of resilience, conscientiousness, self-control, and perseverance (Bashant, 2014). In their preliminary research Duckworth et al. (2007) reasoned that a gritty individual, one who relentlessly persists in the face of adversity and plateaus, and sustains effort levels over long periods, is more likely to achieve highly than another individual of the same intelligence but less grit. Based on their initial ideas about grit, a 12-item Grit Scale was developed for participants to self-report their agreement with statements such as I often set a goal but later choose to pursue a different one and setbacks don’t discourage me using a 5-point Likert scale. In six studies that were conducted, they found that grit accounted for differences in achievement levels beyond the factor of intelligence.

The original 12-item Grit Scale (Grit-O) was later updated by Duckworth & Quinn (2009) into a shorter self-reported survey (Grit-S). The predictive validity of each item in Grit-O was considered to create a subset for Grit-S. Like the analysis in the original paper, the researchers conducted six studies on Grit-S, allowing them to control for age and Big Five personality traits such as conscientiousness and extraversion. They determined that their 8-item Grit Scale was psychometrically stronger than Grit-O and fit the data more accurately.

Since its introduction, grit has gained recognition and popularity as a psychological construct, and many researchers have incorporated it into their studies. Cosgrove et al. (2018) used grit as a factor alongside school absences to explore the relationship with academic outcomes, controlling for physical
fitness. They determined that grit and absences are compelling contributors to academic success, measured by scores on a year-end state subject test. Additionally, Hodge et al. (2018) focused on strengthening the evidence of a relationship between grit and academic outcomes. They further explained the construct of grit as “a student’s ability to effortfully persist in the face of struggle”. Using Grit-S, they explored the relationship between grit, engagement (measured by a modified Utrecht work engagement scale for schools), and productivity (measured by a subset of the jobs demands-resources scale) using approximately 400 Australian university students as participants. They determined that although grit does have an impact on engagement and productivity, the relationship can be best described as grit leading to engagement, and then engagement leading to productivity.

Although grit has been widely accepted as a legitimate predictor of achievement, there are researchers who have concerns and some literature that does not support its validity. Recently, a critical review by Credé (2018) took issue with the current state of grit as a construct, noting that there is no supported argument for combining perseverance and passion into one metric. Credé also suggested that while grit or a subset of its components may be necessary for achievement, they are not necessarily sufficient. That is to say, there may be other factors required by success that are not incorporated into the grit construct. In addition, there are studies that did not find a statistically significant relationship between grit and academic outcomes (Bazelaia et al., 2016; Ivcevic & Brackett, 2014; Stewart, 2015; West et al., 2016). However, one study controlled for previous academic achievement (Bazelaia et al., 2016), which may have been a result of grit, and another study controlled for Big Five personality traits (Ivcevic & Brackett, 2014), of which grit incorporates several factors. When presented in light of the abundance of literature supporting its validity (Bowman et al., 2015; Cosgrove et al., 2018; Cross, 2014; Hodge et al., 2018; Ryan, 2015; Strayhorn, 2014; Wolf & Jia, 2015), it is reasonable to conclude that grit is worth considering as a predictor of academic performance.

2.1.2 Self-Efficacy

Self-efficacy is a concept first introduced by Bandura (1977) and refers to an individual’s belief in their own ability to achieve goals and complete tasks. Note that this is distinct from one’s ability to accomplish goals, but is rather whether one believes themself capable of planning and carrying out the intended task. Bandura (1997) further explained the goal completion process as belief-action-achievement, wherein an individual must believe in their ability to produce desired outcomes before they are incentivized to act. Self-efficacy plays the role in moving from belief to action, and then ability of execution allows the individual to achieve goals. If one is lacking in self-efficacy and does not take the step to action, there is no possibility of the realization of desired goals. When defined formally, there are three distinct dimensions of self-efficacy: level, generality, and strength (Zimmerman, 2000). The level of self-efficacy measures
how dependent it is on the difficulty of the desired outcome, its generality refers to how well it translates to tasks in differing areas, and the strength indicates how confident the individual is in their ability to complete the specified goal.

In the three decades after its introduction by Bandura (1977), self-efficacy became recognized as an important factor in accomplishing goals, especially in a learning context (Linnenbrick & Pintrich, 2003; Mone, 1994; Multon & Brown, 1991; Nease et al., 1999). In a meta-analysis on the impact of self-efficacy on academic performance and persistence, Multon & Brown (1991) examined literature from 1977 to 1988, obtaining 36 relevant studies for performance and 18 studies for persistence. With participants ranging from elementary to college students, they found a statistically significant relationship between self-efficacy and academic performance, accounting for around 14% of performance variance between students. They also discovered that there was a statistically significant relation between self-efficacy and persistence, explaining approximately 12% of the variance between students. Mone (1994) additionally noted that two types of self-efficacy can be considered: process and outcome. Process self-efficacy refers to an individual’s confidence in their subskills required to complete the task at hand, whereas outcome self-efficacy is their belief in their ability to achieve the goal to a satisfactory level. For example, in a learning context, outcome self-efficacy may be a student’s confidence that they can reach a desired level of academic performance, and their process self-efficacy may be their judgment of their required skills such as studying, note-taking, or attending lectures. After a study on 252 university students, Mone concluded that outcome self-efficacy led to more valid results when predicting personal goals and performance, although process self-efficacy significantly affects outcome self-efficacy. Furthermore, Nease et al. (1999) determined that individuals with high self-efficacy were more likely to accept positive feedback as accurate when compared to those with lower self-efficacy, who viewed it as less valid over time. In 2003, Linnenbrick & Pintrich conducted a review of literature with regards to self-efficacy and its effect on motivation and engagement. After examining research on self-efficacy and behavioural, cognitive, and motivational engagement, they concluded that self-efficacy is positively correlated with positive physical and emotional reactions, as well as with motivational beliefs such as value and utility.

A widely-adopted self-efficacy metric was first proposed and validated by Sherer et al. (1982). Their initial scale, which had 36 items, was given to 376 university student participants along with personality metrics. On the completion of a factor analysis, the scale was trimmed down by removing insignificant measures and split into two sections: a general self-efficacy subscale consisting of 17 items, and a social self-efficacy subscale containing 6 items. The general subscale contains statements such as when I set important goals for myself, I rarely achieve them and I feel insecure about my ability to do things, which participants respond to using a 14-point Likert scale. In 1995, Schwarzer & Jerusalem introduced a metric named the Generalized Self-Efficacy Scale (GSES), which assesses an individual’s self-efficacy in a self-administered test that takes 2-3
minutes to fill out. The GSES contains just 10 items, differing from the general subscale developed by Sherer et al. (1982), which contains 17 items. The GSES includes items like *I can always manage to solve difficult problems if I try hard enough* and *it is easy for me to stick to my aims and accomplish my goals* that are assessed on a 4-point Likert scale. Schwarzer & Jerusalem found high internal consistency within five studies, and established concurrent and predictive validity in German studies. Since its proposal in 1995, the GSES has been validated and modified by further literature (Croasmum & Ostrom, 2011; Lown, 2011). Notably, Croasmun & Ostrom (2011) expanded the Likert scale to a 5-point version and added 5 randomly-selected existing questions reworded negatively. They concluded that their variant had strong internal consistency, with Cronbach’s alpha values matching or exceeding the original GSES, indicating that these may be changes worth considering.

### 2.1.3 Engagement

When discussing engagement in a learning setting, it is important to consider the motivation and attention behind students’ actions. Skinner et al. (2009) defined actions as the combination of behaviour, attention, emotion, and goals, meaning that actions are goal-oriented behaviours. Thus, two instances of the same behaviour may be considered different actions if the intention and focus behind them differs, just as two examples of different behaviours may be considered to be the same action if they have the same goal. Skinner et al. thus concluded that motivational constructs of engagement should incorporate metrics of behaviour and goal-orientation. Lee & Reeve (2012) further characterized engagement as the degree to which a student is actively involved in the learning process. Engagement is widely considered to be a multidimensional construct, consisting primarily of behavioural, emotional, and cognitive engagement. Behavioural engagement is students’ attentive and sustained involvement in learning activities in a positive manner, emotional engagement is considered to be students’ positive reactions to learning material, such as interest and enthusiasm, and cognitive engagement refers to the extent to which students use cognitive resources such as planning and revision during learning activity (Christenson et al., 2012; Lee & Reeve, 2012; Pintrich & De Groot, 1990; Skinner & Belmont, 1993; Skinner et al., 2009).

Engagement’s impact on how students interact with their learning environment, build academic and social skills, and challenge themselves to improve cannot be understated. Research has indicated that students with high levels of engagement are more likely to take initiative and persevere on difficult tasks, and show interest and enthusiasm for enrichment activities (Pintrich & De Groot, 1990; Skinner et al., 1990, 2017). In 1990, Pintrich & De Groot conducted a study on 173 seventh graders on the relationship between motivation and engagement, self-regulated learning, and academic achievement. Their analysis indicated that self-efficacy and self-regulation were the best predictors of academic outcome, measured by a variety of grades, reports, and exams. However,
since self-efficacy was not significantly correlated with academic achievement when cognitive engagement was included in the regression analysis, Pintrich & De Groot concluded that self-efficacy plays a role in cognitive engagement, and it is cognitive engagement that is more closely linked with performance. Additionally, Skinner & Belmont (1993) determined that there was a reciprocal relationship between dimensions of teacher behaviour and student motivation in 144 children in Grades 3-5. Students that displayed higher levels of behavioural engagement were found to be more likely to receive teacher interactions such as involvement, structure, and autonomy support when compared to their more disengaged peers. Lee & Reeve (2012) indicated in their research that teachers' ratings of students' behavioural, cognitive, and emotional engagement were predicted by the students' academic achievement. They also noted that the teachers' ratings of the students' behavioural and cognitive engagement were predicted by the students' respective self-reported ratings, but that the teachers' ratings of emotional engagement were not predicted by self-reported ratings. Finally, Skinner et al. (2017) conducted a study on undergraduates in biology, physics, and chemistry to validate the construction of a survey measuring the motivational factors in science education. In their analysis they found a correlation between final course grades and both behavioural and emotional engagement, as well as with both behavioural and emotional disengagement (also known as disaffection). Since this correlation held across both longitudinal data points they collected, Skinner et al. concluded that this is consistent with the concept that higher levels of engagement may be conducive to improved learning and academic performance. Due to its correlation with academic outcomes and skill development, engagement can be effective in measuring the impact of changes in a learning environment.

As part of their study in 1990, Pintrich & De Groot developed a survey titled the Motivated Strategies for Learning Questionnaire (MSLQ), which is designed to capture the motivational beliefs and self-regulated learning strategies of college students. A 56-item survey using a 7-point Likert scale, the MSLQ contains items related to cognitive engagement, and has been used as a basis of measure in subsequent literature (Hsieh et al., 2012; Rodger et al., 2007). The National Survey of Student Engagement (NSSE) can also be used as a metric of university student engagement, and is designed for use by professors and administrators. A study conducted by Kuh et al. (2008) used the NSSE to measure students' engagement in educationally relevant activities, and concluded that student engagement has a correlation with first-year university academic outcomes and a statistically significant positive impact on persistence into the second-year. More recently, Handelsman et al. (2005) proposed a metric of university course engagement, named the Student Course Engagement Questionnaire (SCEQ). The SCEQ consists of 27 items assessed using a 5-point Likert scale, and was validated by Handelsman et al. in two separate studies on undergraduate students. The researchers discovered positive correlations between results on the SCEQ and global engagement, goal orientation, and academic performance. Since its introduction, the SCEQ and its modified versions have been used as
an effective metric by recent educational literature (S. J. Brown et al., 2017; R. A. Rahman et al., 2018; Young & Legister, 2018).

In the preceding sections, this investigation examined several relevant factors in measuring and predicting student learning and achievement, namely grit, self-efficacy, and engagement. Literature indicated that all three factors are strongly correlated with academic success. In the following section, this thesis explores immediate feedback and its effects on the learning process.

2.2 Immediate Feedback

In educational environments, feedback is considered to be a crucial factor in students’ acquisition of knowledge and skills (Hattie & Timperley, 2007; Henderson & Broadbridge, 2007; Hernández, 2012; Hettiarachchi et al., 2014; Holmes, 2015; Joseph & Maguire, 1982; Llorens et al., 2014; Rodríguez et al., 2012; Sancho-Vinuesa et al., 2013, 2017; Shute, 2008). Feedback, in this context, is information that is provided to a learner with regards to their performance or knowledge (Hattie & Timperley, 2007). Hattie & Timperley also asserted that feedback without context is useless, that it must be applied in response to a student’s actions to have benefits. In other words, the feedback provided to students should be a judgment of their work, and should help them to reduce the gap between their current and required level of understanding (S. Brown, 1999). Feedback can generally be described on two axes: formative or evaluative, and immediate or delayed (Hattie & Timperley, 2007; Hettiarachchi et al., 2014; Hernández, 2012; Joseph & Maguire, 1982; Rodgers, 2006; Sancho-Vinuesa et al., 2013; Taras, 2008).

Evaluative (or summative) feedback is provided to learners with the intention of evaluating and reporting on their abilities, often in the form of grades or marks. This is often a required form of feedback in coursework, due to the need to determine if the student has adequately learned the material (Hernández, 2012; Taras, 2005, 2008). In contrast, formative (or descriptive) feedback is relayed to students to provide them with a framework to review and evaluate their mistakes, consider what they have done correctly, and improve their knowledge acquisition (Holmes, 2015; Rodgers, 2006; Sancho-Vinuesa et al., 2017; Taras, 2005, 2008). Evaluative feedback be beneficial on its own (Hernández, 2012) or when combined with formative feedback (Hernández, 2012; Hettiarachchi et al., 2014; Holmes, 2015; Rodríguez et al., 2012). However, some literature suggests that combining formative and evaluative feedback may result in students noting the grade and moving on, or in the formative feedback serving simply to justify the evaluative feedback, thus transferring the focus of the feedback away from methods in which students can improve (E. Brown & Glover, 2006; Sadler, 1989; Yorke, 2007). Taras (2005) asserted that all feedback is evaluative, and that the inclusion of additional information such as how to fix mistakes and improve the submitted work is what makes an assessment formative in addition
to evaluative. In any case, research indicates that while evaluative feedback can add value to the learning process, formative feedback provides learners with the greatest benefits in terms of addressing mistakes and improving understanding (Hattie, 1999; Hattie & Timperley, 2007; Hernández, 2012; Hettiarachchi et al., 2014; Shute, 2008). For example, Hettiarachchi et al. (2014) conducted a study wherein they changed the feedback system of an online university course from evaluative (in the form of multiple choice questions) to formative e-assessment. They found that students’ participation levels increased, and their performance in terms of grades had also improved, indicating that formative feedback may have more power. In 1999, Hattie reported a synthesis of over 500 meta-analyses, which combined data from 20-30 million students. This research concluded that studies with greater effects on student performance provided students with more descriptive feedback about their tasks. Hattie’s synthesis also indicated that feedback is most effective when it provides information and reinforcement to students, and when it is goal-oriented, signifying that formative feedback is more effective than evaluative. Finally, Shute’s review of literature in 2008 concluded that formative feedback should be given in response to the actions of a learner, and should be intended to shape the perception and actions of the student, as well as improve learning and performance.

Immediate feedback is provided to students right after they complete a task, with the intention of promptly correcting any mistakes and avoiding the reinforcement of errors. On the other hand, delayed feedback is given to students after a period of time has elapsed, usually at least 24 hours (Joseph & Maguire, 1982). There is older literature that indicated that delayed feedback is at least as effective as immediate feedback in information retention, and in some cases, more effective (Kulhavy & Anderson, 1972; Sassenrath, 1975; Surber & Anderson, 1975). However, some of these defined immediate feedback as feedback that was issued between questions, and other studies defined immediate feedback as something issued after a completed assignment, which may account for the discrepancies (Johnston, 2015; Joseph & Maguire, 1982). More recent literature concludes that immediate feedback is superior to delayed feedback with regards to student learning and retention, as indicated by improvements in academic performance (Brosvic et al., 2005; DiBattista et al., 2009; Hattie & Timperley, 2007; Kelly et al., 2013; Mendicino et al., 2009; Singh et al., 2011). For example, Brosvic et al. (2005) explored the impact of immediate corrective feedback on academic outcome using 110 undergraduate liberal arts and sciences students. They found that the group of students who received no feedback did the poorest on the final exam when compared to students that received immediate or delayed feedback, and that the students who received immediate feedback achieved the highest on the exam. Likewise, Singh et al. (2011) conducted two studies in eighth grade classrooms, with both studies reaching the conclusion that immediate feedback with tutoring is a significant factor in student knowledge acquisition. Additionally, immediate feedback has been shown to have a positive influence student engagement, and thus on academic performance (Sancho-Vinuesa et al., 2013). In a study on 206 students
in a math course at the Open University of Catalonia, Sancho-Vinuesa et al. (2013) determined that a new strategy based on immediate, automated feedback had a significant impact in reducing the number of dropouts and improving student achievement. With the evidence from the latest literature, this review can conclude that immediate feedback is an important factor in promoting student engagement, confidence, and academic performance.

In this section this thesis explored the role of immediate, formative feedback in educational environments, specifically noting its positive impact on student engagement and academic performance. In the subsequent sections, this thesis reviews test-driven development as a software design process, and considers its effects when integrated with student learning.

2.3 Test-Driven Development

Test-driven development (TDD), also known as test-first development, is a software design technique that directs development focus through testing. In the TDD process, unit tests are written before program code, as opposed to the more conventional method of writing code first and then writing tests to verify its functionality, also referred to as test-last development (K. Beck, 2003; Fucci & Turhan, 2014). TDD can be summarized as an iterative process of the following:

1. Write a unit test for an undeveloped piece of functionality.
2. Run all unit tests. The most recent test should fail.
3. Write the minimum amount of code required to pass the new test.
4. Run all unit tests. The most recent test should pass.
5. Refactor the code to optimize performance and improve quality.

The focus on code quality and optimization by refactoring, along with the resulting suite of applicable tests, leads to the assertion that TDD can produce higher quality code than test-last development (Astels, 2003).

As with any development paradigm, test-driven development is not perfect, and as such there may be drawbacks to its use. Due to the fact that there is more overhead associated with utilizing TDD than with test-last development, some implementations may observe reduced productivity (Dogša & Batić, 2011; Nanthamornphong & Carver, 2017). Additionally, Romano et al. (2017) found in their multi-method study that developers tended to avoid refactoring as frequently as prescribed by the TDD process. There may also be issues with test coverage because a developer writing a test can often be the same one writing the program functionality (Astels, 2003). This means that an error in the code may not be caught if the corresponding test has the same oversight.

Current literature indicates that test-driven development has several associated benefits to the development process. Contrary to previously cited research,
several studies have examined TDD and found no statistical difference in productivity when compared to test-last development (Madeyski, 2009; Pančur & Ciglarić, 2011), and in some cases have noted improved productivity (Bhadauria, 2009; Erdogmus et al., 2005; Fucci & Turhan, 2014; Xu & Li, 2009; Yenduri & Perkins, 2006). In line with the proposal by Astels (2003), the literature also indicates that TDD can have a positive impact on code quality (Bhadauria, 2009; Dogša & Batič, 2011; Nanthaamornphong & Carver, 2017; Xu & Li, 2009; Yenduri & Perkins, 2006). In 2011, Dogša & Batič conducted a multi-case study on three medium-sized project in an industrial setting, wherein two projects did not utilize TDD while the third integrated TDD into its development process. The researchers determined that the developers using TDD produced higher quality code, measuring external code quality by the number of failures or faults per new line of source code. Another potential benefit of test-driven development is that the production of unit tests allows developers to define specifications in an unambiguous manner. This ensures that developers design the tests for the specification, rather than tailoring the tests to the developed program, which can sometimes occur in test-last development (Jones, 2006; Paul, 2016). Finally, TDD has been observed to have a positive effect on code maintainability, requiring less effort expended per change request than test-last development (Dogša & Batič, 2011; Yenduri & Perkins, 2006).

Test-driven development is a form of immediate feedback (Brian et al., 2015; Paul, 2016; Whalley & Philpott, 2011). If the unit test results are returned right away to a learner with the intention of helping the learner to correct mistakes and avoid reinforcing errors, then testing has served the purpose of immediate feedback. Depending on what type of information is returned, this immediate feedback may be classified as formative or evaluative. For example, if a test suite only returns how many tests passed, it may be considered evaluative feedback, whereas if the suite also includes where and what the issues were, it may be an example of formative feedback.

2.4 Test-Driven Development and Student Learning

Given that test-driven development is a type of immediate feedback, instructors may expect to see some of the benefits of immediate feedback by integrating TDD into the student learning process. Many researchers have noted this connection, and their studies have indicated a number of potential benefits that instructors might see if TDD is integrated with curricula.

First, several studies have observed improvements in development measurements of work produced by students using TDD. Coursework developed by test-driven development has been observed to have a negligible or positive impact on productivity when compared to test-last development or coursework with no tests at all. Students’ productivity can be improved with TDD either by reducing the number of hours required to complete a task (Desai et al., 2009;
Yenduri & Perkins, 2006) or by increasing the amount of code developed in the same time period (Erdogmus et al., 2005; Janzen & Saiedian, 2008; Kauffman & Janzen, 2003). In 2005, Melnik & Maurer conducted a survey with 240 participants ranging from college-level diploma to university graduate students, finding that 78% of respondents believed that test-driven development improved their productivity. Similarly, TDD has been shown to have the potential to improve the quality of students’ produced code. This improvement has been measured by an increase in test coverage (Buffardi & Edwards, 2012; Desai et al., 2009; Proulx, 2009), improved code readability (Proulx, 2009), a reduction in program faults (Edwards, 2004; S. M. Rahman, 2007; Yenduri & Perkins, 2006), and a decrease in cyclomatic complexity (Kauffman & Janzen, 2003). In their 2005 survey, Melnik & Maurer also found that 76% of respondents thought that test-driven development led to an improvement in code quality. These benefits regarding program quality can also lead to an increase in academic achievement, usually with respect to the associated assignment, but also in terms of quiz, midterm, and exam grades (Edwards, 2004; Fidge et al., 2013; Janzen & Saiedian, 2008; Kauffman & Janzen, 2003).

Research has also indicated that TDD may have a number of benefits to students regarding knowledge acquisition and retention. When implementing a program using TDD, students may find that the creation of tests allows them to unambiguously formalize assignment specifications. By focusing solely on the project requirements, students are able to break down a daunting task into small, clear objectives. This forces learners to consider their code interactions before beginning, improving their program design and specification understanding (Erdogmus et al., 2005; Fidge et al., 2013; Fridge & Bagui, 2016; Marrero & Settle, 2005; Müller & Hagner, 2002; Paul, 2016). Additionally, research has indicated that TDD may also improve knowledge reinforcement, specifically with regards to improved preparation (Proulx, 2009), faster learning (Erdogmus et al., 2005), and better understanding of course material (Desai et al., 2008; Janzen & Saiedian, 2006; Proulx, 2009). Since students are provided with timely, formative assessments in response to their actions in the TDD process, it follows that students may see some of the benefits of feedback such as improved knowledge acquisition, understanding of material, and academic achievement. As an extension of this, Proulx (2009) observed that their students using TDD curriculum were securing significantly better co-op positions, with employers citing a noticeable improvement in their programming skills.

An appreciable trend reported in literature regarding TDD integration in academia is a reluctance by students to adopt the new development process. These findings appear to hold whether or not the students were required to take part in TDD, i.e., whether testing was required for grades (Braught & Midkiff, 2016; Buffardi & Edwards, 2012; Denny et al., 2011; Fidge et al., 2013; Janzen & Saiedian, 2006, 2008) or if participation was optional (Barriocanal et al., 2002). Buffardi & Edwards (2012) put forth the idea that the behaviour of TDD is considerably different than test-last development, wherein programming the
solution is started immediately, which may account for the adoption resistance of students with prior programming experience. Additionally, Janzen & Saiedian (2006) found that more experienced programmers were less likely to adopt a TDD process than those with less experience. They concluded that younger students may have less resistance to new ideas, and thus stressed the importance of TDD’s implementation in introductory courses. Moreover, several studies conducted found that in their initial introduction to TDD, programmers tended to write unit test suites with poor code coverage, although they were observed to still have better coverage than test-last students (Brian et al., 2015; Buffardi & Edwards, 2012; Fidge et al., 2013). In 2001, Müller & Tichy concluded that due to time considerations, instructors should consider providing students with a skeleton program to begin development with. Further research found some success in introducing students to TDD by providing them with instructor-created unit tests for an initial assignment, and then requiring them to write their own tests for subsequent assignments (Brian et al., 2015; Desai et al., 2009). Paul (2016) outlined a pedagogical process in which students are provided with instructor-generated unit tests as an introductory step in acclimatizing them to TDD. Later, students would modify test cases, then add their own tests to an existing suite, and finally, write their own test suites. Researchers have also observed benefits of TDD when providing students with instructor-generated tests for the entirety of an introductory course (Melnik & Maurer, 2005; Spacco et al., 2006). Notably, Whalley & Philpott (2011) investigated the advantages of TDD, specifically focused on supplying students with unit tests rather than having them write their own. They reasoned that until students have a basic programming ability they are neither ready for testing nor able to see its value. As such, Whalley & Philpott wanted students to view the unit tests as a tool to use, in order to understand the benefits of testing and to acclimatize them with reading tests. They also believed that providing tests would reduce the cognitive load on the introductory learners. It was determined that although there was an increase in course material preparation time, it was well worth the extra effort, since many of the potential benefits of TDD were realized and students were provided with valuable immediate feedback.

Current literature has also indicated that integrating TDD into computer science curriculum may improve students’ confidence in their programming ability and motivation to begin and complete related tasks. Students introduced to TDD expressed an increased confidence in the correctness of their code and in their ability to make modifications (Barriocanal et al., 2002; Desai et al., 2008; Edwards, 2004; Kauffman & Janzen, 2003; Müller & Tichy, 2001; Whalley & Philpott, 2011). For example, Kauffman & Janzen (2003) found that on a 5-point scale with 5 being the most confident, the TDD group of students reported an average of 4.75 with regards to their confidence in their programs’ functionality, compared to the test-last group’s average of 2.5. Studies have also suggested that test-driven development may result in increased engagement with course material (Whalley & Philpott, 2011) and improved motivation to begin projects early (Buffardi & Edwards, 2012; Spacco et al., 2006).
In the two previous sections this review investigated the benefits and drawbacks of TDD, finding that it may positively impact code quality, productivity, and code maintainability, while possibly negatively influencing refactoring or test coverage. This paper also noted that these benefits appear to be conferred on introductory programming students using TDD, improving code quality, productivity, and knowledge acquisition and retention. Another noteworthy benefit of TDD may be an increase in students’ confidence and motivation, which is particularly relevant to this study. In the following section, this review explores Educational Data Mining in order to identify the goals, tasks, and stakeholders of this investigation.

2.5 Educational Data Mining

Educational Data Mining (EDM) is a subset of statistical modelling that focuses on data obtained from an educational setting, which is then used to improve student learning and pedagogical processes (Baker & Yacef, 2009; Siemens & Baker, 2012). According to Romero & Ventura (2013), EDM is an intersection of three primary disciplines: education, computer science, and statistics. As such, EDM also incorporates features of permutations of its primary disciplines, such as computer-based education, data mining and machine learning, and learning analytics.

In their 2009 review, Baker & Yacef identified four main applications of EDM methodology, including:

- Improving student models, which consolidate student information such as motivation, knowledge, and attitudes towards learning. These models can allow learning software to customize their content to individual students to increase student learning.
- Identifying and exploring domain models, which can describe the interrelationships and structure of knowledge in a domain.
- Studying pedagogical processes for the purpose of determining which are most appropriate in a variety of learning environments.
- Using experimental data to improve pedagogical methods, which can be used to create more effective processes and systems.

These applications of EDM can also be explained by exploring the goals of potential stakeholders, of which the two main groups are learners and educators. Learners can benefit from receiving adaptive and personalized feedback, which may improve student learning, whereas educators may better reflect on their students’ learning styles and thus improve their pedagogical methods. Additional potential stakeholders in EDM research are researchers, who can recommend various data mining methods, and administrators, who benefit from improving their educational resources (Romero & Ventura, 2010, 2013).

Taking into account the applications and stakeholders of EDM research, a
number of generalized steps have been proposed. First, the data obtained from a learning management system (LMS) is analyzed and potential correlations between variables are identified. This can be achieved with methods such as classification, regression, clustering, association rule mining, or sequential pattern mining (Baker & Yacef, 2009). Next, models are validated to ensure that they can be properly generalized over various learning contexts, and then are used to predict future learning and student performance. Finally, these predictions can be used to influence and support pedagogical policy decisions (Baker & Yacef, 2009; Romero & Ventura, 2010).

2.6 Literature Review Summary

The purpose of this literature review was to explore the current state of research relevant to this study. Several metrics of student learning and achievement were discussed, establishing that grit, self-efficacy, and engagement are important factors in academic success. Immediate feedback (whether evaluative or formative) has been shown to improve student engagement, confidence, and academic outcomes. Test-driven development was described as an iterative process in which a programmer writes a unit test, writes the code to pass the test, and then repeats. This review established that test-driven development can be considered a form of immediate feedback. Next, current research examining TDD and student learning was explored, finding that students generally produced higher quality code with TDD, along with improved productivity, test coverage, and code readability. Studies also determined that TDD may improve students’ knowledge acquisition and retention, with resulting benefits of increased academic achievement and better co-op positions. Notably, current literature also indicated that TDD may improve students’ confidence and motivation. However, these observations were only tangentially related to the research being conducted, and accordingly may warrant further exploration in this study. Finally, educational data mining was examined, identifying its main applications, generalized steps, and potential stakeholders.
Chapter 3

Methodology

The objective of this study was to perform exploratory research to establish potential relationships between TDD and student achievement. This chapter specifies the goals of this study (Section 3.1), describes the environment that the data were collected in (Section 3.2), and explains which measures were used to obtain the metrics this investigation will be comparing (Section 3.3 and Section 3.4). Finally, a description of relevant metrics and a summary of data are provided (Section 3.5 and Section 3.6).

3.1 Research Questions

To identify potential causal relationships for future research, this study examined the following research questions:

1. Are there any notable correlations between measurements of mental wellness and measurements of engagement with test-driven development activities?
2. Are there any notable correlations between measurements of mental wellness and measurements of success within test-driven development activities?

3.2 Data

244 students were provided with self-assessment surveys (see Section 3.3) and test-driven development tools (see Section 3.4) as part of a regular course they were taking. Usage of these tools and surveys were optional and non-graded, provided to students only to improve their learning and self-reflection. With Research Ethics Board approval, the data were anonymized and used in this study.
3.3 Surveys

Three surveys were given to students to improve their meta-knowledge about how they learn and to increase their self-reflection. The surveys were also intended to give the instructor information about what elements of learning were important to their students. The surveys provided were based on the Grit-S scale by Duckworth & Quinn (2009) (see Appendix A), the Generalized Self-Efficacy Scale by Schwarzer & Jerusalem (1995) (see Appendix B), and the Student Course Engagement Questionnaire by Handelsman et al. (2005) (see Appendix C). Each survey had a number of statements and respondents were asked to answer each question with respect to how they felt the statement applied to them, on a 5-point Likert scale. As per the survey sources, each survey was scored by assigning values from 1 to 5 to each Likert item, summing the results for each statement, and then dividing by the number of statements to give a score between 1 and 5. Values 1 and 5 was assigned to Strongly Disagree or Strongly Agree, depending on if the statement was positively or negatively worded. The grit survey contained five negatively worded questions and the remainder were positively worded. All questions on the self-efficacy and engagement surveys were positively worded.

3.4 Test-Driven Development

Test-driven development was accomplished via a plugin in the learning management system. Virtual Programming Lab (VPL) is an activity plugin for Moodle sites that allows instructors to set up programming assignments from within Moodle. It enables students to interactively edit and run programs from within the web browser and see their compilation and runtime output. VPL also allows instructors to configure tests for the program, and lets students see the test results (VPL, the Virtual Programming lab for Moodle, 2018). Students used the Java language and JUnit to run the instructor’s test cases. JUnit is an open-source framework that allows for unit tests to be easily written and verified for Java programs. JUnit lets the tester create assertions that compare the program results against the expected results, returning whether the test case passed or not (JUnit Frequently Asked Questions, 2018). Figure 3.1 shows a sample of a VPL execution file utilizing the JUnit framework to provide students with unit test feedback when they compile using the VPL activity.

When a student opened the relevant VPL activity, they were provided with the opportunity to load a file from their computer or write their Java files using VPL’s built-in editor. Once the student felt that they reached an appropriate testing point, they saved their file and then executed using the editor’s toolbar. Figure 3.2 shows an example of the toolbar, demonstrating the save, compile, and execute buttons. The VPL activity then compiled the student’s program with the required execution files provided by the instructor. If the program did not compile the student was provided with the compilation errors, whereas if
StudentTest.java

```java
import org.junit.Test;
import org.junit.runner.RunWith;
import org.junit.runner.RunWith;
import org.junit.runner.notification.Failure;
import static org.junit.Assert.*;

import org.junit.Before;
import static org.junit.Assert.*;

public class StudentTest {

    private Student student1;

    @Before
    public void setUp() {
        //this.student1 = new Student("Tester", "McTest", 127001);
        student1 = new Student();
        student1.setFirstName("Tester");
        student1.setLastName("McTest");
        student1.setStudentNumber(127001);
    }

    @Test
    public void checkDefaultConstructor() {
        Student obj = new Student();
        assertTrue(obj instanceof Student);
    }

    @Test
    public void getFirstName() {
        System.out.println("TESTING: getFirstName");
        String result = student1.getFirstName();
        assertEquals("tester", result.toLowerCase());
    }
}
```

Figure 3.1: Sample of a VPL execution file, with a setUp method to be run before each assertion, and two sample unit tests to be run on the students' submitted programs. Students were required to provide the Java file defining the Student class in this example.
it did compile the execution output was displayed instead. This output listed the JUnit tests that were performed and their pass/fail status. From here, the student could make changes to their program code and re-execute, entering the iterative process described in Section 2.3.

3.5 Metrics

Table 3.1 describes the metrics obtained from the three surveys, an abbreviated name for each, and the range of possible results. Table 3.2 and Table 3.3 provide the same information, but for engagement and success with the TDD activities, respectively.

The Skill Acquisition Speed metric outlined in Table 3.3 is based on the assumption that the percentage of test cases passed by a student should increase with each submission, creating a learning curve. Since exponential learning curves have been shown to best model individual learners (Heathcote et al., 2000), this analysis applied a nonlinear regression to each student using Equation 3.1, based on the formula provided by J. E. Beck & Mostow (2008).

\[
\text{Percentage of Test Cases Passed} (t) = -A \cdot e^{-b \cdot t} \quad (3.1)
\]

In Equation 3.1, the free parameter \( A \) represents the percentage of tests passed on the student’s first submission, \( e \) is Euler’s number (a constant with a value of approximately 2.7182), \( b \) is a free parameter representing the speed that the student learns the relevant skill, and \( t \) is the independent variable representing the number of submissions made by the student (J. E. Beck & Mostow, 2008). Three sample measurements of this metric are illustrated in Figure 3.3 for clarity. As depicted, a higher Skill Acquisition Speed (parameter \( b \) in Equation 3.1) indicates a quicker learning rate.

The Total Number of Sessions and Average Session Length metrics in Table 3.2 are based on the definition of sessions from Sheard (2010). In her article, Sheard explained that a timeout mechanism can be used to determine the cutoff times between sessions by assuming that a new session starts if the time between two actions exceeds a specified limit. In this investigation’s metrics, this thesis assumed that a new session began if there was no submission for 30 minutes, the most common timeout limit as per Sheard (2010).

![Figure 3.2: Example of the VPL editor’s toolbar.](image)
<table>
<thead>
<tr>
<th>Metric</th>
<th>Abbreviation</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grit Score</td>
<td>GritSc</td>
<td>Score calculation based on the Grit-S scale by Duckworth &amp; Quinn (2009).</td>
<td>1 to 5</td>
</tr>
<tr>
<td>Engagement Score</td>
<td>EngSc</td>
<td>Score calculation based on the Student Course Engagement Questionnaire by Handelsman et al. (2005).</td>
<td>1 to 5</td>
</tr>
</tbody>
</table>

Table 3.1: Description of survey metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Abbreviation</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Submissions</td>
<td>TotalSub</td>
<td>The number of submissions a student tested throughout the entire semester.</td>
<td>1 or more</td>
</tr>
<tr>
<td>Average Submissions Per Activity</td>
<td>AvgSubPerAct</td>
<td>Total submissions divided by the number of activities the student made at least one submission to.</td>
<td>0 or more</td>
</tr>
<tr>
<td>Total Number of Sessions</td>
<td>TotalNumSsn</td>
<td>The number of distinct development &amp; testing sessions a student used over the semester.</td>
<td>1 or more</td>
</tr>
<tr>
<td>Average Session Length</td>
<td>AvgSsnLen</td>
<td>The total time a student spent using VPL over the semester, in seconds, divided by the number of sessions.</td>
<td>0 or more</td>
</tr>
</tbody>
</table>

Table 3.2: Description of engagement with TDD metrics.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Abbreviation</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Submission</td>
<td>AvgSubPct</td>
<td>A student’s average percentage of test cases passed per submission over the entire semester.</td>
<td>0 to 100</td>
</tr>
<tr>
<td>Percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Submission Percentage</td>
<td>SubPctDiff</td>
<td>The range between a student’s highest percentage of passed test cases and their lowest instance.</td>
<td>0 to 100</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Submission Percentage</td>
<td>SubPctImpr</td>
<td>The range between the percentage of passed test cases of a student’s first submission and their last submission of the semester.</td>
<td>-100 to 100</td>
</tr>
<tr>
<td>Improvement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skill Acquisition</td>
<td>SkAcqSpd</td>
<td>The parameter $b$ in Equation 3.1, representing how quickly a student learns a skill.</td>
<td>Greater than 0</td>
</tr>
<tr>
<td>Speed</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Description of success with TDD metrics.

Not all measurements collected from the Moodle surveys and VPL activities were usable. Some data were missing or filtered out because they did not fit the parameters of the metric. The default data filters are explained in Table 3.4, although the filters were also modified in the analysis.

3.6 Data Summary

There were 200 total responses to the three surveys. There were 88 unique respondents to the grit survey, 44 to the self-efficacy survey, and 68 to the engagement survey (Table 3.5). Overall, there were 127 unique respondents.

130 students used at least one of the five provided VPL activities. The VPL activities collected a total of 1501 submissions over the semester, of which 685 submissions compiled and executed without error. Overall, there were 187 unique individuals who either completed a survey or used one of the TDD
## Metric Measurement Missing If:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Measurement Missing If:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grit Score</td>
<td>Student completed 0 questions on the grit survey.</td>
</tr>
<tr>
<td>Self-Efficacy Score</td>
<td>Student completed 0 questions on the self-efficacy survey.</td>
</tr>
<tr>
<td>Engagement Score</td>
<td>Student completed 0 questions on the engagement survey.</td>
</tr>
</tbody>
</table>

| Total Submissions   | N/A                                                         |
| Average Submissions Per Activity | N/A                                                         |
| Total Number of Sessions | N/A                                                         |
| Average Session Length  | N/A                                                         |

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Submission Percentage</td>
<td>Student made 0 submissions that compiled and executed correctly.</td>
</tr>
<tr>
<td>Submission Percentage Difference</td>
<td>Student made 0 submissions that compiled and executed correctly.</td>
</tr>
<tr>
<td>Submission Percentage Improvement</td>
<td>Student made 0 submissions that compiled and executed correctly.</td>
</tr>
<tr>
<td>Skill Acquisition Speed</td>
<td>Student made &lt;3 submissions, student made 0 submissions with a nonzero score, or the regression function was unable to find a fit.</td>
</tr>
</tbody>
</table>

Table 3.4: Description of the conditions under which a given metric was missing a measurement.

<table>
<thead>
<tr>
<th>Survey</th>
<th>Responses</th>
<th>Mean</th>
<th>Mode</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grit</td>
<td>88</td>
<td>3.15</td>
<td>3.35</td>
<td>1.70</td>
<td>4.80</td>
<td>3.10</td>
</tr>
<tr>
<td>Self-Efficacy</td>
<td>44</td>
<td>3.48</td>
<td>3.60</td>
<td>1.00</td>
<td>4.80</td>
<td>3.80</td>
</tr>
<tr>
<td>Engagement</td>
<td>68</td>
<td>3.76</td>
<td>3.72</td>
<td>2.61</td>
<td>4.91</td>
<td>2.30</td>
</tr>
</tbody>
</table>

Table 3.5: Summary statistics of mental wellness surveys.
activities.

Table 3.5 shows that engagement had the highest average score, followed by self-efficacy and then grit, in terms of both mean and mode. However, self-efficacy did have the largest range of scores. Table 3.6 displays an interesting discrepancy between the means and modes of TDD engagement measurements. The modes of the Total Submissions, Average Submissions Per Activity, Total Number of Sessions, and Average Session length indicate that the most common use case of the VPL activity was the user making only one submission on a single occasion. The means of each metric are higher though, signaling that the students that did not make just one submission contributed enough to raise the average significantly. Similarly, the modes of the Average Submission Percentage, Submission Percentage Difference, and Submission Percentage Improvement indicate that the most common submission score was 100.00%, followed by 0.00%. This could mean that students often passed all test cases when they were able to properly execute their programs.
<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean</th>
<th>Mode</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Submissions</td>
<td>8.02</td>
<td>1</td>
<td>0</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>Average Submissions Per Activity</td>
<td>2.39</td>
<td>1.00</td>
<td>0.00</td>
<td>26.00</td>
<td>26.00</td>
</tr>
<tr>
<td>Total Number of Sessions</td>
<td>1.70</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Average Session Length</td>
<td>431.80</td>
<td>0</td>
<td>0</td>
<td>3177</td>
<td>3177</td>
</tr>
<tr>
<td>Average Submission Percentage</td>
<td>89.25</td>
<td>100.00</td>
<td>14.29</td>
<td>100.00</td>
<td>85.71</td>
</tr>
<tr>
<td>Submission Percentage Difference</td>
<td>94.98</td>
<td>100.00</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Submission Percentage Improvement</td>
<td>34.59</td>
<td>0.00</td>
<td>-100.00</td>
<td>100.00</td>
<td>200.00</td>
</tr>
<tr>
<td>Skill Acquisition Speed</td>
<td>0.44</td>
<td>0.44</td>
<td>0.03</td>
<td>1.86</td>
<td>1.83</td>
</tr>
</tbody>
</table>

Table 3.6: Summary statistics of test-driven development activity.
Chapter 4

Results

This chapter performs an exploratory analysis, then provides an initial correlational analysis (Section 4.1). Next, a method of clustering students by their usage type is explored (Section 4.2) and an analysis is performed on the resulting profiles (Section 4.3). Finally, an analysis of student interactions over time is presented (Section 4.4).

To begin the exploratory analysis, Figure 4.1, Figure 4.2, and Figure 4.3 provide histograms of the results from the three surveys. All three distributions are roughly symmetric over the mean, although each may have a slight right skew, indicating that the average student leaned towards a higher self-reported score.

Figure 4.4, Figure 4.5, and Figure 4.6 provide histograms of three of the primary TDD measurements. Figure 4.4 and Figure 4.5 display a right skew, which can be expected from an opt-in activity. When this information is combined with the modes of Total Submissions and Total Number of Sessions, which are both 1, it becomes apparent that the most frequent use case was that the student submitted once in one session, and did not attempt any more activities. However, it does appear that the students who continued to use the VPL activities did so to a greater extent, bring the mean up to 8.02 and 1.70, respectively. Figure 4.6 indicates a left skew, towards the higher average percentages. Bearing in mind that the mode of this metric is 100.00%, it may be the case that many of the users with only 1 Total Submission simply uploaded their completed program and ran the unit tests as a verification tool, rather than a development tool. Then, once they confirmed that their code passed 100% of the tests, they did not return to any other activities. This investigation considered this possible grouping of users in further analysis.

Finally, Table 4.1 illustrates the number of students that identified as High (greater than or equal to 3) or Low (less than 3) scores in one of the three surveys, compared with their results from another survey. Most of the students that self-reported as a High score in one survey also evaluated themselves as
Figure 4.1: Histogram of the Grit Score measurements, as a score from 1 to 5. The mean of 3.15 is represented by the solid red line and the mode of 3.35 is represented by the dashed blue line.

Figure 4.2: Histogram of the Self-Efficacy Score measurements, as a score from 1 to 5. The mean of 3.48 is represented by the solid red line and the mode of 3.60 is represented by the dashed blue line.
Figure 4.3: Histogram of the Engagement Score measurements, as a score from 1 to 5. The mean of 3.76 is represented by the solid red line and the mode of 3.72 is represented by the dashed blue line.

Figure 4.4: Histogram of the Total Submissions measurements.
Figure 4.5: Histogram of the Number of Sessions measurements.

Figure 4.6: Histogram of the Average Submission Percentage measurements.
High in another survey. Notably, very few of the students that reported as Low in a survey also assessed themselves as Low in another metric. Lastly, the grit survey was most frequently the student’s only completed survey (43 students), compared to engagement (19 students) and self-efficacy (7 students).

To continue this exploration, this thesis examined the correlations between the eleven default metrics outlined in Section 3.5. This analysis used both Pearson product-moment correlation and Spearman’s rank-order correlation when analyzing multiple metrics at once. The Pearson correlation coefficient ($r$) is a measure of the linear relationship between two metrics, whereas Spearman’s correlation coefficient ($r_s$) is the nonparametric version of the Pearson correlation, simply measuring a monotonic relationship between the metric. It was likely that Spearman’s correlation coefficient would provide a more accurate estimation of the relationship between the metrics because they may be ordinal (such as Skill Acquisition Speed and the survey scores) or contain outliers, which Spearman’s handles better than Pearson. However, some relationships have been better described by the Pearson correlation coefficient if the measurements followed a linear, normal distribution. The correlograms in Figure 4.7 and Figure 4.8 provide an illustrative depiction of the correlations between variables. A larger circle and stronger shade indicates a stronger correlation, and the colour indicates a positive or negative correlation.

### 4.1 Initial Analysis

To begin, it is worth noting the issue regarding the relationship between Skill Acquisition Speed and Submission Percentage Difference. In the correlograms this problem is represented by a “?” and is represented by an “NA” in the

<table>
<thead>
<tr>
<th>Metric</th>
<th>Self-Efficacy</th>
<th>Engagement</th>
<th>No Other Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Grit</td>
<td>High</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Engagement</td>
<td>High</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No Other Survey</td>
<td>5</td>
<td>2</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4.1: Cross tabulation of the number of Grit, Self-Efficacy, and Engagement score responses, separated by High (greater than or equal to 3) and Low (less than 3) scores. The “No Other Survey” column/row represents the number of students who completed the indicated survey and neither of the other two surveys. The cell where the “No Other Survey” column intersects with the “No Other Survey” row represents the number of students that interacted with the VPL activities but completed no surveys.
Figure 4.7: Initial correlogram of Pearson correlations.

...scatter plots. The reason that the correlation coefficients could not be calculated is that each of the 19 students that were able to fit an exponential learning curve had a Submission Percentage Difference of 100%. In other words, each student with at least three submissions, where at least one submission had a nonzero score (as per the default filters in Table 3.4), had a minimum submission score of 0% and a maximum score of 100%. This means that the standard deviation of the Submission Percentage Difference is zero, and thus the correlation could not be calculated. It is interesting to note that each of the students that provided enough data points (i.e. submissions) to be fitted with a learning curve improved from a submission score of 0% to a submission score of 100% at some point. This could indicate that students using the TDD activities repeatedly did show an improvement in testing results.

In both Pearson and Spearman’s correlation, the three survey metrics have moderate positive correlations with one another, which is not unexpected. In the Pearson correlation, the relationships between Grit Score and Self-Efficacy Score, Self-Efficacy Score and Engagement Score, and Engagement Score and Grit Score each have a moderate-to-strong positive correlation ($r = 0.69$, 0.52, and 0.61, respectively) compared to a weak-to-moderate positive correlation with Spearman’s ($r_s = 0.51$, 0.32, and 0.43, respectively). In either case, it is not...
unreasonable to conclude that grittier students are more likely to have higher self-efficacy and engagement with course material than less gritty students, and vice versa. Similarly, students with high self-efficacy may be likelier to engage with the course, and vice versa. These correlations were not unanticipated though, since the literature indicated that all three metrics positively correlated with academic performance and achievement (as per Section 2.1).

Another relationship with a moderate-to-strong positive correlation of \( r = 0.66 \) and \( r_s = 0.92 \) is between Total Submissions and Total Number of Sessions. This correlation was not entirely surprising, since a student that has more sessions in the VPL activities is likelier to have more submissions through the nature of spending more time in the testing environment, and vice versa. Likewise, the moderate-to-strong positive correlation between Average Submissions Per Activity and Total Number of Sessions (\( r = 0.61 \) and \( r_s = 0.85 \)) was expected as well through similar logic: a student spending more time in the VPL environment (higher Total Number of Sessions) will be more likely to create a larger number of Total Submissions, and thus, a higher number of Average Submissions Per Activity.
An interesting group of relationships is the strong positive correlations between Average Session Length and Total Submissions, Average Submissions Per Activity, and Total Number of Sessions \( (r_s = 0.84, \ r_s = 0.84, \text{ and } r_s = 0.67, \text{ respectively}) \). This correlation indicates that learners that spend more time in the testing environment in each session are more likely to run more tests overall, and thus more tests per activity, which conforms with the expected outcomes. However, the correlation between Average Session Length and Total Number of Sessions is noteworthy, since they are two unrelated metrics of engagement with test-driven development. This lent some credibility to both metrics as potentially valid methods of measuring engagement with the TDD environment.

The relationships involving Total Submissions, Average Submissions Per Activity, and Total Number of Sessions are intriguing as well. Each of these three metrics has a weak-to-moderate negative correlation with the Average Submission Percentage \( (r_s = -0.44, \ r_s = -0.46, \text{ and } r_s = -0.34, \text{ respectively}) \). This is likely because the Average Submission Percentage skews heavily towards 100%, and the other three metrics each skew towards the lower end. As discussed in Section 3.6, this may be due to the portion of users that wrote their programs outside of the VPL activity, then imported their code to verify its accuracy (i.e. test-last). The Average Submission Percentage metric also weakly, negatively correlates with the Self-Efficacy Score \( (r_s = -0.34) \), the Engagement Score \( (r_s = -0.23) \), and the Average Session Length \( (r_s = -0.26) \). The potential test-last use case was noted and explored further in Section 4.2.

Another noticeable correlation observed by the Pearson coefficient is between the Average Session Length and Skill Acquisition Speed \( (r = 0.45) \). This moderately positive correlation indicates that students that spend more time in each VPL session are likelier to improve their program’s test pass rate more quickly over repeated trials. Since the Skill Acquisition Speed metric treated the submission number as the independent variable in its curve fitting, it may relate to the Average Session Length in a couple potential manners: either students who take more time between submissions improve their pass rate more quickly, or students who complete more submissions in a row (without a timeout ending the session) improve their submissions more quickly.

The last moderate-to-strong positive correlation is between the Engagement Score and Skill Acquisition Speed metrics \( (r_s = 0.55) \). This could imply that more engaged learners are more likely to improve their test results on a submission-by-submission basis. Some of the survey questions that respondents with higher Engagement Scores answered positively to may also lend credence to this proposal, such as really desiring to learn the material is important to me, putting forth effort is important to me, and doing all the homework problems is important to me (see Appendix C).

Finally, there are a few weak correlations between survey metrics and success with TDD metrics, namely between the Grit Score and Skill Acquisition
Speed \((r = 0.21\) and \(r_s = 0.26\)), and between the Self-Efficacy Score and Submission Percentage Improvement \((r = 0.30\) and \(r_s = 0.26\)). Unfortunately, the relatively weak correlations, and the fact that there are not any other noteworthy correlations between survey metrics and engagement or success with TDD metrics, mean that it is difficult to draw likely conclusions from these relationships.

Many of the 20 different correlations examined in this section were not surprising given the expected relationships between metrics, but there were a number of standouts. The correlations between Grit Score and Skill Acquisition Speed, and between Average Session Length and Skill Acquisition Speed were unexpected, and may be worth investigating further. Additionally, the negative correlations between Average Submission Percentage and the 6 discussed metrics were unforeseen, and are potentially interesting for future research.

### 4.2 Data Clustering and User Profiles

Inspired by the possible explanation of a test-last use case in the previous section, this investigation proceeded to examine the correlations and raw data to try to find additional potential user profiles. These user profiles were partitioned by manner in which they interacted with and utilized the TDD activities.

First, this study investigated the use case where the student writes their program outside of the VPL activities, then runs the provided test cases to verify its accuracy. These test-last users should have had a relatively low number of submissions because they only needed to execute the tests for verification. These students should also have had a high Average Submission Percentage, because the programs’ functionality was complete or nearly complete when it is imported into VPL for the first time. By inspecting the scatter plot between Average Submission Percentage and Total Submissions in Figure 4.9, this analysis observed a noticeable grouping of users in the upper-left corner with low Total Submissions but high Average Submission Percentages. It was likely that these students fit the description of the test-last use case.

A basic, naive approach to classifying this user group might consider any users with an Average Submission Percentage of 100% and Total Submissions of 1 to be test-last users. However a clustering approach to classification may find a partitioning method that is more representative of the actual groups. To do this, this analysis examined the results of \(k\)-means, DBSCAN, and OPTICS clustering algorithms.

The \(k\)-means clustering algorithm is a centroid-based method that separates clusters by linear boundaries. This algorithm takes the number of clusters to be identified \((k)\) as input. The result of \(k\)-means on the data from Figure 4.9 is illustrated in Figure 4.10, where the blue circles represent the test-last users.
This clustering result captured the definition of a test-last use case fairly well, although it did include some students with an Average Submission Percentage less than 80%, which was perhaps slightly too low.

The DBSCAN (density-based spatial clustering of applications with noise) algorithm is a density-based procedure that groups data points together if they are within a certain distance of one another (given by input parameter $\epsilon$). This algorithm also takes the minimum number of points to form a cluster as input ($MinPts$). The results of DBSCAN on the data are presented in Figure 4.11. This clustering appeared to match the test-last use case better than the $k$-means algorithm. If this analysis considers the red diamond data points as the cluster of test-last students, DBSCAN has left the students with a sub-80% Average Submission Percentage out. The algorithm has also excluded some students with a high Average Submission Percentage and more than 20 Total Submissions from the test-last cluster, which was also in line with the ideal description.

Finally, the OPTICS (ordering points to identify the clustering structure) algorithm is a density-based process that operates similarly to DBSCAN. The main difference is that OPTICS can account for varying densities within the data, whereas DBSCAN does not. When identifying cluster density, OPTICS notes the reachability of each data point, which is a distance to its closest points.
Figure 4.10: $k$-means clustering of Total Submissions and Average Submission Percentage ($k = 2$). Clusters are indicated by data point colour and shape.

OPTICS has the same input parameters as DBSCAN ($\epsilon$ and $MinPts$), but also has a parameter representing the cutoff point for the reachability of a data point (given by input parameter $\epsilon_c$).

The results of OPTICS on the data are shown in Figure 4.12. The test-last cluster represented by red diamonds is the same as in DBSCAN, but more of the other data points are considered noise in this model. Due to its fit with the description of the test-last user profile, this investigation used the clusters identified by the DBSCAN and OPTICS algorithms to define the test-last cluster. All other data points in Figure 4.11 and Figure 4.12 were considered as part of a test-first use case. These test-first students may have a low Average Submission Percentage, possibly because they developed their program in the VPL activity and thus their submissions started with a low number of passed test cases, bring down their average. Alternately, the test-first students may have a high number of Total Submissions, also due to the fact that they ran many tests while developing their program using the VPL activity.

In the previous clustering analyses, only students with an Average Submission Percentage measurement could be considered for clustering. However, according to the default data filters in Table 3.4, a student did not have an Average Submission Percentage measurement if they did not make at least one submission.
that compiled and executed correctly. This means that there was an additional group of students that did not fit into the existing test-first or test-last clusters. A natural division of this additional group appeared to be whether the student made any submission attempts or not. If the student did not make any submissions, then they did not interact with the VPL activities at all. However, if the student did make at least one submission then they attempted to use the activities, but gave up before they were able to get their program to compile or execute. To summarize, this analysis was able to partition the users into the following groupings by use case:

- **Test-Last**: Students with a high Average Submission Percentage and a low number of Total Submissions, as identified by the DBSCAN and OPTICS clustering algorithms.
- **Test-First**: Students with a low Average Submission Percentage or a high number of Total Submissions, as identified by the DBSCAN and OPTICS clustering algorithms.
- **Sampler**: Students with at least one Total Submission, but no Average Submission Percentage measurement because they were unable to get their program to compile or execute.
- **No-Attempt**: Students with no Total Submissions, and thus no Average Submission Percentage measurement.
By applying the definitions of these four user profiles to the data, this investigation found that it partitions into the subsets outlined in Table 4.2. From this information, this study discerned that of the 130 students that made submissions to at least one VPL activity (Test-Last students, Test-First students, and Samplers), 114 were able to compile and execute a submission (87.7%), which appears to be relatively high for a non-compulsory activity. In the following section, this thesis investigated if there were any significant differences between the user profiles.

<table>
<thead>
<tr>
<th>User Profile</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>All students with measurements</td>
<td>187</td>
</tr>
<tr>
<td>One or more executed submissions</td>
<td>114</td>
</tr>
<tr>
<td>Test-Last</td>
<td>79</td>
</tr>
<tr>
<td>Test-First</td>
<td>35</td>
</tr>
<tr>
<td>No executed submissions</td>
<td>73</td>
</tr>
<tr>
<td>Sampler</td>
<td>16</td>
</tr>
<tr>
<td>No-Attempt</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 4.2: Student counts of the defined user profiles.
4.3 Profile Analysis

To begin the comparison of the four defined user profiles, boxplots are presented in Figure 4.13, Figure 4.14, and Figure 4.15. Each boxplot illustrates the median, minimum, maximum, and interquartile range (IQR) of the user profiles for a different survey metric. The Sampler group appears to have a higher average Grit Score and Engagement Score than the other three user profiles, which was unexpected given the group's definition as students that tried a VPL activity but gave up before creating a working submission. However, since only 16 students in the data were classified into the Sampler profile, it may be the case that the sample size was too small to obtain an accurate estimate of the true average. By investigating further, this study found that only 4 of the Sampler students completed the grit survey, 3 completed the self-efficacy survey, and 3 completed the engagement survey. Considering the small number of measurements, it was likely that this unanticipated result was due to the Sampler group’s susceptibility to noise.

Interestingly, the medians and IQRs for each user profile appear to be relatively similar across metrics, with the exception of the Sampler usage types. This analysis expected that the three survey measurements would be generally higher for students that interacted with the VPL activities consistently, i.e., the Test-Last and Test-First students, so it was notable that they seem to be

![Figure 4.13: Boxplot of Grit Scores, separated by user profile.](image-url)
Figure 4.14: Boxplot of Self-Efficacy Scores, separated by user profile.

Figure 4.15: Boxplot of Engagement Scores, separated by user profile.
so close to the No-Attempt students. It is possible that the aggregate scores of the students’ self-reported surveys do not have any strong correlation with the manner in which they voluntarily used the VPL activities. A potential reason for this may be that the Grit-S survey, Generalized Self-Efficacy Scale, and Student Course Engagement Questionnaire were not originally designed to be used in the context of programming and student learning, but rather as predictors of academic performance. To explore whether this is the case, this study investigated how each individual survey question related to the given user profiles. In Appendix E, each survey question is presented alongside the mean response score for the four user profiles. Given that the distribution of aggregate survey scores in Figure 4.13, Figure 4.14, and Figure 4.15 are overall similar, the majority of the mean responses to individual questions were comparable as well. By identifying which questions differ between user profiles, this study was able to begin to form a new method of evaluating the students’ learning tendencies.

First, the largest difference in means in the grit survey occurred in the “Setbacks don’t discourage me. I don’t give up easily” question. Ignoring the results from the Sampler group because of its low number of measurements (only 4), this analysis found that the Test-First profile has a noticeably higher mean than the Test-First and No-Attempt groups. The boxplot of the responses to this question are provided in Figure 4.16, where the Test-First group has an observably higher third quartile, indicating that students in this group may agree more strongly with this question than the other profiles.

![Question Responses by User Profile](image)

Figure 4.16: Boxplot of responses to “Setbacks don’t discourage me. I don’t give up easily,” separated by user profile.
The next two questions in the grit survey with the largest disparities in responses were “I often set a goal but later choose to pursue a different one” and “My interests change from year to year.” In both of these questions the Test-First group again has the highest score, however, these questions are negatively worded. Students in the Test-First profile may agree more strongly with these statements as well, but it means that they would display lower grittiness than the other profiles in this regard. This is not necessarily at odds with the previous grit question, since Test-First students may change their goals more frequently, but still be less affected by setbacks than the other profiles.

In the self-efficacy survey, there were a few questions where the responses differed between user profiles. The question with the largest gap is “It is easy for me to stick to my aims and accomplish my goals,” where the Test-First profile has the highest mean. While there is a small gap between the Test-First and Test-Last groups, both have a much higher mean than the No-Attempt profile. This may imply that students who have difficulty working towards their goals were less likely to engage with the TDD activities. This reasoning may be strengthened by examining Figure 4.17, where No-Attempt students have a markedly lower first quartile than the other profiles.

The three questions from the self-efficacy survey with the next highest disparities in responses are “I can always manage to solve difficult problems if I try hard enough,” “When I am confronted with a problem, I can usually find

![Boxplot of responses to “It is easy for me to stick to my aims and accomplish my goals,” separated by user profile.](image_url)
several solutions,” and “If I am in trouble, I can usually think of a solution.” In each of these statements, the Test-First students have the highest mean response and the No-Attempt students generally have the lowest. Since all three of these statements have similar themes, it may imply that learners with lower confidence in their ability to find a solution are less likely to take part in VPL activities, or that students with the most confidence may engage more with test-driven development.

Three of the questions in the engagement survey have prominent differences in mean responses: “Thinking about the course between classes,” “Going to the professor’s office hours for questions,” and “Staying up to date on readings.” In the first statement the Test-First students have a considerably high mean response, almost a full point higher than the Test-Last and No-Attempt students. This is illustrated in Figure 4.18, where the minimum response in the Test-First group is no lower than the third quartile of the other user profiles. Interestingly, in the second and third statements the Test-First profile has a remarkably lower mean response than the other groups. When considered together, it may be the case that Test-First students are more likely to reflect on the course material by themselves, opting to use the provided TDD activities to learn rather than consult an instructor or the course literature.

Last, the questions “Coming to class every day” and “Taking good notes in class” in the engagement survey have noticeable disparities between user profiles as well. In both of these questions, the Test-First students agreed most
strongly out of the groups. This was perhaps unexpected, but may suggest that students who attended and engaged with lectures most regularly were likeliest to use the VPL activities for test-driven development. These students might have invested in the concept of TDD more readily because it was discussed and promoted in class.

As a method of further analyzing the differences between user profile responses to individual questions, this investigation examined the distribution of responses to each question. In this section of the analysis the Sampler group was removed from consideration due to its low number of respondents, only three or four students for each survey. Every question response was categorized by the user profile of the respondent, and then the counts of responses were placed in a table with user profile as the row headers and Likert scale answers as column names. Next, a chi-squared test was run on each question’s table to determine which questions differed the most in distribution of responses. The survey questions with a \( p \)-value less than 0.2 were considered sufficiently different between usage types for further examination. There were seven questions with a \( p \)-value less than 0.2, and they are displayed as proportional bar plots in Figure 4.19. One of the questions is from the grit survey (Figure 4.19a), and the remainder are from the self-efficacy survey (Figures 4.19b through 4.19g). Since the most frequent responses differ between user profiles in these questions, it may be possible to use them as a way to classify users into usage types according to their responses.

In Table 4.3, the modes of each of the seven identified questions are presented by usage type, and the totals were examined as a potential metric for classifying users based on their responses. Unfortunately the totals are all similar, ranging from 26 to 29. However, this analysis then looked at the question “Going to the professor’s office hours for questions.”, which was lower for the Test-First group than the others. By considering how well this question predicts the user profile, this investigation converted the question into its negatively worded counterpart (i.e., “Going to the professor’s office hours for questions is not important to me.”). Thus, the Test-First group’s mode of 2 became 4, and the Test-Last group’s mode of 4 became 2. By doing this, the total score for the Test-Last and No-Attempt groups became 26 and 27, respectively, and the total score for the Test-First group became 28-31. After this step, the difference between the total scores had a clearer demarcation, and it should be possible to predict if a student will belong to the Test-First group based on their responses to the seven indicated questions.

In this section, this study examined the aggregate responses to the grit, self-efficacy, and engagement surveys per user profile. Unfortunately, the responses tended to be fairly similar regardless of usage type, which may be because the survey metrics best correlate with academic performance rather than student learning. However, by investigating the survey questions by themselves, this analysis was able to identify a number of areas in which the user profiles differed significantly.
Setbacks don’t discourage me. I don’t give up easily.

Going to the professor’s office hours for questions.

Having fun in class.

Coming to class every day.

Taking good notes in class.

Being organized.

Putting forth effort.

Figure 4.19: Proportional bar plots of Likert scale responses to individual survey questions, separated by user profile.
<table>
<thead>
<tr>
<th>Question</th>
<th>Test-Last</th>
<th>Test-First</th>
<th>No-Attempt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setbacks don’t discourage me. I don’t give up easily.</td>
<td>4</td>
<td>3,4</td>
<td>3</td>
</tr>
<tr>
<td>Going to the professor’s office hours for questions.</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Having fun in class.</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Coming to class every day.</td>
<td>5</td>
<td>4,5</td>
<td>4</td>
</tr>
<tr>
<td>Taking good notes in class.</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Being organized.</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Putting forth effort.</td>
<td>4</td>
<td>4,5</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>26-29</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 4.3: Modes of responses to each survey question identified by the chi-squared tests as having differing distributions.

### 4.4 Temporal Analysis

Each submission made to a VPL activity over the semester was stamped with a date and time. Accordingly, this thesis explored various interactions in Moodle with regards to time. First, the VPL activity submissions were charted in Figure 4.20 and Figure 4.21, comparing the number of submissions in each 6-hour period around the assignment due date. In both figures, the submission time is given in UNIX Epoch time, which is the number of seconds since Jan. 1, 1970 at 12:00am.

In Assignment 1, the number of submissions increased in each 24-hour period up to the due date, then immediately dropped down to no submissions right after the deadline. A noticeable trend in activity submissions is that the most active period of the day appears to be from 8:30pm to 2:30am, followed by 2:30pm to 8:30pm, and then 8:30am to 2:30pm. This could indicate that the students generally did more development in the evening and less in the morning. There is also an observable spike in the period from 2:30am to 8:30am right before the assignment deadline when compared to that period in previous days, which is likely due to students performing last-minute testing. Finally, there is an interesting grouping of submissions approximately three days after the deadline, possibly from students correcting their Assignment 1 programs in preparation of Assignment 2, which builds on the requirements of the first assignment.

For Assignment 2 submissions, the period with the most activity is from 8:30am to 2:30pm on Nov. 24, 2018. Additionally, the morning and afternoon periods in Assignment 2 have a greater ratio of the submissions than in Assignment 1. A possible explanation for this is that the 48-hour period before the deadline is a Saturday and Sunday, so students were able to work
more during the day. Another reason for this differing distribution could be that the students learned from their experience in Assignment 1, and did not procrastinate as much on their program development for Assignment 2. Lastly, there is an unusual amount of submissions that were made up to 36 hours after the Assignment 2 due date. Since Assignment 2 was the last evaluation of their program, there was no need to continue development except out of personal interest. Alternately, the students making the late submissions could have still been developing for evaluation, opting to take a late penalty on their assignment.

Next, this investigation classified each user by the time period in which they were most active, then explored if there were any differences in survey results, engagement with TDD activity metrics, or success with TDD activity metrics. The course under examination took place from September to December in Canada, so the delineations between each time period were chosen to create the largest distinction between daylight hours. As such, the activity periods were selected as follows:

- **Morning**: The 6-hour period from 5:00:00am to 10:59:59am.
- **Midday**: The 6-hour period from 11:00:00am to 4:59:59pm.
- **Evening**: The 6-hour period from 5:00:00pm to 10:59:59pm.
- **Night**: The 6-hour period from 11:00:00pm to 4:59:59am.
Figure 4.21: Distribution of VPL Assignment 2 activity submissions by submission time. Each bin spans a 6-hour period, and the vertical red line indicates the Assignment 2 due date of Nov. 26, 2018 at 8:30am.

Table 4.4 presents the number of users in each activity period after the investigation placed each student in an activity period corresponding to their most frequent submissions. Notably, the majority of the 130 students to make at least one submission were most frequently active during the Night period, indicating that most users developed their assignments later in the day.

Next, the survey and TDD metrics were compared by activity period through the use of boxplots, and the notable results were included in this study. First, the difference in Grit Score between students of different activity periods is very interesting, and is illustrated in Figure 4.22. This boxplot indicates that students that developed most frequently in the morning tended to be grittier.

<table>
<thead>
<tr>
<th>Activity Period</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning</td>
<td>22</td>
</tr>
<tr>
<td>Midday</td>
<td>7</td>
</tr>
<tr>
<td>Evening</td>
<td>32</td>
</tr>
<tr>
<td>Night</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 4.4: Student counts of the defined activity periods.
than students who developed later into the afternoon and night. This result does conform with the definition of grit which states that gritty individuals generally have higher self-control and resilience, which may lend itself to waking up and starting assignment work earlier. Figure 4.23 depicts a similar observation, in that the *Night* users display a lower Self-Efficacy Score than any of the students in earlier activity periods. However, if there is a causal relationship it is difficult to determine if the students worked at night due to low self-efficacy, if they had low self-efficacy because they worked later in the day, or some combination thereof. On the other hand, it appears as if students who engaged with the VPL activities later in the day were more likely have a greater number of sessions. Figure 4.24 demonstrates that the third quartile of *Evening* and *Night* students is higher than nearly the entire *Morning* and *Midday* groups. However, Figure 4.25 shows that the distributions of the Average Session Lengths are similar between all activity periods. This potentially implies that although the students working later in the day had more VPL activity sessions, the students generally developed for the same time period, regardless of the time of day. The exception to this is the *Midday* group, but this may be because the group was sensitive to statistical noise due to its low number of measurements (only 5 students).

Finally, there is a prominent difference in the Total Submissions for each activity period. As seen in Figure 4.26, there is a marked difference in median and third quartile values, increasing as the activity period moves later into the

![Grit Score by Activity Period](image)

Figure 4.22: Boxplot of Grit Scores, separated by activity period.
Figure 4.23: Boxplot of Self-Efficacy Scores, separated by activity period.

Figure 4.24: Boxplot of Total Number of Sessions, separated by activity period.
Figure 4.25: Boxplot of Average Session Length, separated by activity period.

When considering that the Evening and Night groups showed a higher Total Number of Sessions this difference was not entirely unexpected, but it is still noteworthy.

Furthermore, the user profiles defined in Section 4.2 were partially based on the Total Submissions that each user made, so there could have been a relationship between the user profiles and activity periods. To explore this, Figure 4.27 displays the number of students belonging to activity periods within each user profile. Since the activity period classification was derived from the students’ submission times, users with no submissions were not classified by activity period, which means that the No-Attempt group was excluded. Figure 4.27 shows that there is not any substantial difference between the distributions of activity periods, except that the Test-Last group has a lower proportion of Midday users. This could potentially be because the Test-Last students were more likely to verify their program accuracy the night before or morning of the assignment deadlines. It is also possible that the Test-First and Sampler profiles would follow this distribution as well, but noise in the measurements happened to prevent it from becoming evident.
Figure 4.26: Boxplot of Total Submissions, separated by activity period.

Figure 4.27: Bar chart of activity period students, separated by user profile.
Chapter 5

Discussion

This chapter discusses the results from Chapter 4 in the context of the initial research questions (Section 5.1). Next, the limitations of this study are outlined (Section 5.2.1), and potential solutions are provided to inform future research (Section 5.2.2). Finally, this thesis is summarized and the primary outcomes are restated (Section 5.3).

5.1 Research Questions

In this section, this investigation uses the information and insights obtained in the prior section to discuss the research questions developed from the thesis statement.

5.1.1 Are there any notable correlations between measurements of mental wellness and measurements of engagement with test-driven development activities?

In order to inform future research, it was important to identify if there are any noteworthy correlations between the survey scores and how the students interacted with the TDD activities. As noted by Section 4.1 in Figure 4.7, the only initial correlations appear to be between Self-Efficacy Scores and Total Submissions, Average Submissions Per Activity, and perhaps Average Session Length ($r = 0.31$, $r = 0.33$, and $r = 0.22$, respectively). However, these correlations were only weakly positive, so this thesis classified each student by their usage type, i.e. engagement with the VPL activities. Again, no significant differences between profiles were identified, so individual survey questions were considered. This investigation found that there were a number of survey questions that varied considerably, in which the differences between engagement levels could be established, as follows:

- Setbacks don’t discourage me. I don’t give up easily.
- I often set a goal but later choose to pursue a different one.
• My interests change from year to year.
• It is easy for me to stick to my aims and accomplish my goals.
• I can always manage to solve difficult problems if I try hard enough.
• When I am confronted with a problem, I can usually find several solutions.
• If I am in trouble, I can usually think of a solution.
• Thinking about the course between classes is important to me.
• Going to the professor’s office hours for questions is important to me.
• Staying up to date on readings is important to me.
• Coming to class every day is important to me.
• Taking good notes in class is important to me.

By considering the differences in responses, it appears the Test-First students may change their goals most frequently, but are less impacted by setbacks than the other user profiles. The Test-First group may also find it easier to work towards their goals, and have more confidence in their ability to find a solution to a problem. Finally, the Test-First students may also value working in between classes more highly, and agree more strongly that it is important to attend lectures in an engaged manner.

5.1.2 Are there any notable correlations between measurements of mental wellness and measurements of success within test-driven development activities?

Section 4.1 discussed several correlations between survey scores and this study’s measurements of TDD success, providing potential explanations. One of the most prominent of these correlations was between the Engagement Score and Skill Acquisition Speed ($r_s = 0.55$), possibly indicating that engaged students are more inclined to improve their programs between submissions. There is also a weak positive correlation between the Grit Score and Skill Acquisition Speed ($r = 0.21$ and $r_s = 0.26$), potentially signaling that gritty students also display this tendency.

There is another weak correlation between the Self-Efficacy Score and the Submission Percentage Improvement ($r = 0.30$ and $r_s = 0.26$). Since the Submission Percentage Improvement represents the difference between the students’ first and last submission scores, learners with a higher self-efficacy could tend to stick with the VPL activities longer despite a poor first submission, persevering until their program achieves a high submission score.

Finally, this study explored the weak negative correlation between the Self-Efficacy Score and the Average Submission Percentage ($rs = -0.34$) by identifying different user profiles in Section 4.2. It was hypothesized that users with a high Average Submission Percentage had a lower Self-Efficacy Score because they were only using the VPL activities to verify their programs’ accuracy after development, placing them in the Test-Last group. Figure 4.14 indicates that the Test-Last profiles do indeed have a lower average Self-Efficacy Score, and
that the *Test-First* group has the highest.

The user profiles were based off a metric of TDD success (Average Submission Percentage) in addition to a metric of TDD engagement, so the individual survey questions discussed in Section 5.1.1 should apply to this research question as well.

### 5.2 Future Work

The purpose of this exploratory study was to inform future research, so this section discusses the limitations of the investigation in Section 5.2.1, then uses them to suggest improvements to subsequent studies in Section 5.2.2.

#### 5.2.1 Limitations

This investigation was affected by four limitations that should be considered:

1. Participation in mental wellness surveys and VPL activities was voluntary, introducing a selection bias.
2. Test-driven development was not part of the course material, and use of its practice was optional.
3. No control groups meant that it is difficult to attribute any differences to the use of test-driven development.
4. The surveys used showed a correlation with academic performance and code quality, and not necessarily the metrics of this study.

The remainder of this section provides a further explanation of each limitation, and how they potentially impacted the results.

1) In the course offering, the grit, self-efficacy, and engagement surveys were all provided to students to optionally complete on their own time. These surveys were given to allow the students to reflect on themselves and their learning habits. They were also provided to give the instructors more information about what aspects of learning were most important to the students. Likewise, the test-driven development activities were provided as additional resources to potentially aid in learning the course material, but were not required for evaluation. Unfortunately, this bias of self-selection may have also introduced a sampling bias to the results. Since the students that voluntarily engaged in the surveys and activities do not necessarily represent the entire population of computer science students, the capacity to generalize the results to the population was diminished. Similarly, the selection bias introduced impacted the data analysis, hampering the internal validity of the results. Regrettably, these biases could not be avoided in this investigation due to the secondary nature of the collected data.

2) The focus of the course under examination was the introduction to object-oriented programming, and not to promote various testing methods. This means that test-driven development, as it compares to test-first and other
testing techniques, was given only a cursory explanation during the semester. Since the students were not introduced to the potential benefits of TDD, is it possible that the number of students implementing it was lower than it may have otherwise been. This affected the counts of the user profiles in Section 4.2, and subsequent analysis on those groups.

3) With no control group, it became particularly challenging to evaluate the efficacy of different pedagogical methods. A control group would allow this study to determine if there was no effect from TDD, if there was a confounding variable that influenced both groups, or if TDD did indeed have a positive effect on metrics of mental wellness. Unfortunately, as this thesis was based on secondary data, information of a blind or double-blind nature was not accessible for analysis.

4) The Grit-S scale, Generalized Self-Efficacy Scale, and Student Course Engagement Questionnaire have been shown by literature to have a positive correlation with both academic outcomes and code quality. However, they do not have evidence indicating that they are a suitable metric for analysis with the third major factor in student learning, which is the learners’ mental state. This means that it is possible that the results obtained from their correlations was not applicable to this area of pedagogical research, and that the correlations (and lack thereof) noted are not necessarily precise. Despite this, further analysis in Section 4.3 indicated that there were a number of individual questions from the surveys that may have relevance to classifications of the students’ engagement and mental state.

5.2.2 Further Research

Although nothing conclusive was identified, this investigation revealed several directions for potential future research. This section addresses possible measures for correcting for the limitations of this study, and then reviews indicated trends to inform further investigation.

First, making the mental wellness surveys and test-driven development non-optional should account for most self-selection biases in the data. By allowing participating students to decide which aspects they wanted to take part in, the secondary data obtained for this study contained many students who took part in one of the activities but not the other, reducing the amount of data available for correlation. This also threatened the internal and external validity of the study, so requiring participants to complete both activities may allow the results to be better generalized to the population of computer science students.

As discussed in Section 2.4, literature indicates that students are generally reluctant to embrace test-driven development. However, it was noted that less experienced students were more likely to adopt TDD, possibly because they had less resistance to new, different programming techniques. As such, promoting
test-driven development and its benefits as part of the curriculum in introductory courses may result in better comprehension of why it is advantageous, thus leading to improved adoption rates.

To fully explore the impact of TDD, a future experiment may randomly place students into one of three groups: test-first development, test-last development, and no-test development. This will provide control groups, so as long as the other factors remain the same, a conclusive judgment can be drawn about TDD. However, the ethics of placebo groups can be debatable, in that participants may be adversely affected by not receiving the potential benefits of the subject of the study. This can be considered an issue mostly for medical studies, but any experiment operating under the hypothesis that its solution is beneficial should also consider the issue. In the case of clinical studies, placebos are generally only allowed for medical conditions with no proven treatment (Nardini, 2014), preventing participants from being harmed by being denied treatment. Although this area of study is not as high-stakes as medical research, instructors do have a responsibility to provide the best education for their students that they reasonably can. Since test-driven development has been shown by research to improve academic outcomes and code quality, it may not be ethical to deny its use to randomly-selected students for the purpose of establishing its efficacy at improving mental wellness.

Next, although the three surveys used in this investigation may not necessarily be suitable, there were individual statements that showed promise of differentiation between usage types. In future studies, it may be worthwhile to consider developing a new survey, combining the relevant elements of each of the grit, self-efficacy, and engagement surveys. An acceptable start for this has been established in Section 4.3, with the consolidated list of questions summarized in Section 5.1.1. Additionally, it could be advantageous to offer the surveys at multiple points throughout the semester, allowing for a longitudinal study to be performed in addition to latitudinal. A difference in grit measurements over the course of a semester should not be anticipated though, since grit is a personality trait that does not vary much over short periods. However, it could be extremely informative to explore how changes in self-efficacy and engagement over the semester compare with usage rates of test-driven development.

Finally, a number of potential relationships between metrics were discussed in Section ?? . While not conclusive, the most prominent are provided as follows to provide a starting point for further study:

- Grit, self-efficacy, and engagement appeared to be correlated with one another.
- Students with longer VPL session times appeared to improve their program's submission score more rapidly.
- Students that were more engaged appear to improve their program's submission score more rapidly.
• *Test-First* students may change their goals more often, but are less affected by setbacks.
• *Test-First* students may find it easier to persevere at their goals, and have more confidence in their problem-solving.
• *Test-First* students may engage with classes more, and value working between classes more highly.
• Students that worked most frequently earlier in the day tended to have higher Grit Scores.
• Students that worked most frequently later in the day tended to have lower Self-Efficacy Scores.
• Students that worked most frequently later in the day may have engaged in more individual testing sessions.
• Students that worked most frequently later in the day tended to make more overall submissions.

5.3 Conclusion

Students learning to program can encounter difficulty when an unknown error is raised, especially without assistance from an instructor or peer. Immediate, formative feedback in the form of unit tests has been a solution to this issue that has been gaining recognition recently. Introducing test-driven development to students in early programming courses has the advantage of promoting its benefits when learners are most accepting of new development techniques. Finally, the importance of maintaining students’ self-confidence and motivation cannot be understated. As such, this study examined the intersection of introducing unit testing with metrics of student mental wellness, a combination that has not been explored in full yet.

This thesis sought to evaluate whether there are any notable correlations between measurements of student mental wellness and measurements of engagement and success with test-driven development. Although several weak, positive correlations were identified, this investigation could not definitively establish any strong correlations with measurements in mental wellness. However, once the students were classified by the manner in which they interacted with the TDD activities, several strong differences between groups were identified for individual survey questions. This indicated that future studies may be improved by the creation of a new, research-suitable survey involving the established questions.

Finally, though no unambiguous conclusions were drawn, several trends in the students’ grit, self-efficacy, and engagement were noted with respect to usage type and activity period. These results should provide a foundation for the design and implementation of future research into the potential mental benefits of test-driven development.
Bibliography


Bashant, J. (2014, Fall). Developing grit in our students: Why grit is such a desirable trait, and practical strategies for teachers and schools. *Journal for Leadership and Instruction, 13*(2), 14-17.


A survey was given to all participants in the Fall 2018 CIS*2430 course via an optional course Moodle activity. Respondents were asked to answer each question with respect to how they felt the statement applied to them, on a 5-point Likert scale with the following options: Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. This survey was based on the Grit-S survey by Duckworth & Quinn (2009). This appendix lists every question on the grit survey as follows:

1. New ideas and projects sometimes distract me from previous ones.
2. Setbacks don’t discourage me. I don’t give up easily.
3. I often set a goal but later choose to pursue a different one.
4. I am a hard worker.
5. I have difficulty maintaining my focus on projects that take more than a few months to complete.
6. I finish whatever I begin.
7. My interests change from year to year.
8. I am diligent. I never give up.
9. I have been obsessed with a certain idea or project for a short time but later lose interest.
10. I have overcome setbacks to conquer an important challenge.
Appendix B

Self-Efficacy Survey

A survey was given to all participants in the Fall 2018 CIS*2430 course via an optional course Moodle activity. Respondents were asked to answer each question with respect to how they felt the statement applied to them, on a 5-point Likert scale with the following options: Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. This survey was based on the Generalized Self-Efficacy Scale by Schwarzer & Jerusalem (1995). This appendix lists every question on the self-efficacy survey as follows:

1. I can always manage to solve difficult problems if I try hard enough.
2. If someone opposes me, I can find the means and ways to get what I want.
3. It is easy for me to stick to my aims and accomplish my goals.
4. I am confident that I could deal efficiently with unexpected events.
5. Thanks to my resourcefulness, I know how to handle unforeseen situations.
6. I can solve most problems if I invest the necessary effort.
7. I can remain calm when facing difficulties because I can rely on my coping abilities.
8. When I am confronted with a problem, I can usually find several solutions.
9. If I am in trouble, I can usually think of a solution.
10. I can usually handle whatever comes my way.
Appendix C

Engagement Survey

A survey was given to all participants in the Fall 2018 CIS*2430 course via an optional course Moodle activity. Respondents were asked to answer each question with respect to the level to which they felt each statement was important to them, on a 5-point Likert scale with the following options: Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. Students were instructed to read each statement as if it ended with “is important to me”. This survey was based on the Student Course Engagement Questionnaire by Handelsman et al. (2005). This appendix lists every question on the engagement survey as follows:

1. Making sure to study on a regular basis.
2. Putting forth effort.
3. Doing all the homework problems.
4. Staying up to date on readings.
5. Looking over class notes between classes.
7. Taking good notes in class.
8. Listening carefully in class.
9. Coming to class every day.
10. Finding ways to make course material relevant to my life.
11. Applying course material to my life.
12. Finding ways to make the course interesting to me.
13. Thinking about the course between classes.
14. Really desiring to learn the material.
15. Raising my hand in class.
16. Asking questions when I don’t understand the instructor.
17. Having fun in class.
18. Participating actively in small group discussions.
19. Going to the professor’s office hours for questions.
20. Helping fellow students.
21. Getting a good grade.
22. Doing well on tests.
23. Being confident in my ability to do well in the course.
Appendix D

Correlational Scatter Plots

This appendix contains the initial scatter plots of the Pearson and Spearman’s correlations used in Chapter 4. Figure D.1 and Figure D.2 provide scatter plots of each combination of metrics in the lower half, and correlation coefficients of the metrics in the upper half. Where a metric crosses with itself, a histogram is provided.
Figure D.1: Initial scatter plots of Pearson correlations.
Figure D.2: Initial scatter plots of Spearman correlations.
## Appendix E

### Survey Questions by User Profile

This appendix contains the data used in Section 4.3 to identify differentiating survey questions. The average of each survey question is provided, grouped by user profile.

<table>
<thead>
<tr>
<th>Question</th>
<th>Test-Last</th>
<th>Test-First</th>
<th>Sampler</th>
<th>No-Attempt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Mode</td>
<td>Mean</td>
<td>Mode</td>
</tr>
<tr>
<td>New ideas and projects sometimes distract me from previous ones.</td>
<td>3.64</td>
<td>4</td>
<td>3.60</td>
<td>4</td>
</tr>
<tr>
<td>Setbacks don’t discourage me. I don’t give up easily.</td>
<td>3.39</td>
<td>4</td>
<td>3.73</td>
<td>3,4</td>
</tr>
<tr>
<td>I often set a goal but later choose to pursue a different one.</td>
<td>3.00</td>
<td>2,4</td>
<td>3.47</td>
<td>4</td>
</tr>
<tr>
<td>I am a hard worker.</td>
<td>3.73</td>
<td>4</td>
<td>3.93</td>
<td>4,5</td>
</tr>
<tr>
<td>I have difficulty maintaining my focus on projects that take more than a few months to complete.</td>
<td>3.39</td>
<td>4</td>
<td>3.27</td>
<td>3</td>
</tr>
<tr>
<td>I finish whatever I begin.</td>
<td>3.24</td>
<td>3,4</td>
<td>3.13</td>
<td>3</td>
</tr>
<tr>
<td>My interests change from year to year.</td>
<td>3.18</td>
<td>4</td>
<td>3.33</td>
<td>4</td>
</tr>
<tr>
<td>I am diligent. I never give up.</td>
<td>3.39</td>
<td>4</td>
<td>3.60</td>
<td>3,4</td>
</tr>
<tr>
<td>I have been obsessed with a certain idea or project for a short time but later lost interest.</td>
<td>3.45</td>
<td>4</td>
<td>3.47</td>
<td>4</td>
</tr>
</tbody>
</table>
I have overcome setbacks to conquer an important challenge.

<table>
<thead>
<tr>
<th>Survey</th>
<th>Test-Last Mean</th>
<th>Mode</th>
<th>Test-First Mean</th>
<th>Mode</th>
<th>Sampler Mean</th>
<th>Mode</th>
<th>No-Attempt Mean</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can always manage to solve difficult problems if I try hard enough.</td>
<td>3.44</td>
<td>4</td>
<td>3.88</td>
<td>3,4</td>
<td>3.50</td>
<td>2,5</td>
<td>3.76</td>
<td>4</td>
</tr>
<tr>
<td>If someone opposes me, I can find the means and ways to get what I want.</td>
<td>3.38</td>
<td>4</td>
<td>3.38</td>
<td>3</td>
<td>3.67</td>
<td>2,4,5</td>
<td>3.06</td>
<td>3</td>
</tr>
<tr>
<td>It is easy for me to stick to my aims and accomplish my goals.</td>
<td>3.31</td>
<td>4</td>
<td>3.50</td>
<td>4</td>
<td>3.67</td>
<td>3</td>
<td>2.88</td>
<td>2,3,4</td>
</tr>
<tr>
<td>I am confident that I could deal efficiently with unexpected events.</td>
<td>3.38</td>
<td>3</td>
<td>3.50</td>
<td>3,4</td>
<td>3.33</td>
<td>4</td>
<td>3.29</td>
<td>4</td>
</tr>
<tr>
<td>Thanks to my resourcefulness, I know how to handle unforeseen situations.</td>
<td>3.56</td>
<td>4</td>
<td>3.75</td>
<td>4</td>
<td>3.33</td>
<td>4</td>
<td>3.35</td>
<td>4</td>
</tr>
<tr>
<td>I can solve most problems if I invest the necessary effort.</td>
<td>4.00</td>
<td>4</td>
<td>4.12</td>
<td>4,5</td>
<td>4.33</td>
<td>4</td>
<td>3.88</td>
<td>4</td>
</tr>
<tr>
<td>I can remain calm when facing difficulties because I can rely on my coping abilities.</td>
<td>3.31</td>
<td>3</td>
<td>3.50</td>
<td>3</td>
<td>4.00</td>
<td>5</td>
<td>3.29</td>
<td>4</td>
</tr>
<tr>
<td>When I am confronted with a problem, I can usually find several solutions.</td>
<td>3.44</td>
<td>3</td>
<td>3.75</td>
<td>4</td>
<td>4.00</td>
<td>4</td>
<td>3.24</td>
<td>4</td>
</tr>
<tr>
<td>If I am in trouble, I can usually think of a solution.</td>
<td>3.44</td>
<td>4</td>
<td>3.88</td>
<td>4</td>
<td>4.00</td>
<td>4</td>
<td>3.29</td>
<td>4</td>
</tr>
<tr>
<td>I can usually handle whatever comes my way.</td>
<td>3.44</td>
<td>3</td>
<td>3.88</td>
<td>4</td>
<td>3.67</td>
<td>4</td>
<td>3.41</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Survey</th>
<th>Test-Last Mean</th>
<th>Mode</th>
<th>Test-First Mean</th>
<th>Mode</th>
<th>Sampler Mean</th>
<th>Mode</th>
<th>No-Attempt Mean</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement Survey</td>
<td>Test-Last Mean</td>
<td>Mode</td>
<td>Test-First Mean</td>
<td>Mode</td>
<td>Sampler Mean</td>
<td>Mode</td>
<td>No-Attempt Mean</td>
<td>Mode</td>
</tr>
<tr>
<td>Making sure to study on a regular basis.</td>
<td>3.78</td>
<td>4</td>
<td>3.67</td>
<td>4</td>
<td>4.33</td>
<td>4</td>
<td>3.62</td>
<td>4</td>
</tr>
<tr>
<td>Putting forth effort.</td>
<td>4.07</td>
<td>4</td>
<td>4.22</td>
<td>4,5</td>
<td>4.67</td>
<td>5</td>
<td>4.41</td>
<td>5</td>
</tr>
<tr>
<td>Doing all the homework problems.</td>
<td>3.19</td>
<td>3</td>
<td>3.11</td>
<td>3,4</td>
<td>3.33</td>
<td>3</td>
<td>3.52</td>
<td>4</td>
</tr>
</tbody>
</table>

76
<table>
<thead>
<tr>
<th>Activity</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staying up to date on readings.</td>
<td>3.22</td>
<td>3</td>
<td>2.67</td>
<td>3</td>
<td>3.14</td>
</tr>
<tr>
<td>Looking over class notes between classes.</td>
<td>3.07</td>
<td>2,3</td>
<td>3.00</td>
<td>2.3,4</td>
<td>2.3,5</td>
</tr>
<tr>
<td>Being organized.</td>
<td>3.93</td>
<td>4</td>
<td>3.89</td>
<td>5</td>
<td>4.31</td>
</tr>
<tr>
<td>Taking good notes in class.</td>
<td>3.27</td>
<td>3</td>
<td>3.89</td>
<td>4</td>
<td>3.55</td>
</tr>
<tr>
<td>Listening carefully in class.</td>
<td>4.11</td>
<td>5</td>
<td>4.11</td>
<td>4</td>
<td>4.45</td>
</tr>
<tr>
<td>Coming to class every day.</td>
<td>3.78</td>
<td>5</td>
<td>4.33</td>
<td>4,5</td>
<td>3.83</td>
</tr>
<tr>
<td>Finding ways to make course material relevant to my life.</td>
<td>3.63</td>
<td>3</td>
<td>4.00</td>
<td>3,4,5</td>
<td>3.72</td>
</tr>
<tr>
<td>Applying course material to my life.</td>
<td>3.78</td>
<td>4</td>
<td>3.89</td>
<td>4</td>
<td>3.76</td>
</tr>
<tr>
<td>Finding ways to make the course interesting to me.</td>
<td>4.07</td>
<td>4</td>
<td>4.44</td>
<td>4</td>
<td>3.97</td>
</tr>
<tr>
<td>Thinking about the course between classes.</td>
<td>3.44</td>
<td>4</td>
<td>4.38</td>
<td>4</td>
<td>3.58</td>
</tr>
<tr>
<td>Really desiring to learn the material.</td>
<td>4.07</td>
<td>4</td>
<td>4.33</td>
<td>4,5</td>
<td>3.86</td>
</tr>
<tr>
<td>Raising my hand in class.</td>
<td>2.70</td>
<td>2,3</td>
<td>2.67</td>
<td>3</td>
<td>2.86</td>
</tr>
<tr>
<td>Asking questions when I don't understand the instructor.</td>
<td>3.74</td>
<td>4</td>
<td>4.00</td>
<td>4</td>
<td>3.69</td>
</tr>
<tr>
<td>Having fun in class.</td>
<td>3.56</td>
<td>4</td>
<td>4.00</td>
<td>4</td>
<td>3.66</td>
</tr>
<tr>
<td>Participating actively in small group discussions.</td>
<td>3.37</td>
<td>4</td>
<td>3.44</td>
<td>3</td>
<td>3.28</td>
</tr>
<tr>
<td>Going to the professor’s office hours for questions.</td>
<td>3.44</td>
<td>4</td>
<td>2.89</td>
<td>2</td>
<td>3.62</td>
</tr>
<tr>
<td>Helping fellow students.</td>
<td>4.15</td>
<td>4</td>
<td>4.00</td>
<td>3,4,5</td>
<td>4.31</td>
</tr>
<tr>
<td>Getting a good grade.</td>
<td>4.07</td>
<td>4</td>
<td>4.44</td>
<td>4</td>
<td>4.21</td>
</tr>
<tr>
<td>Doing well on tests.</td>
<td>4.07</td>
<td>4</td>
<td>4.44</td>
<td>4</td>
<td>4.14</td>
</tr>
<tr>
<td>Being confident in my ability to do well in the course.</td>
<td>4.11</td>
<td>4</td>
<td>4.44</td>
<td>5</td>
<td>4.14</td>
</tr>
</tbody>
</table>