Prediction of Student Success or Failure Without Domain Models or Formal Assessments

by

Julia Baldwin

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

© Julia Baldwin, August, 2013
ABSTRACT

PREDICTION OF STUDENT SUCCESS OR FAILURE
WITHOUT DOMAIN MODELS OR FORMAL
ASSESSMENTS

Julia Baldwin
University of Guelph, 2013

Advisor:
Dr. Judi McCuaig

Models of student knowledge provide intelligent tutoring systems (ITS) with the required information to make informed adaptations that improve student learning. Student models are often based on a domain model and can be cumbersome to build. This work demonstrates that it is possible to create a model of student interactions with a Learning Management System (LMS) that can be used to predict student success at any point during the semester without a domain model or the use of formal evaluations. This research shows that it is possible to predict student success or failure early and accurately for students who will do very poorly or very well in the course. This work describes the framework of a domain independent student model that can predict student success throughout the semester.
Acknowledgements

Many people were involved in the completion of this thesis. I’d like to thank my advisory committee for their time and effort in this work. Dr. Zoe Zhu has been very supportive and excited about my research. Edits and re-organization from Dr. Blair Nonnecke were immensely helpful and contributed to the readability and cohesion of the thesis.

I would like to thank my thesis advisor Dr. Judi McCuaig. Her support and belief in my abilities were vital to the completion and quality of this work. Judi helped me to expand my abilities in ways I didn’t know were possible, and provided her very own adaptive instruction perfectly tailored to my needs. Judi has been a wonderful advisor, friend and mentor to me throughout my degree. She has been an inspiring role model and her teachings have gone far beyond my research. I will be forever grateful for what I have learned from her not just in academics, but in life itself.

I’d like to acknowledge Josh Smith who tirelessly and voluntarily edited this thesis and patiently detailed every issue that he found.

My parents have played an important role in my journey. I’d like to thank my mother, Jane Baldwin, who was always very supportive of my education. My father, Barrow Baldwin, has been very supportive and a great person to bounce ideas off of. He has been a source of inspiration for me. Despite my best efforts, and to my great delight, we have become more
similar than I had ever expected.

My husband, Andrew Berry, played a crucial role in my degree. His unconditional love and support were integral to my success. He spent many late nights helping edit this document on a very tight schedule and was always available to discuss my ideas and to encourage me.
Contents

List of Figures ix

List of Tables x

1 Introduction 1

2 Literature Review 6

2.1 Student Modelling ................................. 7

2.1.1 Model Types .................................. 8

2.1.2 Model Creation and Maintenance ................ 14

2.1.3 Validating Student Models ..................... 17

2.2 Grade Prediction Using EDM ...................... 19

2.2.1 Grade Prediction Studies ....................... 20

2.3 Classifiers ........................................ 25

2.4 Tools for Educational Data Mining ................ 27

2.5 Data Collection and Preparation .................. 28

2.5.1 Data Selection .................................. 30
## 3 Experimental Design

3.1 Design of the Experiment ........................................... 42
3.2 Apparatus and Materials .............................................. 43
   3.2.1 Participants .................................................. 43
   3.2.2 The Course ..................................................... 43
   3.2.3 The Course Website ........................................... 44
   3.2.4 Data Mining Tool .............................................. 44
   3.2.5 Classifier type ................................................ 45
   3.2.6 Classifier Parameters ....................................... 46
   3.2.7 Classification Groups ....................................... 49
3.3 Procedure .......................................................... 50
   3.3.1 Data Collection .............................................. 51
   3.3.2 Cleaning ....................................................... 55
   3.3.3 Derived Data ................................................ 57
   3.3.4 Amalgamation of Data ..................................... 59
   3.3.5 Analysis Plan ............................................... 60

## 4 Results and Discussion

4.1 Data Exploration .................................................. 64
4.2 Individual Weeks .................................................. 66
4.3 Expanding the Data Pool ........................................... 67
4.4 Combining Weeks ........................................... 69

4.5 Analysis of Aggregated and Combined Data .............. 71

4.6 The Removal of Lab Grades ................................. 74

4.7 In-Depth Analysis ........................................... 78

4.7.1 Weekly Accuracy .......................................... 78

4.7.2 Accuracy by Category ..................................... 79

4.7.3 Identifying the Most Predictive Data ..................... 81

4.7.4 Relevant Data ............................................... 88

5 Conclusions, Limitations and Future Work .................. 91

5.1 Limitations and Future Work ................................. 95

5.1.1 Validation .................................................. 96

5.1.1.1 Validation Across Domains and LMS ............. 97

5.1.2 Suggested Improvements ................................... 98

5.1.3 Changes to the Data ....................................... 100

5.1.4 Improvements to the Self-Evaluations .................... 102

5.1.5 Automated Data Collection and Cleaning ............... 103

5.1.6 Prediction .................................................. 104

5.2 Conclusion .................................................... 105

Bibliography ...................................................... 106

A Ethics Documents .............................................. 113
A.1
Ethics Application .................................................. 113

A.2
Study Consent Form ................................................. 123

A.3
Data Collection Plan ............................................... 127

B Course Website .................................................. 131
List of Figures

3.1 Default Parameters for the j48 Tree as Seen in Weka .............. 46
3.2 Data Sources ................................................. 51
3.3 Moodle Log Data ............................................. 51
3.4 Weekly Self-Evaluation Questions ................................. 53

4.1 Spearman Correlation Chart of the Week 4 Data ................. 66
4.2 Decision Tree From Week 2 ...................................... 83
4.3 Decision Tree From Week 8 ...................................... 85
4.4 Decision Tree From Week 11 ..................................... 87

B.1 Course Information Section of Moodle ........................... 132
B.2 Discussions, Feedback and Voting Section of Moodle .......... 132
B.3 Assignments Section of Moodle .................................. 133
B.4 Quizzes Section of Moodle ....................................... 133
B.5 Lessons Section of Moodle ....................................... 134
B.6 Problem Sets and Labs Section of Moodle ...................... 135
List of Tables

2.1 Data Gathered in Knowledge Prediction Studies . . . . . . . . . 32
2.2 Data Gathered in Participant Affect Studies . . . . . . . . . . 34
2.3 Data Gathered in Grade Prediction Studies . . . . . . . . . . 36

3.1 Trials to Test the minNumObj Parameter . . . . . . . . . . 47
3.2 Participant Removal . . . . . . . . . . . . . . . . . . . . . . . 57
3.3 Data Captured By Type . . . . . . . . . . . . . . . . . . . . . 60

4.1 Weekly Data . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
4.2 Classification Accuracy for Individual Weeks . . . . . . . . . . 67
4.3 Classification Accuracy for Multiple Weeks . . . . . . . . . . 68
4.4 Classification Accuracy for Multiple Weeks Using a Median . . . 70
4.5 Classification Accuracy for Multiple Weeks Using a Mean . . . 71
4.6 Classification Accuracy by Combination Method . . . . . . . . 72
4.7 Additional Data . . . . . . . . . . . . . . . . . . . . . . . . . . 73
4.8 Comparison of Lab Changes . . . . . . . . . . . . . . . . . . . 76
4.9 Accuracy By Week Using Average Weeks and Lab Completion . 79
4.10 Accuracy By Category ............................................. 80
4.11 Data Used By the Classifier ................................. 90
Chapter 1

Introduction

In the majority of educational settings, the instructor is the most qualified person to predict whether or not a student is going to succeed. Instructors make these predictions based on their experience with the course and its material, interactions with the student, and the results of formal student evaluations. Large class sizes are increasingly common at most higher education institutions resulting in students receiving reduced or even non-existent individual attention from their instructors. As classes grow in size, it becomes less and less feasible for instructors to know each student well enough to monitor individual progress[6]. In these cases, another method of predicting student success or failure early in the course is necessary. With such prediction, teachers have the opportunity to intervene or provide specialized assistance.
Intelligent tutoring systems (ITS) are computerized learning environments that utilize information about students in the form of student models to provide adaptive instruction to students [15]. ITS can help instructors manage the difficulties presented by large class sizes, and tailor a personalized learning experience for each student by providing individualized course content. An ITS can also alert instructors to possible problems students may be having with specific course material as well as inform instructors on who may require extra help or attention. In addition to or in place of extra attention from the instructor, an ITS can help to meet the needs of at-risk students by providing extra help and resources. The needs of students who are ahead in a course can be addressed by providing enrichment material to improve their learning experience.

ITS rely on both a domain model and a student model. Domain models are comprised of information about the domain of the course, the concepts within it and the relationships between these concepts [5]. Student models within ITS are composed of a student’s knowledge of concepts within the course [5]. Both types of models are time consuming for researchers and instructors to build. Knowledge-based student models depend on a domain model to represent the concepts of the course and the links between them. Because of this dependance, student models can not transfer across domains, a separate domain and student model be created must be created for each domain of instruction.
One aspect of student knowledge that can be predicted is student success or failure within the course. Success or failure can be measured as the final grade a student will receive in the course. Identifying students in the extreme grade categories of needing enrichment materials or extra help can help ITS to target the students who will most benefit from adaptations.

A domain independent student model could offer instructors and ITS with a reusable method to obtain the knowledge required to create a tailored learning experience for students. One possible model can be constructed by utilizing characteristics of student behaviour instead of knowledge assessments. These behavioural markers can be used to develop a student model that enables the prediction of student success in the course from observations of student behaviour. To create a behavioural model, data regarding student interaction with learning materials is needed.

Learning management systems (LMS) are customarily web applications designed to provide an environment in which students are presented with course material, interactive demonstrations, and evaluations [35]. LMS capture student interactions with the system in data logs. Through the use of data mining, connections can be identified between student behaviour and student success. For example, Romero et al. used data from Moodle and a data mining algorithm to generate decision trees which predicted whether or not a student would pass or fail the course [37].

The success or failure of a student can be predicted using a student
model generated using his or her behaviour. Likewise, a LMS can capture the behaviour of an individual student. Access to a student model and the intelligence to make adaptations allows the extension of a LMS into an ITS which can adapt to student success or failure.

This thesis has several goals:

1. Show that it is possible to create a behavioural student model that is predictive of student success or failure in the course.
2. Show that it is possible to predict student success without the use of a domain model or the formal evaluations of students.
3. Identify how early in the semester an accurate prediction of student success or failure can be made.
4. Determine behaviours that are predictive of student success or failure.
5. Explore how to leverage data from multiple weeks of student interactions for use in classification.

The thesis statement for this research is that *the prediction of the success or failure of students can be achieved without the use of a domain model or formal assessments of student learning.*

The next chapter begins with a review of the literature about student models, grade prediction, and Educational Data Mining (EDM). The experimental process, setting, data collection and cleaning is described in Chapter 3. Chapter 4 discusses predictions based on individual weeks of the course and combined weeks of the course throughout the semester. The results
include an analysis of the effect of lab grade, prediction accuracy by category and an exploration of the most predictive data. Conclusions about the results and suggestions for future work in the field are discussed in Chapter 5.
Chapter 2

Literature Review

ITS require an understanding of the student in order to adapt to the student’s needs. To gain this understanding, ITS use a model of the student. A model is a representation of attributes about a particular thing, or in the case of an ITS, a model of a particular student. Student models can be used to make predictions about student learning in ITS. Traditional student models are explicit, domain based, and time-intensive to create. EDM offers a set of techniques that can create implicit models of students without the need for a domain model.

This literature review examines mechanisms for constructing student models, discusses the literature concerning the prediction of student grades, and outlines the tools and techniques used in EDM.
2.1 Student Modelling

Student models can represent information about student knowledge, goals, demographics, behaviour or affect. Student models are a critical part of ITS used to make decision about and on behalf of a student. The student’s understanding of specific course concepts can comprise a model of student knowledge. Student behaviour can contribute to models of affect as well as models of student knowledge or success. Prediction of student grade using a model of student behaviour has been researched and accomplished by Macfadyen et al., Zafra et al., Minaei-Bidgoli et al., Hung et al., Kotstiantis et al. and Romero et al. [29, 25, 35, 28, 42, 18]. The prediction of student affect using a model of student behaviour has been completed by Cocea et al. and Munoz-Organero et al. [30, 9, 8]. Weber et al. and Baena et al. have completed studies using student models of knowledge based on course concepts [3, 39].

Student models of affect use data about student behaviour and interactions with a LMS to predict a student’s state of affect. Student motivation is primarily predicted using behavioural data. Cocea et al. used multiple algorithms to classify student’s data into three different categories of motivation which composed of engaged, neutral and disengaged using data from the log files of a LMS [8].

There are several common methods researchers use to model student knowledge. Knowledge may be represented as an explicit student model of
knowledge that can be compared to the model of a known expert through the use of a structured domain model. This comparison may be in the form of either knowledge known by the student, or knowledge missing from the student’s understanding of the course. The domain model forms the basis for the representation of both the student model, as well as the expert model. Knowledge may also be estimated using an implicit student model based on student behaviours.

2.1.1 Model Types

Student knowledge can be modelled both explicitly and implicitly. Explicit modelling of student knowledge relies on a domain model which student knowledge is mapped onto. An implicit model of student knowledge relies on a model of student behaviour which infers student knowledge. Because the goal of this work is to predict student knowledge without the use of a domain model it is important to understand the differences between the model types.

Explicit student models of knowledge are based upon a structured domain model composed of course concepts. Domain models in ITS contain information about the materials that comprise the content of the course [5].

Explicit models are most frequently employed in online courses, and in a variety of fields, such as programming, chess or math [3, 39, 22]. Each concept within the model may be represented either as known, not known
or as some probability of being known by the student [3, 10]. Explicit student knowledge models represent a student’s knowledge of each concept contained within the domain of study.

Relationships between domain concepts can then be used to achieve a more complex view of student knowledge [11, 39, 3]. Concept relationships define prerequisite connections between the concepts of a course. In some models, student knowledge of concepts can be inferred based upon the dependency of one concept to another [11]. Weber et al. created a student model that contains both concepts and links to prerequisite concepts. These relationships enable the student model to make recursive knowledge assumptions regarding linked concepts [39]. Recursive assumptions of knowledge enable the student model to more comprehensively model student knowledge by assuming that prerequisite concepts of already known material must also be known.

Within explicit modelling, expert models provide researchers with a model of what a student would be expected to know to attain expertise in the given course. In the expert model, all of the knowledge within the domain is often represented as known [13]. A student’s knowledge is considered to be a subset of expert knowledge and can be measured against it [5, 13]. As such, expert models can be compared to a student model in order to determine a student’s knowledge [38]. Desmarais et al. consider knowledge to be the sum of concepts that the student comprehends within
a given domain [10]. This measurement of knowledge provides a detailed
assessment of student knowledge based on the student’s knowledge of each
concept that can be provided to an ITS.

Instead of measuring a student’s total knowledge, the researcher Elsom-
Cook models a student’s lack of knowledge as compared to the knowledge
of an expert with perturbation based modelling [13]. Perturbation based
modelling models student knowledge about the domain that is not found
within the expert model. Incorrect knowledge or understanding is called a
“bug” and helps to define both what the student understands and what the
student understands incorrectly [13]. In order to create a student model
that could be utilized within a wider range of domain models, Vassileva
proposed a student model which makes use of a combination of knowledge
measurement and perturbation modelling. The proposed model represents
information about the concepts being learned as well as bugs in the student’s
knowledge [38].

Explicit models incorporate a variety of features about the student to
model knowledge. Explicit student knowledge models contain the concepts
of the course and student interactions with them as features. Interactions
include actions such as evaluations and the viewing of content [10, 39, 3].
Weber et al. measured student knowledge based on groups of related con-
cepts within the curriculum [39]. Student knowledge was modelled us-
ing four states for each group of concepts: the visited state, the learned
state, the inferred state, and the known state. The visited state represented whether or not a student had visited informational pages within a group of concepts. The learned state contained the results of formal student evaluations, in relation to a group of concepts. The inferred state represented whether or not a group of concepts had the potential to be inferred as known based on other learned states. The known state represented whether or not the student had manually indicated that he or she had knowledge of a group of concepts.

Student knowledge can be modelled implicitly without the use of a domain model. Implicit models of student knowledge contain information about student behaviour and interactions, as well as rules for predicting student knowledge or success. Such student models do not require the creation or use of a domain model. The lack of a domain model means that implicit student models can be applied to multiple domains without the need to create a domain model. The prediction of knowledge can be measured by a single number on a scale, or by the categorization of students into categories such as an expert or a beginner [38]. These categories or scales may be created by researchers or derived from an external measurement of success or knowledge.

Many measures of student success are possible. Because final grade represents how successfully a student has completed the course, it is often used as a measurement of student success or failure. Final grade can be
represented as a linear scale or broken down into categories such as pass or fail. Implicit student models of knowledge may also be used to model student knowledge within an application, e.g., skill level is often used as a representation of student knowledge.

Implicit student models of knowledge do not model a student’s mastery of individual concepts. Instead, a broader picture is formed of student knowledge based on behaviours and interactions. These student models can be updated on the fly [19, 21], over a period of time, or after a predetermined set of interactions [22, 26]. Classification or prediction can be completed using a subset of the student’s interactions [22], using all of a student’s history [26], or by using a combination of both [19, 21, 14].

In implicit student modelling studies, features are selected from a pool of available data in order to facilitate a more accurate classification of the overall knowledge possessed by each student [14, 19, 21]. Such models often rely on the researcher’s evaluation of prior research and the features that they believe will be most predictive in the determination of student knowledge [21]. Researchers may also use prior research to determine a large set of possible features and use a formal method to find the most predictive features [19, 14].

Ghazarian et al. identified more than 70 features from related studies that could be used in the prediction of knowledge [14]. A small pilot study was completed with three novice and three skilled computer students to
explore additional features. Using observations on these subjects during a
think aloud experiment, new features with the potential to be predictive
were identified. They plotted the average value of each feature for students
of different skill levels. These plots were used to remove features with no
correlation to student skill level. This removal left 35 features which were
analyzed with an information gain ranking, which ranks features based upon
their entropy reduction property.

Hurst et al. collected a large set of possible features which they provided
to an algorithm that performed a combinatoric optimization to determine
what subset of features were most predictive of knowledge when combined
[19]. This approach took into account the problems of overfitting, where
multiple features rely on the same underlying data and do not provide
increased predictive power over using just one of the related features.

Researchers often rely on classification algorithms to determine and
make use of the most predictive variables. Hung et al., Thai Nghe et al.
and Romero et al. used a C4.5 decision tree algorithm, which determined
the most predictive features and used them to create a decision tree that
classified students [37, 31, 18]. The C4.5 algorithm uses an extension of
the information gain ranking system called gain ratio which maximizes the
information gained by partitioning the data on a given attribute [17].

Explicit models provide a greater level of depth of information about
student knowledge, while implicit behaviour based models can be easily
moved between domains. Explicit student models based on a domain model are more precise and flexible than implicit student knowledge models, which are based on student behaviour. The effectiveness of an implicit student model suffers when applied to ITS. The lack of detail of the information available about the student may reduce the effectiveness of ITS based on implicit student models [11]. Domain-based explicit student models are complex to design, as relationships between concepts which make up a model of the domain must be created [11]. The creation of a new model for a knowledge unit modelling system is both difficult and time consuming [11]. Implicit models built using student behaviour are more easily transferred between different settings and domains as they are not reliant on a domain model [11].

2.1.2 Model Creation and Maintenance

Student models are created by collecting structured data about a student or a student’s knowledge. Some models make use of bootstrapping to increase the early effectiveness of the model. Bootstrapping involves using the results of student self-evaluation questions or by using the results of domain related questions or problems posed to the student. The results of these questions can be used to initialize more of the student model by making use of stereotypes to estimate traits or knowledge of similar students. A model is initialized with either no data, data from questions given to students, or
data extrapolated from a stereotype. The model is updated as additional data are collected.

Bootstrapping helps to create an initial student model that more accurately reflects a student’s knowledge than an empty, uninitialized student model. In Baena et al.’s study of student knowledge within a chess course, the student provided a self-evaluation of his or her overall chess knowledge [3]. This self-evaluation was used to determine the initial assumption of knowledge within the student model. Similarly, Kalyuga and Sweller created a baseline assumption of initial knowledge by asking students to solve three equations and analyzing the solutions to create an initial student model [22]. Weber et al. did not initialize student overall knowledge and began with a model that represented no mastery of any content [39]. Students were then able to indicate specific concepts that were known, which filled in the more precise details in the student model. These methods provide researchers with details about each student to bootstrap to an initial student model.

Stereotyping is a method used to extrapolate an initial student model using only small amounts of information about a student [23]. Stereotypes represent a set of traits that often co-occur; each set of co-occurring traits makes up a stereotype [33]. Kay et al. asked students a series of questions related to their programming experience and knowledge [24]. Each answer helped to fill in facets of the student model. Using known information about
the student a stereotype was extrapolated that provided more information to the student model. Researchers may also use the features of an average student to initialize student models. In these cases a stereotype of an average student is created and used as the starting point for each new student model [33].

How often a student model is updated can affect how it is used by an ITS. Dynamic student models change as students complete activities and learn new information or concepts. Static student models are created after all of a student’s data have been collected, measure knowledge only once, and are not created during a student’s knowledge acquisition [26, 33]. To date, models built using EDM are typically reliant on the availability of data from an entire semester with the intent that the resulting model can be used in subsequent semesters for different students. The use of the model in future semesters assumes that cohorts of students have the same characteristic behaviours and that the original models can be used accurately when data from part of a semester is available. [29, 35, 28, 42, 18]. Dynamic student models allow for adaptations based on the student model throughout the course.

Linton et al. created a static student model to suggest appropriate instructional topics to each student [26]. The student model was not created until the student had finished using the software. Prior to the creation of the student model, data were collected for each student for between three
and eleven months. These data were stored until the end of the experiment at which point the aggregated data were used to create a student model based on the student’s expertise over a period of time.

Conversely, dynamic student models update frequently and keep track of student knowledge or predictions about student knowledge as it changes [39, 3, 24, 19, 22, 21, 38, 33]. Systems that measure student knowledge explicitly are often updated as the student moves through the course material. Models may update when the student views content, completes tasks or undergoes an evaluation [39, 3, 24]. When modelling application knowledge, Hurst et al. evaluated the student’s mouse and menu interactions to determine if an action was a novice or expert action [19]. This information was used to update the model of the student’s skill level every time an interaction occurred. The resulting dynamic model of application knowledge represented the student’s acquisition of knowledge as they gained more skill with the software.

2.1.3 Validating Student Models

Some researchers have validated their student models with varied success [22, 14, 19, 21, 26]. One method of validating the model is to evaluate the success of the adaptations that are based on the student model. Kalyuga et al. used both a control group and an experimental group to validate their student model [22]. Students in the experimental group were presented with
an adaptive ordering of the learning material based on their student model. Students in the control group were all presented with the same information with no adaptive ordering. Kalyuga et al. compared the results of the final test of each group and found that the group receiving the adaptive ordering of the course material fared better on the final test of the material than the control group. The prediction was not evaluated independently first to determine how effectively students were classified. This approach to validation is problematic because both the classification and the adaptations are being evaluated together. By validating both methods separately the accuracy or helpfulness of the model and adaptations could be individually assessed and optimized.

The accuracy of the classification of students based on a behavioural model can be validated by withholding the data of a number of students from the classification process [14, 19]. Using ten-fold cross-validation, the data are divided evenly into ten parts in which the class variable is represented evenly across each part. Each part is then excluded and used as the testing data to determine the accuracy of the classifier which is trained on the nine parts that are not excluded. This holdout process is then repeated using each of the ten-folds as the testing data [41]. The accuracy of the classifier is calculated as the average percentage of correctly classified instances of the testing data for each iteration. Using ten cross fold validation the classifier is tested on unseen data which validates the ability of the classifier to work on unknown data. Minaei-Bidgoli et al. and Thai Nghe et al. use
ten-fold cross-validation without stratification [29, 31]. Romero et al., and Zafra et al. make use of stratified ten-fold cross-validation which represents class variables evenly across all folds [42, 35]. Cross fold validation more accurately validates the student model than evaluating the adaptations stemming from it. The withholding method only measures accuracy of unknown data only in the specific set of conditions it is created. The model is not validated across domains or types of applications. The model has not been proven to work with a different set of students from a different background or who are performing different tasks within the system.

Ghazarian et al. ran a second experiment to validate the knowledge classification of students in the original experiment [14]. Two skilled users and one novice user participated in the validation experiment. The classifiers created previously were tested on the interaction data from these participants. Ghazarian et al. found that two of the three classifiers correctly classified the participants skill most of the time. This validation suffered from a lack of participants. This trial does test the classifier directly and with a new set of participants.

2.2 Grade Prediction Using EDM

In a 2009 review of the educational data mining (EDM) field, Baker et al. defined EDM as “an emerging discipline, concerned with developing methods
for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in” [4].

The EDM community has tried to provide instructors and ITS with pertinent information about student success or failure using data mining to classify students based upon a model of student behaviour. Success or failure is often measured as student final grade in the course based on a student model of grades, behaviours and interactions with the course LMS or website. Student grade is used to measure success or failure because it is indicative of how a student has done in a course and is easy to measure and compare.

EDM provides tools for creating implicit models of students that can be used to predict student success or failure in a course. The next subsection explores several EDM studies in which student final grade was predicted with the goal of understanding how these studies created models and how the prediction was achieved.

2.2.1 Grade Prediction Studies

The final grade of students can be predicted using a linear scale of final grade or by creating categories that segment the final grade. Most EDM grade prediction studies classified students into categories. Three studies classified students into two different groups. Zafra et al. and Romero et al.
both classified students based on those who would fail the course and those who would pass [42, 36]. Macfadyen et al. also categorized students into two groups; those whose grades were less than 60% and those whose grades were greater than 60%.

Some researchers used multiple category distributions. Minaei-bidgoli et al. classified students into two, three or nine categories derived from final grade [29]. Thai Nghe et al. used two, three or four categories [31]. Thai Nghe et al. found large imbalances in the distributions of the categories and found the accuracy of prediction in the smaller categories was much lower than the accuracy in the larger categories. To increase accuracy in the minority classes they used the resample function in the EDM tool Weka which oversamples the minority class and undersamples the majority class to create a more balanced distribution for training an algorithm.

Romero et al. used four classification groups which were rebalanced because they were too diverse in size [36]. To do this they used random over-sampling which involved copying randomly chosen instances from minority classes into the data set until all classes had the same number of instances. Romero et al. found that rebalancing was beneficial in most cases. The number of categories students are classified into changes both the accuracy and granularity of a final grade prediction. Categories that are more balanced often benefit from a higher accuracy across all classes.

Thai Nghe et al. completed two grade classification studies to predict
student GPA based on each student’s prior GPA and demographic information [31]. They compared classifying students onto a continuous scale, into four categories and into two sets of three categories using two different grade distributions. The M5P model tree, the BayesNet Bayesian Network and the J48 decision tree (the open source implementation of the C4.5 decision tree) were compared. They found that the J48 algorithm consistently provided the best results. Overall prediction accuracy into three categories derived from final grade using rebalanced data was 86% for the first study and 74% for the second study.

A study done by Minaei-Bidgoli et al. classified students into categories of final grade in two, three or nine categories [29]. Classification was completed using data from the entire course. Student final grade was classified with a combination of multiple classifiers optimized with a genetic algorithm with a 95% confidence interval into two classes with a 94.09%+/−2.84% accuracy rating, into three classes with a 72.13%+/−0.39% accuracy rating and into nine classes with a 62.25%+/−0.63% accuracy rating.

A study completed by Kotsiantis predicted the final grade of participants [25]. This study is the only grade prediction study discussed that created a dynamic prediction of success that changed throughout the semester. Success was predicted using data from the beginning of the semester and adding more data at multiple points to compare prediction accuracy at different times during the semester. Multiple regression techniques were
compared based upon their ability to predict final grade and M5 rules had the highest accuracy. Student final grade in the course was predicted at various points in the course using M5 rules with a mean absolute error ranging from 1.83 at the beginning of the course and 1.21 at the end of the course.

Three studies were published by Romero et al. all of which use the same data seen later in Table 2.3. The paper entitled “Web Usage Mining for Predicting Final Marks of Students That Use Moodle Courses” explored classifying students into four categories derived from final grades using data from the entire course [35]. Several different classification techniques were compared. The results from this study indicate that there is not one single algorithm will provide the best classification in all cases.

Another study by Romero et al. titled “Data Mining Algorithms to Classify Students” compared 25 EDM algorithms to predict student final grade based on data from the entire course [36]. These algorithms were selected because of their availability in Keel, the chosen data mining software package. Romero et al. determined that the best algorithms to use when using categorical data were the two decision tree algorithms; CART and C4.5. The most effective algorithms for data which were rebalanced to have equivalently sized prediction categories were found to be Corcoran, XCS, AprioriC which are rule induction algorithms and MaxLogicBoost which is a fuzzy rule learning algorithm. The best algorithms when using numerical
data were found to be CART (a decision tree algorithm), GAP (a fuzzy rule learning algorithm), GGP (a rule induction algorithm) and NNEP (a neural network).

Another study by Romero et al. titled “Data mining in course management systems: Moodle case study and tutorial” used the C4.5 decision tree algorithm to generate decision trees that predicted whether or not a student would pass or fail the course [36]. Classification was completed using data from the entire course. Data were held out to validate the classifier but classification accuracy was not discussed.

A study completed by Zafra predicted if a student would pass or fail a course using data from the entirety of the course [42]. They achieved a 74.29% accuracy of prediction using the evolutionary algorithm G3P-MI.

A study done by Macfadyen et al. predicted student final grade, in particular students at risk of failing. Data from interactions throughout the course was used [28]. Macfadyen et al. used a logistic regression model to correctly classify students into two categories 73.7% of the time.

Hung et al. determined common relationships between interaction data and used a decision tree to create a predictive model based on final grade based on course interactions throughout the entire course [18]. No measurements of accuracy were used in this exploratory study.

While the grade prediction studies discussed in this section provide a reasonably accurate grade prediction, all but one of the studies require an
entire semester worth of data to predict student grade. To make use of grade prediction in an ITS, it must be possible to make an accurate prediction throughout the semester. The studies discussed in this section commonly used an entire semester of data, classified students into between 2 and 4 groups, and used a variety of classifiers. The next section of this thesis discusses common EDM classifiers and the characteristics of each.

## 2.3 Classifiers

The classifiers considered for this research are the classifiers outlined in a recent paper by Hämäläinen et al. which reviews the use of different classifiers that are appropriate for EDM [20]. The following classifiers were discussed: decision trees, Bayesian networks, neural networks, K-nearest neighbour classifiers, support vector machines and regression based methods.

Decision trees represent a set of rules in tree form. There is one rule for each leaf node, this rule is obtained from tracing the path from the root to the leaf of the tree. The learning algorithm for a decision tree partitions the data until a terminal criteria is reached and a leaf is created. The best known algorithms for decision trees are ID3 and C4.5 [20].

Bayesian classifiers represent statistical dependencies as a graph structure. Each vertex of the graph corresponds to an attribute and is connected
to incoming edges which define which attributes it depends upon. The strength of these dependencies is defined by conditional probabilities [20].

Neural Networks are often found in pattern recognition. Feed-forward Neural Networks (FFNN) are the most used type of neural networks. FFNN consist of a layer of input nodes, a layer of output nodes and one or more layers of hidden nodes. Nodes in the hidden layers are connected to nodes in the other layers and the edges represent individual weights [20].

K-Nearest Neighbour Classifiers do not build an explicit model for the data, they approximate it both locally and implicitly. This approximation is done by classifying objects based on the class value of K similar data points [20]. Support vector machines (SVMs) concentrate on class boundaries. The goal of an SVM is to find the data points that define the class boundaries [20].

Although linear regression is not a classification method it is often used in the field of EDM. For linear regression to be used, the predictor variable must be numeric. Linear regression attempts to find the function of one attribute to another [20].

Hämäläinen et al. describe the most appropriate classification methods for different data [20]. Different classifiers are appropriate for different studies based on the data used, the number of participants, and the output desired. The prediction of categorical data are restricted to using decision trees, Bayesian networks, feed forward neural networks, nearest neighbours
classifiers and support vector machines for data mining. For particularly small data sets the best accuracy can be obtained using a naive Bayesian classifier or linear regression. Support vector machines can also be used for particularly small data sets if the parameters are properly set. For data with missing values Bayesian classifiers, FFNN and nearest neighbour models are most robust. Decision trees, Bayesian classifiers and nearest neighbour classifiers are able to handle data that is both numerical and categorical. Decision trees and Bayesian networks create the most easily understood output for humans to comprehend.

Many of the classifiers used for prediction have specifiable parameters that can be customized to optimize classification accuracy based on classifier input. In the majority of studies there was no discussion of which parameters were used or why they were selected. Kotsiantis et al. used default parameters for the algorithms used in the study in order to minimize expert bias [25]. Romero et al. used default parameters as well but did not explain the reasoning for doing so [36]. Thai Nghe et al. tuned their parameters but did not explain how or why this was done.

### 2.4 Tools for Educational Data Mining

A variety of tools have been used in the studies discussed, including custom built tools. The two tools that are used most often are Keel and Weka,
which are both open source java based EDM tools.

In a study 2010 study, Romero et al. used a custom tool built on the Keel framework [35]. Kotstiantis et al. also built their own tool for use in their research [25]. Hung et al. used both Weka and Knime for their research [18]. Zafra et al. used Weka to classify students [42]. In their 2008 Moodle case study Romero et al. use both Weka and Keel [36]. These were chosen because both software packages are free, implemented in Java and use the same external representation of the dataset (the Attribute-Relation File Format). Thai Nghe et al. began with a list of 30 data mining tools which they filtered down to three; Weka, Orange and Yale [31]. Weka was chosen due to its wider range of algorithms and support for very large data sets.

2.5 Data Collection and Preparation

Student modelling studies collect large amounts of data from student interactions with LMS. This section discusses the data collected from eleven different student modelling studies. The studies are presented in three categories based on the type of model and what is being measured:

- Explicit student knowledge modelling studies [3, 39].
- Affect prediction studies based on a model of student behaviour [30, 9, 8].
Grade prediction studies based on a model of student behaviour [29, 25, 35, 28, 42, 18].

While each study makes use of different data, there are common categories. The data found in these studies can be grouped into the following categories:

- Self-evaluations (used in 3 studies)
- Successful task completion (used in 3 studies)
- Outcome of a first attempt (used in 1 study)
- Number of total attempts (used in 4 studies)
- Correctness of scores or grades (used in 6 studies)
- Visits to a page (used in 10 studies)
- Interaction date and time (used in 5 studies)
- Time spent on a problem or page (used in 5 studies)
- Help requests (used in 2 studies)
- Order of interactions (used in 1 study)
- Demographic information (used in 1 study)

These categories provide a rich variety of information about student learning and behaviour. Self-evaluations enable researchers to ask students to provide feedback about some feature of their learning such as their confidence or perception of knowledge. Successful task completion represents when a student has completed a task successfully. The outcome of the first attempt of a task may be studied by researchers as may be the of total attempts before a task is completed successfully. Data may be gathered
about correctness of scores or grades on assignments or learning modules. Almost all studies gather data about student visits to either specific pages or a group of pages. Data about the date and time of interactions and the total time spent on problems or pages may be gathered. The number of times a student asks for help may also be calculated in some systems. In one study the order students interacted with the LMS was considered. Basic demographic information may provide details about students such as their gender or program.

2.5.1 Data Selection

The choice of what data to gather varied between researchers in the field. Each study discussed in this thesis investigated a different combination of data. It is common for researchers to not provide reasoning as to their original choice and gather what data are readily available and justify the choice with the outcome of the experiment. For example, Minaei-Bidgoli and Dráždilová gathered data based upon what data were readily available to them [29, 12]. Munoz-Organero et al. gathered information about the number and type of participant interactions to predict the motivation of students [30]. The researchers reported that the decision to gather these data was motivated by two prior studies which used this type of data to predict motivation. Zafra et al. gathered data about student interactions with quizzes, assignments and forums based upon unspecified prior research.
which they posit has shown the strength of these three activities in other studies [42]. Romero et al. gathered data from quizzes, assignments and forums [35]. These data were chosen because of its importance in the course as discussed in their prior research. Most often data selection is based on the availability of data instead of an active choice. The researchers discussed in this literature review did not provide an analysis as to whether or not the data being gathered are the best data for answering the question.

Explicit Modelling Studies
Two of the studies discussed model an explicit view of the student knowledge within a course and the concepts within it. This detailed student model enables the LMS to intervene in the student’s learning process based upon a model of the student’s knowledge of individual concepts [3, 39].

Table 2.1 displays the data gathered by each study. Neither of the explicit student modelling studies gathered data about the outcome of the first attempt at a particular task, the total number of attempts at a particular task, the interaction date and time, the order of interactions or demographic information.
A study by Baena et al. which gathered data about a student’s knowledge of chess concepts gathered data about the number of help requests by the student, successful task completion, student scores or grades as well as time spent on a particular task, page or using the system [3].

Weber et al. introduced an interactive educational system called ELMART that models a student’s knowledge of concepts in an electronic textbook [39]. They gathered data about successful item completion and visits to a page. Both studies gathered self-evaluation data from students [3, 39].

**Affect Prediction Studies**

Three of the student modelling studies use data about student behaviour for the purpose of predicting affect [30, 9, 8]. Affect refers to the student’s feelings or emotions. The affect prediction studies discussed here predict disengagement and motivation. Table 2.2 displays the data gathered by each study. The data from most commonly gathered to least is: visits to a
page, correctness, self-evaluations, total number of attempts, help requests and time spent on a problem or page. None of the affect prediction studies gathered data about successful task completion, the outcome of the first attempt, the date and time of interactions, order of interactions, or demographic information.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>self-evaluations</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Successful Task Completion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome of First Attempt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Attempts</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correctness(Scores or Grades)</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Visits to a Page</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Interaction Date and Time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help Requests</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Order of Interactions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Spent on a Problem or Page</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Demographic Information</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Data Gathered in Participant Affect Studies
Munoz-Organero et al. conducted a study that used student interaction patterns to predict the motivation of students in an online course [30]. This study gathered data about self-evaluations and the date and time of interactions. A study by Cocea et al. that predicted student disengagement within a course was the only affect prediction study that considered the number of failed attempts before a student succeeded on coursework [9]. Another study by Cocea, this time studying motivation was the only study to gather data about the time spent on a problem or page [8].

Two of the three studies in this section gathered data about correctness of activities or evaluations using scores or grades [9, 8]. All of the studies that predicted student affect gathered data about page visits [36, 9, 8].

**Grade Prediction Studies**

Six studies predicted the final grade or success of the student using a student model of behavioural data about the participant [29, 25, 35, 28, 42, 18]. Table 2.3 displays the data gathered by each study. The data from most commonly gathered to least are: visits to a page, the date and time of interactions, the total number of attempts, correctness, time spent on a problem or page, successful task completion, the outcome of a first attempt, the order of interactions and demographic information. None of the grade prediction studies discussed in this thesis use data from self-evaluations, or help request data.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>self-evaluations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Successful Task Completion</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome of First Attempt</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Number of Attempts</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correctness (Scores or Grades)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Visits to a Page</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Interaction Date and Time</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Help Requests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order of Interactions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Spent on a Problem or Page</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographic Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 2.3: Data Gathered in Grade Prediction Studies
A study by Minaei-Bidgoli et al. is the only research featured in this literature review to gather data relating to the order of interactions with the LMS or to consider the outcome of a student’s first attempt at a problem [29]. The only study referenced in this thesis to use demographic information about participants was a grade prediction study completed by Kotsiantis [25].

Studies by Minaei-Bidgoli, Kotsiantis and Romero gathered data about grades or correctness on a lesson or evaluation [29, 25, 35]. A study completed by Zafra and a study completed by Romero et al. collected data about successful item completion or the number of times an item had been successfully completed [42, 35]. A study by Macfadyen et al. and studies by Zafra et al. and Minaei-Bidgoli et al. gathered information about the time a student spent on a particular problem or page [28, 42, 29].

Five of the six researchers discussed in this section gathered information about the date and time of student interactions with the LMS [29, 35, 28, 42, 18]. All of the grade prediction studies discussed in this thesis use page visits, either to represent that a page had been visited or to count the number of times a page or certain type of page was visited in the LMS [29, 25, 35, 28, 42, 18].
2.5.2 Data Preparation

Unprocessed data gathered directly from a LMS may not be suitable for use in a student model. Many of the research papers in this area do not discuss how data are transformed from LMS databases to usable data. Among those that do, discretization is often discussed. Discretization is the process of dividing numerical data into categories in order to increase comprehensibility [36].

Aggregation is another method of making data more meaningful by using summary or aggregation operations on the data to create new variables [17]. Cocea et al. divided sessions into sequences of ten minutes or less which allowed for analysis of a combined and smaller set of interactions [8]. Hung et al. aggregated requests within a session into a single set of variables [18]. Macfadyen et al. used aggregation to create counts for how often students performed certain tasks [28]. Similarly, Dráždilová created event sets from combinations of attributes and created activity sets from the events over thirty minute time periods [12].

Data from LMS often has useless and missing data. Data cleaning is applied to remove both noise and inconsistencies in data. Missing data may be disguised if students enter incorrect values for mandatory fields when they do not want to answer a question [17].

Data cleaning is another part of data mining that is often not discussed or not detailed in depth within published research. Hung et al. cleaned
their log file by removing what they considered to be useless, irregular or missing data [18]. In order to study only normal learning events, aggregated sessions were removed if a student had stayed idle or accidentally closed the browser. Macfadyen et al. removed the data of any student who had not completed all of the coursework [28].

The anonymity of study participants is also an important part of preprocessing data. This process is called anonymizing. It is important to be careful about how data are anonymized, otherwise it may be possible to reconstruct the original participant identities. Dráždilová et al. used ID numbers to represent each student and removed IP information [12].

2.6 Summary

Student modelling plays a vital role in the intelligence of ITS. Student models that require a domain model provide in depth details about knowledge, but are neither transferable or reusable. Reusable models constructed via EDM techniques can predict student success but usually require an entire semester of data to bootstrap the modelling. An ITS system capable of dynamically modelling students without the requirement for a domain model would help to facilitate the broad adoption of ITS in educational settings.

Chapter 3 outlines the design of a study which reflects the common practices for EDM that have been outlined in this chapter. Student final grade
is often used to represent a student’s success or failure within the course. Final grade is most often aggregated into categories or less commonly used as a linear scale. To predict student final grade, the most commonly collected data were visits to a page, followed by information about the data and time of interactions. Weka, an open source standalone EDM package was the most commonly used tool to predict student final grade. The most prevalent classifiers were decision trees and regression based methods.

The results from the study show that EDM techniques can be used to create models that allow prediction of student success or failure without the use of a domain model or the results of any formal assessments. The study also shows that models can be successfully created using data from as little as one week of the semester which is a first step in understanding how to dynamically create models using EDM.
Chapter 3

Experimental Design

The thesis of this research is that the prediction of the success or failure of students can be achieved without the use of a domain model or formal assessments of student learning. This chapter describes a study using data mining techniques to predict student success or failure. Data for the study were captured and analyzed in order to determine if student success in the course could be predicted; and if so how accurately and how early it could be determined. The prediction of student success or failure was accomplished using correlations to look for relationships of interest and delving deeper with the use of data mining to identify predictive relationships.
3.1 Design of the Experiment

Relationships in the data were explored by investigating correlations using the Spearman correlation. This investigating suggested an experiment into the predictive relationships of the variables and the students final grades using the C4.5 decision tree algorithm.

To explore how to leverage data from multiple weeks of student interaction, the following three questions were proposed for examination through experimentation:

- Question 1: Can an individual week’s worth of data accurately predict student success?
- Question 2: Can an aggregation of multiple weeks’ worth of data accurately predict student success?
- Question 3: Can a combination of multiple weeks’ worth of data accurately predict student success?

Question 1 was addressed by using an individual week’s worth of data as discussed in Section 3.1.4 and examining each individual week using a decision tree to determine how accurately student success could be predicted using only a single week’s worth of data. Question 2 was addressed by using both the data from each week and prior weeks. These data were mined using decision trees to determine how accurately student success could be predicted using multiple weeks’ worth of data. Question 3 was addressed
by mathematically combining the data of the week being studied and each week prior. The data from the combined weeks were examined using a decision tree to determine how accurately student success can be predicted using combined weeks of data. The detailed results from this experiment are presented in Chapter 4.

3.2 Apparatus and Materials

3.2.1 Participants

The data for this study were collected during the Winter 2011 offering of the course CIS*2500: Intermediate Programming at the University of Guelph. Data were collected from 152 participants. This is a first year, second semester programming course and as such, the majority of students were in their second semester and most were computer science, engineering and science majors. Students were encouraged to participate in the study with a drawing of two fifty dollar gift certificates at a local electronics retailer.

3.2.2 The Course

CIS*2500 consisted of three one hour lectures per week, over a 12 week semester. Eleven weeks of data were collected as the final week of the semester did not include labs or a full set of classes. Grades for CIS*2500
were weighted heavily towards assignments and labs. Students completed 11 labs worth a combined total of 20% towards their final grade. Each lab was individually valued at 2% and students were allowed to drop their lowest lab mark. Students also completed four assignments worth a combined total of 30% of their final grade. A classroom response system utilizing small wireless devices called clickers allowed students to answer questions in class. Marks were assigned to students using student participation in answering the clicker questions posed during class. These marks accounted for a combined total of 10% of the student’s final grade. Students completed four formal assessments: three quizzes worth a combined total of 20% and a final exam worth 20%.

3.2.3 The Course Website

In addition to class lectures, the students of CIS*2500 used Moodle, an open source LMS accessible to students via the web [1]. For details of the course website, please visit appendix B.

3.2.4 Data Mining Tool

Three data mining software packages were considered to use for the classification of students based upon student final grade. Rattle [40], an open source R add-on used for data mining was considered but discarded due to
its inability to batch process data, lack of options and lack of documentation. Next considered were Keel [2], and Weka [16], both of which are standalone data mining packages. Both packages provided more functionality and algorithms than Rattle as well as the batch processing of data. Weka was chosen for this research due to its prominence in the EDM field, the high quality of documentation available and the suite’s stability.

### 3.2.5 Classifier type

This research required a classifier in order to classify students based on their success or failure. Regression based methods and feed forward neural networks were dismissed, as these methods do not support both numerical and categorical data [20]. Decision trees were the most commonly used classifier found in other studies and were often used by prominent researchers in the field [36]. Decision trees were chosen for use in this analysis due to their prominence in similar research, their appropriateness for the available data and their simplicity and flexibility [20].

There are multiple implementations of decision trees available. Due to its popularity in similar research and the improvements made over the ID3 algorithm, the C4.5 implementation was selected for use in this research. The Java implementation of the C4.5 decision tree, the open source J48 implementation was used for classification[27].
3.2.6 Classifier Parameters

The J48 implementation of the C4.5 decision tree has many parameters that determine exactly how the tree is built. For example, a parameter can be set which determines the minimum number of instances that are required for any leaf in the tree. The default parameters for the j48 decision tree can be seen in Figure 3.1.

![Figure 3.1: Default Parameters for the j48 Tree as Seen in Weka](image)

Figure 3.1: Default Parameters for the j48 Tree as Seen in Weka
In order to determine the most effective parameters to use, each parameter needed to be tested. Parameters were changed one at a time by either toggling between true or false or increasing and decreasing the parameter while holding all other parameters to the default setting for each week’s set of data. The average of the classification accuracy from all of the weeks with the single parameter changed was then compared to the average of the classification accuracy of all of the weeks using only the default parameters provided by Weka. Each parameter change which caused an increase in the average classification rate of correctly classifying students based on their final grade was selected for further study. For example when testing the minNumObj parameter the values found in Table 3.1 were use for each trial.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>binarySplits</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>ConfidenceFactor</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>debug</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>minNumObj</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>numFolds</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>reducedErrorPruning</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>SaveInstanceData</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>seed</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>subtreeRaising</td>
<td>true</td>
<td>true</td>
<td>true</td>
</tr>
<tr>
<td>unpunmed</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>useLaplace</td>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
</tbody>
</table>

Table 3.1: Trials to Test the minNumObj Parameter

Modifying the following parameters were found to cause an increase in accuracy prediction and were subsequently selected for further study:

47
• Confidence.
• Minimum number of instances per leaf.
• The performance of subtree raising.

The confidence limit for tree pruning attempts to forecast and account for the variance of data within the decision tree. Confidence may vary between 0 and 0.5. The j48 algorithm will not perform any modifications if the confidence value is greater than 0.5. A lower confidence value will result in more aggressive pruning and a more generalized tree while a larger confidence value will result in less pruning and a more specific tree.

The minimum number of instances per leaf dictates the smallest number of instances that a leaf can be composed of. This parameter may be increased to cope with noisy data. A higher minimum number of instances per leaf will result in a smaller decision tree.

Subtree raising is a type of post pruning. When post pruning is used the decision tree is pruned after creation is complete instead of during creation. This method of pruning requires computational work to create the classifier because a tree must first be built and then pruned. Subtree raising in particular does exactly as the name suggests; nodes are be moved upwards by replacing their parent node. Of the two post pruning methods of J48 trees subtree raising is the most computationally intensive [41].

To find an appropriate confidence value, confidence was tested in steps of 0.05 from 0.05 to 0.5 on all weeks of data. The accuracy of prediction
was found to be highest at 0.05 and 0.2. Given that a confidence of 0.05 would incur aggressive pruning and a very general tree, a confidence of 0.2 was selected. Using a confidence value of 0.2, the minimum number of instances per leaf was tested from 1 to 5 using steps of size 1. The highest prediction rate was found when the minimum number of objects was 3. Using these new parameters, the use of sub tree raising increased prediction slightly as compared to not using it. Given the small increase and the lack of issues with performance, sub tree raising was used. This process of parameter testing led to the finalized parameters for data mining with the J48 implementation. These parameters were:

- Confidence = 0.2.
- Minimum number of instances per leaf = 3.
- The performance of sub tree raising = true.

These parameters were combined with the following default parameters found in 3.1 which had previously been tested and found not to lead in an increase in accuracy as a result of their modification.

### 3.2.7 Classification Groups

The J48 decision tree was trained on the student data using Weka [16]. A classifier requires categories in which to place the individuals being classified. Students were categorized into three groups based on final grade. The three categories were: *exceeds expectations*, *meets expectations* and *failed*
expectations. These categories allow for the testing of a classifier’s accuracy in identifying students at the top and bottom of the academic spectrum in a class. It is those students who most often need instructor attention either for enrichment or extra help and who may not get the attention they need in a large class. Students with marks of 80 or higher were placed in the exceeds expectations category, students with marks of between 50 and 79 were placed in the meets expectations category and students with marks lower than 50 were placed in the fails expectations group. The use of these three categories had the additional advantage that each category was balanced with a similar number of participants in each group. Very few students received marks of B, C or D and the three category solution allowed the categories to be somewhat equal in size. Using this breakdown there were 49 students placed into the exceeds expectations category, 41 students into the meets expectations category and 28 students into the fails expectations category.

### 3.3 Procedure

Data for this experiment were collected from various sources. The data were then cleaned and amalgamated into a usable format for analysis.
3.3.1 Data Collection

Moodle log data were the primary data source for this research. However, data from other sources as listed in Figure 3.2 were also collected in order to permit the investigation of as many potentially predictive relationships as possible.

1. Moodle interaction logs
2. weekly questions
3. clicker data
4. timing data
5. grades spreadsheet
6. demographic data

Figure 3.2: Data Sources

1. Moodle interaction logs recorded all visits by the students to the LMS. These logs consisted of a list of the pages each student had visited and at what time the visit occurred. An example of these logs can be seen in Figure 3.3.

Figure 3.3: Moodle Log Data
2. Weekly questions consisted of a short self assessment and quiz which students were to complete each week. Students were not evaluated on their answers to the weekly questions sets but students were required to complete the problem set in order to have access to the weekly lab’s description. The quiz consisted of seven questions related to the subject material of CIS*2500 which were randomly selected from a test bank. Two easy questions, three medium questions and two difficult questions were presented to each student. Students were given the option to select “I haven’t learned this yet” in order to prevent guessing. These questions were drawn from the entire set of topics for the course as opposed to being drawn from the restricted set of topics covered to date.

As part of the weekly questions students were asked to self evaluate their expertise and confidence in three different ways. They were asked to rank their level of C programming expertise, their confidence in their ability to program and to estimate their letter grade on the final exam if they were to take it right away. The exact format of these questions and answers is found in Figure 3.4
Figure 3.4: Weekly Self-Evaluation Questions

Please rank your level of C programming expertise on a scale of 1 to 10.

- 1 represents no knowledge of C programming
- 3 represents complete understanding of the C programming theory and skills from CIS 1500
- 10 represents complete understanding of C programming theory and skills required for CIS 2500

Answer: __________

If you were to take the CIS 2500 final exam right now what letter grade do you believe you would achieve?

Select one:
- a. A+
- b. A
- c. A−
- d. B+
- e. B
- f. B−
- g. C+
- h. C
- i. C−
- j. D+
- k. D
- l. D−
- m. F

Please rank yourself on a scale of 1 to 10 with 1 being no confidence in your ability to program in C and 10 being complete confidence in your ability to program in C and meet challenges given to you in CIS 2500.

Answer: __________

Figure 3.4: Weekly Self-Evaluation Questions
3. Clickers were used in class at least once a week at times when the
professor of the course felt that in class participation would be constructive to the lesson being taught. Students were given a participation grade of either one or zero for using their clickers which was used to represent attendance in this research.

The length of time taken to complete exams and quizzes (the timing data) were collected manually. All of the students in the exam room started the quiz or exam at the same time which was recorded. The time each student handed in his or her quiz or exam was recorded on the exam and later recorded into a spreadsheet. Some students wrote exams at the Center for Student with Disabilities, and as such no exam times were recorded for them.

Grades were collected from the instructor’s gradesheets and consisted of 11 weekly lab grades and the final grade in the course. Assignment and quiz grades were not used in order to avoid using evaluative data to classify student success or failure.

The class list contained a list of the students in the course and their contact information. In this list could be found their basic demographic information consisting of their gender, semester level and program, as well as their minor if they had declared one. The basic class list was provided by the university and augmented with the gender of students as they became known to the instructor.
3.3.2 Cleaning

The data were cleaned prior to analysis to ensure that outliers and erroneous data did not unduly impact the results of the analysis. The data were anonymized to ensure compliance with the commitment of anonymity that had been given to the ethics board. Cleaning and anonymizing consisted of the following steps:

- Removal of identifying data
- Abstraction
- Reparation of data entry errors
- Removal of outliers

The data were sanitized by removing any identifying information about the students. Participant names and emails were removed from all collected data. A randomly generated participant identifier (PID) was assigned to each student who had consented to the study and kept in a map file. Using this map file student identification numbers in all data were replaced with each student’s unique participant identifier. The data of 28 students who did not complete the course and did not have a final grade were also removed. Once this process was complete the map file was deleted leaving no link between a student and their participant ID.

The ethics approval for this research found in Appendix A required the abstraction of all time and date information. Precise time and date
stamps were converted to the nearest sixth of the day and fell into one of six categories:

- Early Morning (4-7:59 am)
- Morning (8-11:59am)
- Afternoon (Noon-3:59 pm)
- Evening (4-7:59 pm)
- Night (8-11:59 pm)
- Late Night (midnight-3:59 am).

The ethics approval for this research also required that final grades be converted into letter grades. This abstraction was completed so that someone who knew exactly when a student had performed a task or a grade they had received could not use that information to identify the student.

The data of two students were removed from the data set due to their unusable self-evaluations. For example, one student entered “12345” when a number between 0 and 10 was requested. One student entered interpretable self-evaluations such as “7-8”. Data such as this that could be interpreted were manually repaired to something more meaningful instead of being removed, in the case of the input “7-8” the self-evaluation was changed to “7.5”.

Data that were clearly out of range or incorrect were removed or repaired as necessary. The data from one participant was removed entirely because
the number of website visits was dramatically out of range. This participant’s weekly website visits often doubled that of the next highest number of visits. The problem set was expected to take less than ten minutes for a student to complete, however, multiple students took several days to complete it. Problem set times ranged from 1 minute to an entire week; any problem set time above 60 minutes was adjusted to 60 minutes. There were 118 participants remaining after the removal of students who had no final grade, had unusable input and the student who had an excessive number of website visits as shown in Table 3.2.

<table>
<thead>
<tr>
<th>Reason</th>
<th># of Participants Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student did not complete the course</td>
<td>28</td>
</tr>
<tr>
<td>Unusable self-evaluations</td>
<td>2</td>
</tr>
<tr>
<td>Excessive number of website visits</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.2: Participant Removal

### 3.3.3 Derived Data

Raw data were used to derive more meaningful counts and metrics that could be used by data mining algorithm. Derived data were generated from Moodle interaction logs, clicker data, timing data, and weekly questions.

Assignment viewing dates, the total number of weekly visits and the number of days each week a student was active on the website were derived from the Moodle interaction logs. Each time a student visited a page on the website, it was counted as a single interaction. The different actions
a user could take were not used as data for this study. The first time students viewed an assignment was located and compared to the due date. The viewing date and the due date were subtracted to determine how many days before the assignment was due the student viewed the details of the assignment. The total number of visits to the website each week was calculated. The number of days of the week the student visited at least once was used to create a days active value.

Attendance scores were derived from the clicker data. During each week the clickers were used in one to three classes. Student participant was a binary count for each class clickers were used in, creating an attendance score of between one and three. These scores were normalized so that the attendance score of each week was out of a total score of one. In weeks where the clickers were only used once a student would either receive an attendance score of zero or one. In weeks where clickers were used three times students would have an attendance score of either zero, one third, two thirds or one.

Problem set scores, self-evaluation scores, the time taken to complete the problem set and the date students first viewed their labs were derived from the problem set data. Problem sets were designed to provide a snapshot of the student’s knowledge of course material at the time they were taken. Given that the problem set represented what a student should know at the end of the course, it was expected students would do poorly at the beginning
of the course and would perform better as they learned the material. A problem set score out of 7 was calculated. Completion of the problem set was required before students could view their lab, as such the date the problem set was completed was considered the first day a student viewed their lab and was used to calculate how early students had viewed their labs.

### 3.3.4 Amalgamation of Data

To prepare for analysis; the cleaned data were amalgamated into sets of data consisting of seven days worth of data. The weeks corresponded to each week of the course. Saturday-Friday was considered to be one week of the course because labs were due on Fridays. Each individual set of data are referred to as a week of data for the remainder of this document. Weeks were the finest granularity of data collection as many events such as labs and self-evaluations occurred each week. Other events occurred in specific weeks such as manually collected quiz timings and assignment viewing dates. Occasional data were placed in the corresponding weeks.

Some data, such as the self-evaluations and weekly website visit counts were included each week. Other data were not available weekly and were included only for the relevant week. Quiz timing data were available in weeks four and seven. Assignment viewing dates were available in weeks one, three, six and ten. Demographic data, such as gender and program,
were included in each week to provide additional information about the students. Demographic data were consistent across all weeks. Some data such as the time taken to complete the final exam did not fit properly in any given week and were not used.

The amalgamation and creation of derived variables resulted in the data found in Table 3.3 as data available in each week, data that were not available in all weeks and were only available in some weeks and static data that stayed the same throughout the semester.

<table>
<thead>
<tr>
<th>Weekly Data</th>
<th>Occasional Data</th>
<th>Static Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekly visits</td>
<td>Time taken to complete quizzes</td>
<td>Final grade</td>
</tr>
<tr>
<td>Days active</td>
<td>Time taken to complete final exam</td>
<td>Student gender</td>
</tr>
<tr>
<td>Attendance</td>
<td>Date assignment viewed</td>
<td>Student semester level</td>
</tr>
<tr>
<td>Problem set score</td>
<td></td>
<td>Student program</td>
</tr>
<tr>
<td>Time taken to complete problem set</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date lab viewed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>self-evaluation of confidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>self-evaluation of expertise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>self-evaluation of final grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab mark</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Data Captured By Type

### 3.3.5 Analysis Plan

In order to understand the data, and thus ensure that the most relevant data were given to the data mining algorithms, an analysis of the correlation between the different pieces of data was conducted. Exploratory work was completed using the gathered data in order to determine if any predictive
relationships could be found within the data. This was completed using Rattle [40], an add-on for the stats package R. The variables from each week of data were compared to each other using the Spearman correlation method. The Spearman method was chosen due to the fact that it does not assume a normal distribution and can handle ordinal data. In order to comply with the ethics board at the University of Guelph, final grade data were abstracted into letter grades. To obtain a numerical value which could then be used in correlations, the midpoint grade of each letter category was used in place of the final grade of each student.

In order to predict the success or failure of students through the use of a classifier as early in the semester as possible, it was necessary to determine the best way to use the data from prior weeks to predict early success or failure. For this purpose, three questions about classifying the data were addressed. These were:

- Question 1: Can an individual week’s worth of data accurately predict student success?
- Question 2: Can an aggregation of multiple weeks’ worth of data accurately predict student success?
- Question 3: Can a combination of multiple weeks’ worth of data accurately predict student success?

These three experimental questions were addressed individually using Weka, the C4.5 algorithm and the cleaned, amalgamated data. Classification ac-
accuracy was calculated using ten-fold stratified cross-validation. The average classification accuracy was then determined using the repeated process which provided ten classifier accuracy measurements.

The accuracy of data mining algorithms can be measured using either a measure of precision or a measure of recall. Precision is measured separately for each category. It is a measure based on the number of instances that are correct given the set of instances returned by the classifier. Recall is also measured separately for each category. It is a measure based on the number of correctly classified instances given the total number of correct instances in the entire data set. Precision and recall are usually reported as percentages. For example, when considering students who are classified by the classifier as exceeds expectations students:

- Precision would be calculated as the number of exceeds expectation students classified correctly as exceeds expectations divided by the total number of students classified as exceeds expectations.
- Recall would be calculated as the number of exceeds expectations students classified correctly classified as exceeds expectations divided by the total number of exceeds expectations students.

For the purposes of this research accuracy is reported using the measure of recall.
Chapter 4

Results and Discussion

Interaction data, demographic data, grades and self-evaluations from 118 students in a second year computer science class were used to determine if it is possible to predict the success or failure of students. This chapter outlines the results of data mining. The data from each week used for this data mining are found in Table 4.1. First, an exploration of the correlations between the various pieces of data is discussed. Then the results of applying the C4.5 algorithm to individual weeks as well as multiple weeks are outlined. The removal of lab grade and multiple methods of doing so are then discussed. Finally an in-depth analysis is completed that discusses accuracy by category, the decision trees resulting from the C4.5 decision tree and a list of the most predictive data.

This work shows that prediction is possible as early as the 7th week
of the semester with an accuracy of over 70% without the use of the results from formal assessments. The accuracy of prediction varies by class, throughout the semester students are classified into the exceeds expectations with an average accuracy of 78.4%, into the meets expectations category with an average accuracy of 47.6% and into the fails expectations category with an average accuracy of 72.5%.

### 4.1 Data Exploration

The data were explored using Spearman correlations to help identify interesting relationships between the different elements of data. Figure 4.1 shows

<table>
<thead>
<tr>
<th>Data</th>
<th>W1</th>
<th>W2</th>
<th>W3</th>
<th>W4</th>
<th>W5</th>
<th>W6</th>
<th>W7</th>
<th>W8</th>
<th>W9</th>
<th>W10</th>
<th>W11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Gender</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Program</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Semlevel</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PSCompleted</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PSTime</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PSScore</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Expertise</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FinalConfidence</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Confidence</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>LabViewed</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DaysActive</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>TotalVisits</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>LabGrade</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>DateAssignmentViewed</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>QuizTiming</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FinalGrade(category)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4.1: Weekly Data
the correlations found when the data for week 4 were explored. In the figure each variable is found along both the X and Y axis. Positive correlation is represented by the intensity of the blue and the slant to the right while a negative correlation is represented by the intensity of the red and the slant to the left. As can be seen for week 4, the strongest positive correlation was between student’s self reports of expertise and confidence, while attendance and semester level were found to be most negatively correlated.

During week 4, the self-evaluation of expertise, lab grade and number of days active were most correlated to the final grade. Many positive correlations were found with final grade. All variables except the semester level of a student and the date they viewed their assignment were found to be positively correlated to the final grade. Exploration of the data from other weeks showed similar correlations. The exploration of correlations provided evidence that much of the data collected were correlated to final grade and suitable for further investigation of predictive relationships in this thesis. Because the correlations were used in an exploratory way with no intention of using correlation as final results, no calculation of significance or in-depth analysis was done for each week. The results found using the Spearman correlations were enough encouragement to spur an investigation into prediction.
### 4.2 Individual Weeks

Question one asked if an individual week’s worth of data could accurately predict student success. To answer this question, classification was done on each week individually. Each single week of data were used to predict ...
student success into each of the three categories of *exceeds expectations, meets expectations and fails expectations*. Table 4.2 shows the number of correctly classified students by final grade category in each individual week using a C4.5 decision tree and stratified ten-fold cross-validation.

<table>
<thead>
<tr>
<th>Week</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.11</td>
</tr>
<tr>
<td>2</td>
<td>62.58</td>
</tr>
<tr>
<td>3</td>
<td>55.08</td>
</tr>
<tr>
<td>4</td>
<td>61.59</td>
</tr>
<tr>
<td>5</td>
<td>61.52</td>
</tr>
<tr>
<td>6</td>
<td>59.92</td>
</tr>
<tr>
<td>7</td>
<td>58.26</td>
</tr>
<tr>
<td>8</td>
<td>67.88</td>
</tr>
<tr>
<td>9</td>
<td>70.30</td>
</tr>
<tr>
<td>10</td>
<td>68.71</td>
</tr>
<tr>
<td>11</td>
<td>53.33</td>
</tr>
<tr>
<td>Average</td>
<td>60.21</td>
</tr>
</tbody>
</table>

Table 4.2: Classification Accuracy for Individual Weeks

The accuracy of prediction using individual weeks ranged from 43.11% in week 1 to 70.30% in week 9. The average accuracy of classifying students using an individual week of data was 60.21%.

### 4.3 Expanding the Data Pool

Question two asked if an aggregation of multiple weeks’ worth of data could accurately predict student success. To answer this question, weekly data remained separate while utilizing the data pool available from each new week. For this classification, the data from all weeks prior to the week being
studied were included. For example, for week 4 the data included weekly website visits from week 1, week 2, week 3, and week 4. Each variable from each subsection was treated as a separate variable. The data used for classification were the data from the week being studied and all prior weeks as well as the final grade category of each student. Table 4.3 shows the number of correctly classified students by final grade category using the data from multiple weeks, a C4.5 decision tree, and stratified ten-fold cross-validation. In the early weeks, classification accuracy using multiple weeks was less than that obtained using the latest week individually. Accuracy was lower for the combined for weeks 1 to 2, weeks 1 to 3 and weeks 1 to 4 as compared to the individual weeks of 2,3, and 4 respectively. For all other weeks the additional data provided an increase in classification accuracy. The average accuracy classifying with an aggregation of the weeks was 63.43% for multiple weeks, an increase of 3.22% over the classification accuracy of individual weeks.

<table>
<thead>
<tr>
<th>Weeks 1 to X</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>55.83</td>
</tr>
<tr>
<td>3</td>
<td>50.76</td>
</tr>
<tr>
<td>4</td>
<td>54.02</td>
</tr>
<tr>
<td>5</td>
<td>62.65</td>
</tr>
<tr>
<td>6</td>
<td>65.30</td>
</tr>
<tr>
<td>7</td>
<td>65.23</td>
</tr>
<tr>
<td>8</td>
<td>66.06</td>
</tr>
<tr>
<td>9</td>
<td>66.17</td>
</tr>
<tr>
<td>10</td>
<td>72.05</td>
</tr>
<tr>
<td>11</td>
<td>76.21</td>
</tr>
<tr>
<td>Average</td>
<td>63.43</td>
</tr>
</tbody>
</table>

Table 4.3: Classification Accuracy for Multiple Weeks
4.4 Combining Weeks

Question three asked if a combination of multiple weeks’ worth of data could accurately predict student success. To answer this question, central data values were calculated for each variable using the individual values for the current week and previous weeks. The central tendency was evaluated using both mean and median.

All prior week’s data were combined into a single variable. For example, when classifying week 4 the average or median of the total number of website visits from weeks 1, 2, 3, and 4 was calculated. Each variable from each week leading up to and including the week being studied were combined in two different ways, using the mean and the median to determine an appropriate method of combination. Students were classified using the median information of the week being studied and all prior weeks as well as their final grade category of *exceeds expectations, meets expectations or fails expectations*. Table 4.4 outlines the classification accuracy that was found using the median data, a C4.5 decision tree and stratified ten-fold cross-validation.

Similar to multiple week’s classification accuracy, the accuracy using combined weeks and a median was less than that of individual weeks early in the semester. Accuracy was lower for weeks 1 to 2 and weeks 1 to 3. For the remaining weeks using the median proved to be more accurate than both single weeks and multiple weeks. The average classification accuracy
Combination of weeks 1 to X | Accuracy(%)  
--- | ---  
2 | 60.76  
3 | 54.85  
4 | 66.74  
5 | 65.08  
6 | 71.06  
7 | 67.35  
8 | 78.71  
9 | 75.16  
10 | 73.56  
11 | 77.88  
Average | 69.16  

Table 4.4: Classification Accuracy for Multiple Weeks Using a Median

using data combined using a median was 69.16%, an increase of 8.95% over the average for single weeks and an increase of 5.73% over the average classification accuracy using multiple weeks.

Next students were classified using the mean information of the week being studied and all prior weeks as well as their final grade category of *exceeds expectations, meets expectations or fails expectations*. Table 4.5 outlines the classification accuracy that was found using the mean data, a C4.5 decision tree and stratified ten-fold cross-validation.

Classification accuracy was highest in all but four of the weeks investigated using multiple weeks’ worth of data combined with a mean. In weeks 1 to 2 single week classification provided higher accuracy and the median method provided the same accuracy. This is due to the fact that the mean and median of only two numbers is always equivalent. In weeks 1 to 4 the classification accuracy using the mean was outperformed by both median
and single week methods. In weeks 1 to 6 and weeks 1 to 8 the median data had a higher classification accuracy. Classifying students using combined weeks using the mean yielded an average classification accuracy of 71.29%, an increase of 11.08% over using single weeks to classify, an increase of 7.86% over using multiple weeks and an increase of 2.13% over using the median of previous weeks.

4.5 Analysis of Aggregated and Combined Data

The classification accuracies of individual weeks, aggregations of weeks, and combinations of weeks using a median as well as a mean have been determined. The accuracy of classification is important when selecting a method to be used by an ITS. Table 4.6 shows the percentage of correctly
classified students in each week or combinations of weeks using each method.

<table>
<thead>
<tr>
<th>Weeks</th>
<th>Single Week Acc.</th>
<th>Aggregated Week Acc.</th>
<th>Acc. w/ a Median</th>
<th>Acc. w/ a Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>62.58</td>
<td>55.83</td>
<td>60.76</td>
<td>60.76</td>
</tr>
<tr>
<td>3</td>
<td>55.08</td>
<td>50.76</td>
<td>54.85</td>
<td>58.41</td>
</tr>
<tr>
<td>4</td>
<td>61.59</td>
<td>54.02</td>
<td>66.74</td>
<td>56.74</td>
</tr>
<tr>
<td>5</td>
<td>61.52</td>
<td>62.66</td>
<td>65.08</td>
<td>72.73</td>
</tr>
<tr>
<td>6</td>
<td>59.92</td>
<td>65.30</td>
<td>71.06</td>
<td>70.98</td>
</tr>
<tr>
<td>7</td>
<td>58.26</td>
<td>65.28</td>
<td>67.35</td>
<td>78.64</td>
</tr>
<tr>
<td>8</td>
<td>67.88</td>
<td>66.06</td>
<td>79.17</td>
<td>72.66</td>
</tr>
<tr>
<td>9</td>
<td>70.30</td>
<td>66.14</td>
<td>75.15</td>
<td>79.62</td>
</tr>
<tr>
<td>10</td>
<td>68.71</td>
<td>72.05</td>
<td>73.56</td>
<td>81.14</td>
</tr>
<tr>
<td>11</td>
<td>53.33</td>
<td>76.21</td>
<td>77.88</td>
<td>81.21</td>
</tr>
<tr>
<td>Average</td>
<td>60.21</td>
<td>63.42</td>
<td>69.16</td>
<td>71.29</td>
</tr>
</tbody>
</table>

Table 4.6: Classification Accuracy by Combination Method

As can be seen in Table 4.6, the lowest average classification accuracy of 60.21% is found using single weeks. This accuracy does not come as a surprise as the classifier was provided with only one week of data which means there were less available data than when data from prior weeks were included. Using the single week method, it was not expected that accuracy would increase substantially week by week as prior week data were not included and therefore the data pool was not enlarged in subsequent weeks. Weeks 2, 4, 5, 8, 9 and 10 all provided a greater classification accuracy than 60%. As can be seen in Table 4.7, many of these weeks had additional data. These additional data were made up of either the length of time taken to write a quiz or how early the student viewed the assignment before it was due. Of the weeks with a higher accuracy than 60%, only weeks 10 and
11 did not benefit from additional information. One difference between the weeks that might explain why classification accuracy went up through the semester could be that students were able improve the accuracy of their self-evaluations throughout the semester. One could also theorize that student behaviour is exaggerated later in the semester, a student who does not check the course website much early in the semester may be even less likely to visit as the semester goes on.

<table>
<thead>
<tr>
<th>Week</th>
<th>Quiz Timing</th>
<th>Assignment Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Table 4.7: Additional Data

Using an aggregation of weeks provided a small increase in average accuracy to 63.42%. Given the increase in available data, it seemed likely that accuracy would be higher using aggregated data. The aggregation of weeks outperforms individual weeks in all weeks but week 2. It also seemed likely that as the semester went on and more information was available to the classifier, classification accuracy would increase. This expected increase in accuracy was the observed trend when using the aggregated data.
Combined weeks yielded a larger increase in classification accuracy than aggregated weeks. Classification accuracy was at its maximum using the mean value for each attribute from prior weeks. Classification accuracy using the mean was the highest of all methods 6 out of 10 times. Students could be classified into three final grade pool classes with over 70% accuracy each and every week after the first four weeks using the mean value for each attribute of prior weeks.

With the preliminary analysis it was concluded that further analysis into research question 3 showed the most promise for deeper exploration. Furthermore it was determined that the combination method of using the mean was most suitable for exploration.

4.6 The Removal of Lab Grades

One of the novel components of this research is the thesis that predictions can be made about student success without the use of formal assessments. The initial classifications of the data allowed the classifier to use the grade students received on their weekly labs. The completion of labs was representative of the student’s activity level in the course and at 2% a lab the grade percentage was low for each lab. When individual weeks were being considered, the 2% of the final grade per week that overlapped between lab grade and final grade was not a significant issue. With the use of averages,
the lab grade average represented 20% of a student’s final grade by the final week. When multiple weeks are considered, the use of lab grades does not allow predictions to be made about student success without the use of formal assessments. Three types of modifications were considered to address the overlap between the lab and final grades:

- Complete removal of lab data from the data provided to the classifier.
- Removal of the lab grade from the final grade.
- Usage of lab completion metrics in place of lab grades.

Using each of these three methods the classifier was re-created using a C4.5 decision tree and ten-fold stratified cross-validation for each week. The classification accuracy using this new data was compared to the accuracy of the unmodified case. The unmodified case used the averaged weeks with lab grade included in the final grade and unmodified lab marks provided to the classifier. A comparison of the classification accuracy garnered with these changes and the unmodified case can be found in Table 4.8.
The overlap in lab grade and final grade was first addressed by the complete removal of all lab grades from the information provided to the classifier. All other input remained the same and the same final grade category was used to categorize students.

The removal of the lab grade resulted in a drop of average classification accuracy by 16.11%. As can be seen in Table 4.8, classification accuracy increased as the semester proceeded and peaked at 65.38% in week 9.

The lab grade overlap was next examined by the removal of the lab grade from each student’s final grades. After this grade was removed the student’s grades were normalized again to be out of 100% and converted back into the three letter grade categories that were being used; exceeds expectations, meets expectations or fails expectations. The data provided to the classifier was not changed. The lab grade was still included as it was no longer a component of the final grade.

<table>
<thead>
<tr>
<th>Weeks</th>
<th>Unmodified</th>
<th>No Lab Grade</th>
<th>Final Grade W/out Labs</th>
<th>Lab Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 2</td>
<td>60.76</td>
<td>48.18</td>
<td>52.58</td>
<td>55.77</td>
</tr>
<tr>
<td>Week 3</td>
<td>58.41</td>
<td>44.02</td>
<td>54.92</td>
<td>59.09</td>
</tr>
<tr>
<td>Week 4</td>
<td>56.74</td>
<td>46.52</td>
<td>64.32</td>
<td>53.26</td>
</tr>
<tr>
<td>Week 5</td>
<td>72.73</td>
<td>51.82</td>
<td>71.14</td>
<td>59.17</td>
</tr>
<tr>
<td>Week 6</td>
<td>70.98</td>
<td>51.67</td>
<td>63.41</td>
<td>65.68</td>
</tr>
<tr>
<td>Week 7</td>
<td>78.64</td>
<td>56.89</td>
<td>72.73</td>
<td>70.08</td>
</tr>
<tr>
<td>Week 8</td>
<td>72.65</td>
<td>60.98</td>
<td>68.48</td>
<td>73.64</td>
</tr>
<tr>
<td>Week 9</td>
<td>79.62</td>
<td>65.38</td>
<td>62.73</td>
<td>75.38</td>
</tr>
<tr>
<td>Week 10</td>
<td>81.14</td>
<td>64.47</td>
<td>72.12</td>
<td>77.80</td>
</tr>
<tr>
<td>Week 11</td>
<td>81.21</td>
<td>61.89</td>
<td>66.97</td>
<td>71.97</td>
</tr>
<tr>
<td>Average</td>
<td><strong>71.29</strong></td>
<td><strong>55.18</strong></td>
<td><strong>64.94</strong></td>
<td><strong>66.08</strong></td>
</tr>
</tbody>
</table>

Table 4.8: Comparison of Lab Changes
Removing the lab grade from the final grade yielded an average classification accuracy of 64.94%. Accuracy was not markedly lower than using average data with lab grades in the earlier weeks. As the weeks progressed accuracy did not increase much and became noticeably lower than the accuracy of the unmodified case. The average loss in accuracy from the unmodified case was 6.35%.

Finally, the overlap of lab grades in the classifying data and in the final grade was addressed by removing lab grades and replacing them with a measure will henceforth be referred to as lab completion. Lab completion is a binary representation indicating whether or not students had completed their lab. Students who received a lab grade greater than zero were considered to have completed the lab. Students with a lab grade of zero were considered not to have completed the lab. All other aspects including the final grade categories remained the same.

The change from lab grades to lab completion resulted in a drop of average classification accuracy of 5.20%. As can be seen in Table 4.8, classification accuracy generally increased each week, and as of week 7 was consistently over 70%. These results led to the conclusion that a lab completion metric is useful when the goal is detecting success or failure as early as possibly without the use of grade data.
4.7 In-Depth Analysis

4.7.1 Weekly Accuracy

Prediction of student success or failure was completed using averaging to include prior weeks and lab completion in the place of lab grades. The weekly accuracy varied from the previous analysis as they are calculated slightly differently. The batch experimentation of classification accuracy using Weka used the classification accuracy from each of the ten-folds and averaged that to determine a final classification accuracy for each week. It is important to note that with 118 students there were 8 folds each containing the data of 12 students and 2 folds each containing the data of 11 students. Because each fold was not equal in size, it means that there were minor differences between the weighting of the classification accuracy of each student. When classifying individually on each week, Weka provided the total number of students classified correctly and incorrectly from each fold, summed them and determined accuracy in this manner. The resulting weekly accuracies are shown in Table 4.9.
<table>
<thead>
<tr>
<th>Week</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>55.10</td>
</tr>
<tr>
<td>3</td>
<td>59.30</td>
</tr>
<tr>
<td>4</td>
<td>53.40</td>
</tr>
<tr>
<td>5</td>
<td>59.30</td>
</tr>
<tr>
<td>6</td>
<td>66.10</td>
</tr>
<tr>
<td>7</td>
<td>70.30</td>
</tr>
<tr>
<td>8</td>
<td>73.70</td>
</tr>
<tr>
<td>9</td>
<td>75.40</td>
</tr>
<tr>
<td>10</td>
<td>78.00</td>
</tr>
<tr>
<td>11</td>
<td>72.00</td>
</tr>
<tr>
<td>Average</td>
<td><strong>66.26</strong></td>
</tr>
</tbody>
</table>

Table 4.9: Accuracy By Week Using Average Weeks and Lab Completion

As can be seen in Table 4.9, accuracy increased throughout the semester, peaking in week 10. In the final five weeks of the semester students were consistently classified with an over 70% accuracy rate.

### 4.7.2 Accuracy by Category

The focus of this thesis is on those students in the extreme categories. These are the students who will be placed in the *exceeds expectations* category; those who received an A, and the students who be placed in the *fails expectations* category; those received an F. The classification accuracy of the model built on averaging prior weeks and using lab completion was broken down based upon its accuracy predicting each category of *exceeds expectations, meets expectations or fails expectations*.

Table 4.10 details the weekly accuracy of the classifier at predicting each class. The classification accuracy of the extremes was markedly higher than
the classification accuracy of students classified as *meets expectations*. It is interesting to note that the classification accuracy of the *exceeds and fails expectation* students does not increase substantially throughout the semester. The gains in overall classification are driven by an increase in accuracy of classifying the *meets expectations* students. With the exception of week 3, in any given week, students in the *exceeds expectations* category were classified correctly with an accuracy rate of over 77%. The classification accuracy of students in the *fails expectations* category was lower than that of the *exceeds expectations* category, however it was more constant throughout the semester. In any given week throughout the semester, students in the *fails expectations* category could be predicted with an over 64% accuracy, and in 7 out of 10 weeks an over 70% accuracy.

<table>
<thead>
<tr>
<th>Week</th>
<th>Total Classification</th>
<th><em>Exceeds Expec.</em></th>
<th><em>Meets Expec.</em></th>
<th><em>Fails Expec.</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 2</td>
<td>55.1</td>
<td>77.6</td>
<td>14.6</td>
<td>75.0</td>
</tr>
<tr>
<td>Week 3</td>
<td>59.3</td>
<td>40.8</td>
<td>78.0</td>
<td>64.3</td>
</tr>
<tr>
<td>Week 4</td>
<td>53.4</td>
<td>83.7</td>
<td>9.8</td>
<td>64.3</td>
</tr>
<tr>
<td>Week 5</td>
<td>59.3</td>
<td>98.0</td>
<td>4.9</td>
<td>71.4</td>
</tr>
<tr>
<td>Week 6</td>
<td>66.1</td>
<td>77.6</td>
<td>46.3</td>
<td>75.0</td>
</tr>
<tr>
<td>Week 7</td>
<td>70.3</td>
<td>85.7</td>
<td>48.8</td>
<td>75.0</td>
</tr>
<tr>
<td>Week 8</td>
<td>73.7</td>
<td>77.6</td>
<td>73.2</td>
<td>67.9</td>
</tr>
<tr>
<td>Week 9</td>
<td>75.4</td>
<td>77.6</td>
<td>73.2</td>
<td>75.0</td>
</tr>
<tr>
<td>Week 10</td>
<td>78.0</td>
<td>85.7</td>
<td>63.4</td>
<td>85.7</td>
</tr>
<tr>
<td>Week 11</td>
<td>72.0</td>
<td>79.6</td>
<td>63.4</td>
<td>71.4</td>
</tr>
<tr>
<td>Average</td>
<td>66.3</td>
<td>78.4</td>
<td>47.6</td>
<td>72.5</td>
</tr>
</tbody>
</table>

Table 4.10: Accuracy By Category
4.7.3 Identifying the Most Predictive Data

The J48 classifier creates decision trees based on the data which represent rules for classifying students into their final grade categories. Through understanding the classification of these decision trees the data required for intelligent systems can be better identified. Three of these trees will be discussed to examine how the classification is done. The decision trees shown in this section have been created using the entire data set without cross-validation. The accuracies shown here are higher previously reported as a result of all of the data being used.

Decision trees are composed of conditionals and classifications. Conditionals are represented by parallelograms with two possible outcomes. These outcomes are shown on the lines stemming out from the circle. Classifications are represented as terminal nodes and contain the classification category. They contain the number of correctly classified instances on the left and the number of incorrectly classified instances on the right. Classification categories with only one number represent classifications with no incorrectly classified students. Each classification category can be traced back up to the root of the tree to represent a rule.

Week 2

Figure 4.2 shows the decision tree created to classify students at the end of week 2. The decision tree in Figure 4.2 has been created using the averages
of the data from week 1 and week 2. From this decision tree for week 2, five rules can be deduced:

- IF a student has completed less than half of his or her labs THEN the student will be classified under the **fails expectations** category.
- IF a student has completed more than half of his or her labs AND has viewed his or her assignment three or fewer days before it was due THEN the student will be classified under the **fails expectations** category.
- IF a student has completed more than half of his or her labs AND has viewed his or her assignment four or more days before it was due AND was active for 4.5 or less days per week AND visited the website 57.5 or less times THEN the student will be classified under the **exceeds expectations** category.
- IF a student has completed more than half of his or her labs AND has viewed his or her assignment 4 or more days before it was due AND were active for 4.5 or less days per week AND visited the more than 57.5 times THEN the student will be classified in the **meets expectations** category.
- IF a student has completed more than half of his or her labs AND has viewed his or her assignment 4 or more days before it was due AND were active for more than 4.5 days per week THEN the student is classified in the **exceeds expectations** category.
Students who did not complete more than half of their homework (labs) in the first two weeks in the course were classified under the fails expectations category with a 75% accuracy rating. For students who did complete more than half of their labs, if they first viewed their assignment close to the deadline they were also classified under the fails expectations category with a 100% accuracy rating. It is important to note that this fails expectations classification is a small category comprising of only three students. Students who completed more than half of their homework and viewed their assignment early were classified as either exceeds or meets expectations students. Of these students, those with more days active or with less days...
active and less total visits were classified as *exceeds expectations* students and those who had less days active but more total visits were classified as *meets expectations* students.

**Week 8**

The 8th week corresponds with the 40th day of classes, which at the University of Guelph is the last day during which a student may withdraw from a course without penalty. By the 40th class day, one would expect the classifier to have sufficient data to have a strong understanding of how well a student will do in the course. Figure 4.3 shows the decision tree for week 8. It can be immediately seen that those who have completed half or less of their labs were classified as students under the *meets expectations* or *fails expectations* categories. Of those students, those with low confidence in their final grade, or those who reported a high degree of confidence but who had a low attendance rate were classified as *fails expectations* students. Students who self-reported a high final confidence and had a high rate of attendance were classified as *meets expectations* students. Students with higher than 50% lab completion throughout the semester were grouped once more on the basis of lab completion. Those with a very high lab completion amount were classified as *exceeds expectations* students. Students with between 50% lab completion and 87.5% lab completion were differentiated by their attendance. Those with an attendance of less than 29.1% were classified under the *fails expectations* category, while those with attendance
of greater than 29.1% were classified as meets expectations students. Two features that stand out in Figure 4.3 are lab completion and attendance. In each case a higher lab completion or attendance rate indicated a higher mark.

Figure 4.3: Decision Tree From Week 8

Week 11

In Figure 4.4, it can be seen that lab completion is an important factor when classifying students. Students with a lab completion of less than 45% were classified as fails expectations students. Those with a lab completion of between 45% and 64% were classified under the meets expectations category. Male students with a lab completion between 64% and 82% were differentiated by their attendance. Those with an attendance rate of higher than 80% were classified under the exceeds expectations category and those
with an attendance rate of less than or equal to \(80\%\) were classified under the \textit{meets expectations} category. Female students with a lab completion of between \(64\%\) and \(82\%\) are classified in the \textit{meets expectations} category. Given that this rule only applied to four students it is not clear if this rule was due to an anomaly in the data or if gender was predictive of success. Students who had a lab completion of higher than \(82\%\) were again grouped by their attendance. Students with an attendance rate of at least \(38\%\) were classified under the \textit{exceeds expectations} category while those with a lower attendance rates were classified under the \textit{meets expectations} category. In Figure 4.4, it is evident once more that lab completion and attendance were indicative of student success and that higher in class attendance and lab completion rates result in the student being classified as a higher grade category.
Figure 4.4: Decision Tree From Week 11
4.7.4 Relevant Data

The decision trees for each week were different and relied on different data selected by the J48 algorithm. Understanding the data used in the prediction of student success in the course was important in determining which pieces of data were relevant and worth capturing either over time or on the fly. The data that were selected by the classifier each week are enumerated in Table 4.11. The five most commonly used pieces of data were lab completion, attendance in class, assignment viewing date, days active and the time taken to complete the problem set. Lab completion were by far the most commonly used data, they were found in every week’s classifier, often more than once. Lab completion accounted for 44.68% of the data used. Furthermore, lab completion was the conditional at the top of every tree for each week. Every rule from every decision tree had at least one component derived from lab completion. The prominence of lab completion in the rules further indicated their importance in the prediction of student success or failure. Following lab completion, with a weighting of 12.77% was attendance, which were selected for use in the decision trees for four weeks. How early a student viewed his or her assignment composed 8.51% of the data selected by the classifier. The number of days a student was active as well as the length of time taken to complete his or her problem set were selected by the classifier 6.38% of the time. Students self-reported confidence, gender and problem set score each accounted for 4.26% of the
data selected by the classifier. Self-reported final confidence, time taken to complete the quiz, total visits to the website and how early the student viewed the lab were found only once throughout the weeks, each accounting for a total of 2.13% of the data selected by the classifier. Of the data provided to the J48 algorithm only expertise, semester level, and program were not selected by the classifier for use in any of the weeks. In summary, the most important factors when predicting student success in this study, as many might theorize from experience, were the completion of homework and attendance in class.
<table>
<thead>
<tr>
<th>Week</th>
<th>Lab</th>
<th>Attendance</th>
<th>Assign Viewed</th>
<th>Days Active</th>
<th>PS Time</th>
<th>Conf.</th>
<th>PS #</th>
<th>Gender</th>
<th>Final Conf.</th>
<th>Quiz Time</th>
<th>Visits</th>
<th>Lab View</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Week 3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Week 8</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Week 11</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

| % of Total | 44.08 | 12.77 | 8.51 | 6.38 | 6.38 | 4.26 | 4.26 | 4.26 | 2.13 | 2.13 | 2.13 | 2.13 |

Table 4.11: Data Used By the Classifier
Chapter 5

Conclusions, Limitations and Future Work

Early prediction of the success or failure of students in the course was achieved without the use of formal evaluations or a domain model. Beginning with the second week, students were correctly classified at a higher rate than if they were randomly assigned to one of the three final grade categories. Leveraging data from prior weeks yielded a higher classification accuracy. As seen in Table 4.9, average weekly accuracy increased to 66.26% when the mean of all prior weeks was used.

During the first five weeks of the semester the classifier reached an accuracy of nearly 60%. Week 6 was a turning point where the classifier reached a 66.10%. The drop date at the University of Guelph is the 40th class day
which occurs at the end of the 8th week of class. After this point students may no longer drop the course without penalty. As of the seventh week of classes students were classified into final grade categories with over 70% accuracy. Clearly, student success or failure can be predicted early in the course without the use of formal evaluations. While the temptation is to compare these results with the results of previous work this is problematic due to varying classification category sizes and the varying data that were collected in other studies. It is also problematic to compare the results of this work due to the more dynamic weekly classifications found in this work.

As seen previously in Table 4.10, it was easier to accurately classify students in the edge categories of fails expectations or exceeds expectations. Students in the edge categories could be accurately classified from the very first week of the course. With the exception of classification during week 3, all classifications of the exceeds expectations and fails expectations categories had a greater than 60% classification accuracy. Classification of early success or failure was more accurate throughout the semester when considering only the edge categories of those students who will either fail the course or receive a grade over 80%.

Analysis of the most predictive variables of success or failure yielded interesting results. As detailed in Table 4.11 the top five variables used most often in classification of success or failure were:

- Lab completion.
• Attendance.
• How early assignments were viewed.
• Days active.
• Time spent on the problem set.

Lab completion was by far the most predictive of student success, followed by attendance. The predictive nature of lab completion and attendance may not come as a surprise to some educators but provides a solid understanding about which data are most important for studies like this to gather. Four of the five most meaningful data can be easily gathered from a LMS. Attendance data were not gathered from the LMS but with the use of in class clickers were also easily obtained. It is not necessary to have intrusive data collection in order to have an accurate and early prediction of student success or failure.

It seemed likely that student self-evaluations of expertise, final grade and confidence would be predictive when classifying student success. However, only 4.26% of the classifier’s selected data were comprised of student self-reported confidence in the course. Self-reported confidence in the final grade they would receive comprised only 2.13%, and self-reported expertise in the course material was one of only three variables that was not selected for use by the classifier at all. The time taken by the student to complete self-evaluation questions and the problem set questions turned out to be more predictive of success than the answers given. The time taken to complete
the problem set might be predictive because it indicates how much effort students put towards the self-evaluations and may have been reflective of their effort throughout the rest of the course.

How early students viewed their labs and assignments was expected to be predictive of success or failure. Assignment viewing dates were selected by the classifier for use in 8.51% of the decision trees created, while lab viewing dates were selected in only 2.13% of cases. It is interesting that assignment viewing dates were much more useful than lab starting dates even though there was less data available due to the fact that there were only four assignments while there were eleven labs. The importance of assignment viewing dates may have been due to the fact that the time required to complete assignments was much longer than time required to complete labs. A student viewing a lab the night before still had time to complete the lab, if they were viewing an assignment the night before it was very unlikely that they would have time to complete the assignment properly.

The ability to predict success or failure could be used to provide instructors with an early warning of who may need enrichment opportunities or extra help. This prediction is particularly important in larger classes where instructors may not have the opportunity to assess each student individually. Instructors may utilize this warning to make adaptations for students who are at risk of failing or those who may need enrichment opportunities.
to maintain motivation.

A prediction of success or failure can also be used to help students to be more aware of how they are doing in the course. Extra resources can be made available to students who might need extra help and suggestions for activities based on predicted success can be provided. Self directed learning requires this feedback and this thesis presents a method of providing such feedback.

This work can produce a student model of predicted success or failure in the course without domain knowledge. Student models based on domain models require a new domain model for every domain they are used in. These domain models are time consuming and complex to create. A student model that does not rely on domain specific data can be used to inform an adaptive system across domains. An ITS can then utilize these predictions to adapt course content and flow to student needs.

5.1 Limitations and Future Work

Future work could include the validation of the work in this thesis in other courses and domains, exploring possible additional data of interest or other attributes that can be predicted and making use of the adaptations that are made possible through the prediction of student success. There are also several limitations to this work that should be addressed in future work.
including the bimodal distribution of grades in the course, the removal of students who did not complete the course, the assumption of lab viewing data, inconsistencies of the forum viewing data, the method by which lab completion was calculated, and the ineffectiveness of the self-evaluations. The validation of this work is vital as replicating the results will allow this work to be used in ITS in other courses as well as other fields.

5.1.1 Validation

This study was completed using data gathered from only one class. The grades from the class did not have a standard distribution, making it harder to generalize the results. This thesis has shown that it was possible to predict student success in the winter 2011 offering of CIS 2500: Intermediate Programming at the University of Guelph. The validation of these methods with additional classes would allow this research to be generalized to a wider variety of circumstances. The first step for this validation would be to complete the study again in another offering of CIS 2500 at the University of Guelph. Repeating this study would provide evidence that this methodology can be repeated with success. To further expand the scope of the results of this thesis, studies could be completed in other computer science classes at the University of Guelph in differing years and topics within computer science. A study in other computer science courses would allow the conclusion that it is possible to predict the success or failure of students
in computer science classes across years at the University of Guelph. Due to the bimodal distribution of grades in the course studied in this thesis this work must be replicated in other courses to ensure that the method will work in courses with different distributions of final grade. In other studies, categories may need to change based on the distributions as the *fails expectations* category would be rather small in a course where only a few students fail.

5.1.1.1 Validation Across Domains and LMS

This work was intended to be unique in that it can be replicated across classroom domains such as computer science, Math and English. The only data gathered specific to the course were the problem set answers which could be removed altogether given that they were not one of the more predictive features in this study. Perhaps the most important expansion of the scope of this research will be across domains. The prediction of student success as completed in this thesis needs to be applied to classes in other domains to validate that it is possible to use LMS interactions to predict student success or failure in different domains. If this work can be validated in other domains will allow for a domain independent prediction of a students’ success or failure within a course.

The re-usability of this work can be further tested by completing similar studies at different universities and different levels of education such as
high school and college to ensure that success or failure can be accurately predicted across teaching styles and cultures. This work focused specifically on interaction data from Moodle. This study should be replicated using other LMS, which will allow for models to be created in other LMS as well as provide confidence that this method is LMS independent and would work in future systems.

5.1.2 Suggested Improvements

The way some of the data were gathered and amalgamated during this study can be improved in future studies. The completion of problem sets was required before students could view their labs. The first viewing date of labs was assumed to be the date students completed their problem set and unlocked access to their lab. Students may have actually first viewed their lab on a later date than they completed their problem set resulting in their lab viewing date being represented incorrectly. The possible inconsistency with lab viewing date may reduce the accuracy of the classifier. Moodle interaction logs contain student every student interaction with the website. In future studies these should be used to find each lab and determine the date it was viewed.

The problem sets were intended to provide an alternate mechanism for assessing student expertise, but did not function in that capacity. The problem sets were ungraded and unpopular with students resulting in data
that were not predictive of student success. Additionally, students who did complete the problem sets often left the interactive tool open for many hours even though the time to complete them should have been less than 10 minutes. In order to make use of the data that were collected from the problem sets, the maximum completion time was capped at 60 minutes. This may have affected the utility of the timing data to the decision tree and is a part of this work that should be revisited in future work.

Students were able to sign up to receive forum posts to their email instead of viewing each thread online. The number of days students were active as well as the total weekly number of visits may be artificially lower for students who signed up to receive forum posts in their email as compared to those who used the website. This research does not differentiate between students who read the forums through the course website and those who received daily digests of forum posts. Students who received emails of forum posts may have read just as many posts as students who used the website, but the data used for this study only reflected this activity for students who read forum posts on the website. This inconsistency means the classifier has less accurate data and could have a reduced prediction accuracy. Consistent information about student visits to the forum would also mean that in future work the number of forum posts read could be explored as a predictive feature when classifying success or failure in a course.

The most predictive data used by the classifier was lab completion.
Lab completion was a binary representation of whether or not the student had received a mark greater than zero for the lab. As such, students who submitted a lab but received a lab grade of zero were considered to not have completed their labs. More in depth data about lab completion would be worth investigating. Properly representing those who did complete the lab but received a zero may be useful in future work.

Different types of student interaction with the LMS such as visits to forum pages or lesson pages were not differentiated. Aggregated counts were used to create a value for the number of pages a student viewed and the number of days they were active on the website. This research has not investigated whether different types of interactions such visits to the forum are more predictive than others. Investigating the different type of website interactions by students may provide predictive data in future work.

5.1.3 Changes to the Data

Despite the large amount of data gathered it would be interesting to pursue data that were not available in this study in an attempt to increase the accuracy of the classifier.

In CIS*2500 students worked on their labs at home or in the labs and it was not possible to monitor their progress or the time they took to complete their work. Data regarding student’s work on their labs and assignments could provide more predictive data to the classifier. The number of hours
spent on an assignment or lab and the time and frequency for each work session could be very predictive of student success.

Other data worth considering might be student grades from prior courses. For example, perhaps the classifier would differentiate between a student with a prior average above 90% and a student with a history of failing who both view the lab only a couple of days before it is due. This differentiation could result in a higher classification accuracy. The classifier could consider both a student’s average as well as a student’s marks in specific courses. Grades in similar domains might be more predictive of student success in the course than grades in unrelated courses.

How students learn changes how they can best use the resources provided to them. Information about learning styles and working habits might provide the classifier with a greater understanding of each student and how they learn. Gathering data in a variety of different settings could provide insight into how students learn best.

One goal of this study was to find predictive relationships to a student’s final grade without using contributing grades. As such, the grades from formal evaluations were not used to predict student success or failure. Future work could include a student’s grades from formal evaluations to increase the accuracy of prediction.

A study focused on determining how much data are actually needed to create an accurate classifier could be instructive. Determining the change
in classification accuracy if a subset composed of the X most predictive
variables were used across all weeks would help to determine if less data
are sufficient. Collecting less data would make it easier to transfer the this
research across domains and course types that may not have all of the data
collected in this study available.

5.1.4 Improvements to the Self-Evaluations

The value of the self-evaluations questions used in this study is suspect
due to their poor predictivity. An examination of the data from the self-
evaluation questions shows that many participants submitted low quality
data for the self-evaluations. Sometimes the data point was unusable be-
cause it was not an appropriate value, sometimes it was simply a value at
one extreme of the scale or another and sometimes no value was inputted.
It is possible that these questions were not taken seriously and considered to
be a roadblock to viewing the lab and therefore not answered thoughtfully.
This may have been due to the time it took to complete the questions, lack
of incentive for students to complete the questions properly or because the
questions were confusing.

The self-evaluation questions for this study were developed for this re-
search to determine confidence and expertise in a way that was domain
independent. The self-evaluation questions were limited to three questions
as they were mandatory for all students in the class and done weekly. Self
efficacy studies are a proven method of determining a person’s judgement of their capability within a specific area [32] [7]. Further research into developing a self efficacy instrument that can determine a student’s judgement of his or her expertise and capability within a specific domain while the questions themselves remain domain independent would help to provide more reliable self-evaluation scores. This self efficacy instrument would either need to be comprised of a much smaller set of questions than most self efficacy questionnaires, gathered less frequently than the weekly self-evaluations of this study, or used in a study where participants were willing to complete a larger self-evaluation questionnaire each week.

5.1.5 Automated Data Collection and Cleaning

The amalgamation and cleaning of the data for this study was done manually and was an intensive process. An automated system of collecting, amalgamating and cleaning the data would allow for easier replication of this work and make the use of it as part of an adaptive system much more feasible. Automated data collection and cleaning is a problem for many researchers in the EDM field. Attempts have been made to create systems to automatically collect, amalgamate and clean data. The EDM workbench is an attempt to solve this problem but is prototypical at best and not very robust [34]. A working system of this nature would increase the speed of research in the field and make it easier to create dynamic models that update
on the fly.

5.1.6 Prediction

This research only considered data from participants who completed the entire course. Students who drop a course would also benefit from early intervention. It would be interesting to determine if students who dropped the course can be predicted using the approach as described in this thesis.

Student success was predicted using weeks of data in order to move towards a dynamic student model using EDM techniques. A prediction of student success or failure is available at the end of each week and uses data from all prior weeks. Future work should include a dynamic model that updates a prediction of student success or failure on the fly as students learn and use course materials.

This research classified students into predefined final grade categories using prediction. Clustering involves grouping objects into classes of other similar objects and can be used to find students with similar learning characteristics [37]. Clustering would identify students with similar data and cluster them together, perhaps allowing for a better understanding of the types of students in a class. Using clusters instructors could more easily create groups of either similar or different students. Instructors would also be able to examine the clusters and characteristics of each cluster to have a better understanding of the types of students in the class.
5.2 Conclusion

This study has proven that the prediction of success or failure without the use of a domain model or formal evaluations is possible. The validation of this work and subsequent adaptations made on the predictions of success or failure will allow personalized learning environments for students. Students in large classes will be able to have a personalized learning environment without extra effort on the part of the instructor. If this can lead to even a small decrease in students who fail a course it will be an important step forward in the design and delivery of courses using ITS.
Bibliography


Appendix A

Ethics Documents

The following sections contain the documents submitted to ethics including the ethics application, consent form, and the data collection plan.

A.1

Ethics Application
University of Guelph Research Ethics Board (REB)
FACULTY AND GRADUATE
Application to Involve Human Participants in Research

Please refer to the University of Guelph Research Ethics Guidelines, found at http://www.uoguelph.ca/research/forms_policies_procedures/human_participants.shtml before completing and submitting this application. If you have questions about this form, please contact the Research Ethics Coordinator, Sandra Auld at ext. 56606, or reb@uoguelph.ca.

Send this form and all accompanying material by email, as attachments, to reb@uoguelph.ca. One hard copy of the signed signature page should be forwarded to the Research Ethics Coordinator, Office of Research, University of Guelph, 437 University Centre, Guelph, ON, N1G 2W1.

If you want to change a previously approved protocol, please complete the “Change Request” form, available at http://www.uoguelph.ca/research/forms_policies_procedures/human_participants.shtml.

Date: 2010-12-04 (For OR use only) Protocol#: 

SECTION A – GENERAL INFORMATION

1. **Title of the Research Project:** Factors for predicting the expertise of in-course students

2. **Investigator Information**

<table>
<thead>
<tr>
<th>Name &amp; position</th>
<th>Dept./Address</th>
<th>Phone No.</th>
<th>E-Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Faculty with Principal Responsibility</strong>*:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dr. Judi McCuaig</td>
<td>Computer Science</td>
<td>519-8244120 x58534</td>
<td><a href="mailto:judi@uoguelph.ca">judi@uoguelph.ca</a></td>
</tr>
<tr>
<td><strong>Faculty: Co-Investigator(s)</strong>:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Student: Investigator(s)**: |                  |                    |                      |
| Julia Baldwin           | Computer Science |                    | baldwinj@uoguelph.ca  |

| **Other: Investigator(s)**: |                  |                    |                      |

* must be advisor of any student investigators.

3. **Proposed Date**
   a) of commencement: January 10, 2011
   b) of completion: May 30, 2011

Note: The commencement date should be the date the researcher expects to actually begin interacting with human participants (including recruitment). The completion date should be the date that the researcher expects that interaction with human participants, including any feedback or follow-up, will be complete.

4. **Indicate the location(s) where the research will be conducted:**

University of Guelph
Other (please specify site):
5. **Other Research Ethics Board Approval**

   a) Is this a multi-centred study? □ □
   
   b) Has any other institutional Ethics Board approved this project? □ □

   c) If Yes, please provide the following information:
      - Title of the project approved elsewhere:
      - Name of the Other Institution:
      - Name of the Other Board:
      - Date of the Decision:
      - A contact name and phone number for the other Board:
      - OR
      - A copy of the clearance certificate / approval

   d) Will any other Research Ethics Board be asked for approval? □ □
      - If Yes, please specify:

6. **Level of the Project**

   - Faculty Research □
   - PhD Thesis □
   - Masters Thesis □
   - Honours Thesis □
   - Class Project □
   - Internship □
   - Practicum □
   - Other (please specify):

7. **Funding of the Project**

   a) Is this project currently funded? □ □

   b) Period of Funding: From __________ To __________

   b) Agency or Sponsor (funded or applied for)
      - CIHR: □
      - NSERC: □
      - SSHRC: □
      - Other (please specify):

   Note: Please specify the complete title of the funding source. For example, “NSERC Discovery Grant”.

   **NOTE:** If the funding source changes, or if a previously unfunded project receives funding, you must submit a Change Form to the Research Ethics Coordinator.

8. **Conflict of Interest**

   a) Will the researcher(s), members of the research team, and/or their partners or immediate family members:

      i) Receive any personal benefits (for example a financial benefit such as remuneration, intellectual property rights, rights of employment, consultancies, board membership, share ownership, stock options etc.) as a result of or connected to this study? □ □ □ □ □
ii) If Yes, please describe the benefits below. (Do not include conference and travel expense coverage, possible academic promotion, or other benefits which are integral to the general conduct of research.)

N/A

b) Describe any restrictions regarding access to or disclosure of information (during or at the end of the study) that the sponsor has placed on the investigator(s).

N/A

c) Discuss the possibility of commercialization of the research findings.

This study will be examining the relationship between different types of data that can be collected about students. As such, the results of this study are not directly commercializable, however the results of the study will be used in the development of intelligent tutoring systems, which may be suitable for commercialization at some point in the future.

SECTION B – SUMMARY OF THE PROPOSED RESEARCH

9. Rationale

Describe the purpose and background rationale for the proposed project, as well as the hypotheses(is)/research questions to be examined.

Intelligent tutoring systems rely on knowledge about the learner to make decisions about the information that should be presented to the learner at any specific point in time. Often the tutoring system is making decisions based on little more than a record of how long a learner has spent with material about a specific subject. This study will determine if commonly available information about learners can be used to predict the learner’s expertise with particular topics. Specifically, experimenters will collect information that is normally available to instructors via online learning environments and in-class clickers, such as start times and hand in times for assignments, labs and online lessons, seating position within the classroom, lecture attendance, participation and successes in in-class activities, self assessments of expertise and objective assessments of expertise. The assessments of expertise (self and objective) will be compared to the other data collected to determine which, if any of the available data has a relationship with student expertise. The expertise assessments will also be compared with student letter-grades to determine if there is a correlation between in-course assessment evaluations and expertise assessments. Specifically the hypothesis for this study is that some of the easily collectible metadata about student interactions with course material can be used to predict or assess the expertise of the student. The purpose of the study is to identify which metadata are most suitable.

10. Methodology

Describe sequentially, and in detail, all procedures in which the research participants will be involved (e.g., paper and pencil tasks, interviews, surveys, questionnaires, physical assessments, physiological tests, time requirements etc.)

Note: Attach a copy of all questionnaire(s), interview guides or other test instruments. These should be on University of Guelph letterhead if they are intended for public dispersal.

The experimental subjects for this study will be the students in CIS 2500 in the Winter 2011 semester. Because the study activities are also the course assignments, quizzes and labs, all students will complete exactly the same activities, regardless of whether they decide to participate in the study or not. Metadata will be collected and processed only for those students who sign the consent form for the study. The consent form (attached) will be available in lectures and labs for the first week of classes (three lecture slots).
Consent forms will be printed, not electronic. Students will be invited to participate via in-class announcements and via email. The announcement used will be the first three paragraphs of the consent form. (Intro, Purpose and Procedures)

The only study-specific activity that is required of participants is to answer three self-assessment questions along with their weekly problem-set (self-assessment questions attached). These questions are designed to provide experimenters with a baseline for identifying participant expertise, but will also serve pedagogically to encourage student self-reflection. As such, all students will be required to answer the self-reflection questions, even those who are not participating in the study. A draft course outline for this class is attached to this proposal that describes all of the lab, exam, and assignment activity. It is important to note that the weekly problem set has NO impact on students’ final grades, it is simply a set of practise questions. Metadata will be collected, for participants, from each of these activities throughout the semester. A detailed description of all data that will be collected during this study is attached in a separate document (Data Collection Plan)

11. Experience

What is your experience with this kind of research?

Professor McCuaig has conducted usability and interaction studies many times. Julia Baldwin has been involved in several experiments within the research group as a coach and experimental design consultant.

12. Participants

Describe the number of participants and important characteristics (such as age, gender, location, affiliation, etc.)

Participants will be students who are taking CIS 2500 in the Winter 2011 semester. The course is open to students who have completed CIS 1500. It has an estimated enrollment of 180 students. Most of the students are in the College of Physical and Engineering Sciences (CPES).

13. Recruitment

a) Describe how and from what sources the participants will be recruited, including any relationship between the investigator(s) and participant(s) (e.g., instructor-student; manager-employee).

Note: Attach a copy of any poster(s), advertisement(s) or letter(s) to be used for recruitment.

Participants will be recruited exclusively from the above-mentioned course. Recruitment will consist of an explanation of how the data will be collected and how it will be used and the safeguards that will be in place to protect the identity of individuals. The consent form will be used as the structure for the recruitment talk. The first three sections of the consent form will be read as the recruitment speech.

b) How and where will you contact these participants?

Participants will be recruited during the first week of classes, in class and during lab sessions.

c) Time required of participants: hour(s) on 1 occasion(s).

No extra time is required of participants. Participants simply complete the course in which they have registered.

d) Are participants proficient in the language in which the survey is being conducted? Yes   No

If not, is translation available? Yes   No
14. **Compensation**

   a) Will participants receive compensation for participation?  
      i) Financial  
      ii) Non-financial  

   b) If Yes to either i) or ii) above, please provide details.

   There is no payment for participation. Participant’s names will be entered into a draw for a $50 gift certificate to Future Shop. Chances of winning a certificate depends on the number of participants and the number of certificates donated. At least two certificates will be drawn for.

   c) If participants choose to withdraw, how will you deal with compensation?

   Non-participants and participants who withdraw will be treated the same way. Their name will not be entered into the draw for the Future Shop gift certificates.

---

**SECTION C – DESCRIPTION OF THE RISKS AND BENEFITS OF THE PROPOSED RESEARCH**

15. **Possible Risks**

   a) Indicate if the participants might experience any of the following risks:  
      i) Physical risk (including any bodily contact or administration of any substance)?  
      ii) Psychological risks (including feeling demeaned, embarrassed worried or upset)?  
      iii) Social risks (including possible loss of status, privacy and/or reputation)?  
      iv) Is there any deception involved?  
      v) Are any possible risks to participants greater than those the participants might encounter in their everyday life?

   b) If you answered Yes to any of points i) through v) above, please explain the risk.

   N/A

   c) Describe how the risks will be managed (including an explanation as to why alternative approaches could not be used).

   N/A

16. **Possible Benefits**

   Discuss any potential direct benefits to the participants from their involvement in the project. Comment on the (potential) benefits to the scientific community/ society that would justify involvement of participants in this study.
Students will have the opportunity to reflect on their personal progress at many times in this course. I believe that the extra attention paid to student self-efficacy through the weekly problem sets will positively affect student engagement with the course.

SECTION D – THE INFORMED CONSENT PROCESS

17. The Consent Process

a) Describe the process that the investigator(s) will be using to obtain informed consent, including a description of who will be obtaining the informed consent. If there will be no written consent form, explain why.

Consent forms will be distributed in class and during lab times during the first week of class.

For information about the required elements in the letter of information and the consent form, please refer to “Instructions for the Preparing Information and Consent Letters” and the sample consent form available at [http://www.uoguelph.ca/research/forms_policies_procedures/human_participants.shtml](http://www.uoguelph.ca/research/forms_policies_procedures/human_participants.shtml).

*Note: Attach a copy of the Letter of Information (if applicable), the Consent Form (if applicable), the content of any telephone script (if applicable) and any other material which will be used in the informed consent process. If the document will be made public, please ensure that it is on University of Guelph letterhead.*

Consent form attached. Student Id numbers are requested on the consent form to ensure that we collect data from the correct participants. CIS 2500 is a large class that may contain students with duplicate names, so name alone may not allow us to distinguish between students who have signed consent forms and those who have not. Since ALL CIS 2500 students will be doing all of the activities associated with the course and the study, it is important that we correctly discriminate between students who sign the consent form and students who do not. Student ID numbers are the only way to ensure correct discrimination.

b) Will the information provided to the participants be complete and accurate? Yes ☒ No ☐

If no, please describe the nature and extent of the deception involved. Include how and when the deception will be revealed, and describe the specialized training of the person who will administer this feedback. It is recommended that participants have the opportunity to sign a second consent form, following debriefing when the deception is revealed, to ensure a fully informed consent.

*Note: Attach a copy of the debriefing feedback and, if necessary, a copy of the second consent form on University of Guelph letterhead.*

N/A

18. Consent by an authorized party

If the participants are minors or for other reasons are not competent to consent, describe the proposed alternate source of consent, including any permission / information letter to be provided to the person(s) providing the alternate consent.

Participants will be past the age of majority.
19. **Alternatives to prior individual consent**

If obtaining individual participant consent prior to starting the research project is not appropriate for this research, please explain and provide details for a proposed alternative consent process.

N/A

20. **Participant feedback**

Explain what feedback/information will be provided to the participants after participation in the project. (For example, a more complete description of the purpose of the research, or access to the results of the research).

*Note: Please provide a copy of the written information, if applicable.*

When classes conclude, a participant may ask for (and will receive) a copy of all data collected about him or her. This will only be possible immediately after classes conclude because the key that links student ID to data will be destroyed at the end of the semester. All students will be given access to the research results as soon as possible. Preliminary results will be presented in a public seminar during the Summer 2011 semester. All students (whether or not they participate in the study) will be invited to attend. Final results will be published and made public later in 2011 or early 2012.

21. **Participant withdrawal**

a) Describe how the participants will be informed of their right to withdraw from the project. Outline the procedures that will be followed to allow the participants to exercise this right.

Participants who wish to withdraw will be required to send the experimenter an email stating that desire.

b) Indicate what will be done with the participant’s data and any consequences for the participant of withdrawing from the study.

There are no consequences for withdrawing. The data of withdrawn participants will be discarded completely.

c) If the participants will not have the right to withdraw from the project, please explain.

N/A

**SECTION E – CONFIDENTIALITY**

22. **Ensuring confidentiality**

a) Will all participants be anonymous? ❌ Yes ☑

b) Will all data be treated as confidential? ☑ No ❌

*Please note the difference:* Participants’ identity/data will be confidential if an assigned ID code or number is used, but it will not be anonymous. Anonymous data cannot be traced back to an individual participant.

c) Describe the procedures to be used to ensure anonymity of participants and/or confidentiality of data both during the conduct of the research and in the release of its findings.

While the data for this study cannot be anonymized, we will put procedures in place to ensure that
participants remain anonymous and that the identity of a participant cannot be deduced from the data. Those procedures are described in detail in the document that describes the data that will be collected during this study (attached).

d) Explain how written records, video/audio tapes and questionnaires will be secured, and provide details of their final disposal or storage.

All physical artifacts (paper quizzes, consent forms, etc) will be stored in a locked office in a locked filing cabinet. Those artifacts will be destroyed at the end of the Summer 2011 semester (giving time for data analysis after the collection period). Electronic data will be stored on the password-protected server for the ICC research group. The data collected for this study will be the basis of several research projects for the next 2-3 years, including sabbatical research that is planned by the primary investigator. Because of the importance of this data set, the researchers would like to keep the data for 5 years, with the understanding that future uses of the data will require separate ethics approval. The five year time period is noted on the consent form. Since participants will not be identified, there should be no impact on them as a result of keeping the data for that length of time.

e) If participant anonymity or confidentiality is not appropriate to this research project, explain, providing details of how all participants will be advised of the fact that data will not be anonymous or confidential.

N/A

SECTION F – MONITORING ONGOING RESEARCH

23. Annual Review and Adverse Events

a) Minimum protocol review requires the completion of a “Renewal/Completed Status Report” at least annually. Indicate whether any additional monitoring or review would be appropriate for this project.

Note: It is the investigator’s responsibility to notify the REB using the “Renewal/Completed Status Report” when the project is completed, or if it is cancelled. The form is available at http://www.uoguelph.ca/research/forms_policies_procedures/human_participants.shtml.

N/A

b) Adverse events (unanticipated negative consequences or results affecting participants) must be reported to the Research Ethics Board and the Research Ethics Coordinator as soon as possible.

24. Additional Information

(Use an additional page if more space is required to complete any sections of the form, or if there is any other information relevant to the project that you wish to provide to the Research Ethics Board.)

N/A
SECTION G – SIGNATURES

Responsible Faculty Assurance:

I, Judi  have the ultimate responsibility for the conduct of the study described in this application including my responsibilities as an advisor to any students involved in this project. I have read and am responsible for the content of this application. If any changes are made in the above arrangements of procedures, or adverse events are observed, I will bring these to the attention of the Research Ethics Coordinator.

Signature

Date 2010-12-08
A.2

Study Consent Form
CONSENT TO PARTICIPATE IN RESEARCH  

Factors for Predicting the Expertise of in-course Students

You are asked to participate in a research study conducted by Dr. Judi McCuaig and Ms. Julia Baldwin, from the School of Computer Science at the University of Guelph. The results of this study will be used in Julia Baldwin’s MSc thesis and in research that Dr. McCuaig plans to conduct during her sabbatical in Fall 2011.

If you have any questions or concerns about the research, please feel free to contact Dr. McCuaig. (extension 58534, judi@uoguelph.ca)

PURPOSE OF THE STUDY

This study will determine whether data about student interactions with course activities such as assignments, quizzes, exams, and labs has an identifiable relationship with student success in mastering course material.

PROCEDURES

If you volunteer to participate in this study, we would ask you to do the following things:

You will simply carry out the activities of the course as you normally would. We ask you to be honest when asked to supply information about your learning habits or estimations of your own understanding. Self reports will never affect your grade in the course.

The experimenters will collect information about your work and study habits from svn, the clickers, and moodle as the course progresses. All of this information will have your name and student number removed so that you cannot be identified. To protect your identity, your grades in course activities will be converted to letter grades prior to their use in the study. The researchers will use this data to improve tutoring systems and educational software such as Courselink. The data will be kept for 5 years, but your name and student ID number will not be recorded with the data.

POTENTIAL RISKS AND DISCOMFORTS

There are no foreseeable risks associated with participating in this study. All students in CIS 2500 (W11) will be required to complete all of the course activities regardless of their participation in the study.

POTENTIAL BENEFITS TO PARTICIPANTS AND/OR TO SOCIETY

At the end of the semester, you will be given the opportunity to request a copy of the data we collect about you. This will permit you to reflect on your own study habits and how they affect your achievement at university.

Intelligent tutoring systems rely on knowledge about learners to make decisions about the information that should be presented to the learner at any specific point in time. Often the tutoring system is making decisions based on little more than a record of how long a learner has spent with material about a specific subject. This study will determine if commonly available information about learners can be used to predict the learner’s expertise with particular topics.

PAYMENT FOR PARTICIPATION

There is no payment for participation. Participant’s names will be entered into a draw for a $50 gift certificate to Future Shop. Chances of winning a certificate depends on the number of participants and the number of certificates donated. At least two certificates will be drawn for.
CONFIDENTIALITY

Every effort will be made to ensure confidentiality of any identifying information that is obtained in connection with this study.

Data about you will be kept confidential at all times. Your name and student number will be replaced by an experimental ID number prior to the data being stored. Once the semester is over, all records that match your name and ID number to that experimental ID number will be destroyed.

Information about you that is used in this study will be made less granular where appropriate to protect your identity. For example, times will be recorded as Morning, Afternoon, Evening, Night, Late Night or Early Morning rather than exact times and grades will be recorded as letter grades rather than percentages.

PARTICIPATION AND WITHDRAWAL

You can choose whether to be in this study or not and your choice will not affect your grade in CIS 2500. If you volunteer to be in this study, you may withdraw at any time in the semester without consequences of any kind. If you withdraw, your data will be removed from the study. The investigator may withdraw you from this research if circumstances arise that warrant doing so.

RIGHTS OF RESEARCH PARTICIPANTS

You may withdraw your consent at any time and discontinue participation without penalty. You are not waiving any legal claims, rights or remedies because of your participation in this research study. This study has been reviewed and received ethics clearance through the University of Guelph Research Ethics Board. If you have questions regarding your rights as a research participant, contact:

Research Ethics Coordinator  Telephone: (519) 824-4120, ext. 56606
University of Guelph  E-mail: sauld@uoguelph.ca
437 University Centre  Fax: (519) 821-5236
Guelph, ON  N1G 2W1
I have read the information provided for the study “Factors for predicting the expertise of in-course students” as described herein. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

Name of Participant (please print)

____________________________________
Signature of Participant

____________________________________
University of Guelph Student Number

SIGNATURE OF WITNESS

Name of Witness (please print)

____________________________________
Signature of Witness

Date
A.3

Data Collection Plan
Factors for predicting expertise in in-course students
Data Collection Plan

Three types of data will be collected for this study. The first set of data includes metadata about the student’s habits with respect to starting and finishing labs and assignments and about their habits when studying from online materials. It also includes information about student attendance at lectures and labs and about where students sit in the lecture hall. The second type of data consists of weekly measurements of expertise. These measurements come from the weekly problem set and the self-reflection questions. The third type of data is student letter grades for the course. This data will be collected as students are graded. The tables below expand the descriptions for each type of data; indicate from where it will be collected and the transformation (if any) that will be applied to the data to protect student identities.

It is important to note that at no time will student names or UofGuelph ID numbers be used to identify experimental participants. All data will be recorded using ID numbers generated specifically for this experiment (assigned randomly to individuals at the beginning of the study). All data that is captured for this study will be filtered to remove student identifiers and replace them with the experimental ID number prior to storage. At the end of the data collection period, the key that is used to map student id to experimental id will be destroyed. During the data collection period, that key will reside only on the password-protected server for the research group.

Table 1- Metadata about Course Activities

<table>
<thead>
<tr>
<th>Short Name</th>
<th>Description</th>
<th>Identity Protection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment Submission Time and Assignment Revisions</td>
<td>All assignments for this course will be handed in via svn (a version control system common in software development). SVN records the timestamp each time something is saved to the system. The most recent timestamp for assignments will be retrieved from svn after the assignment due date. The first timestamp for assignment submission will also be retrieved (svn permits and encourages multiple revisions), as will the number of revisions between first and final submission.</td>
<td>Rather than using exact times, the submission timestamp will be converted to a date and the nearest sixth of the day. Early Morning (4-7:59 am), Morning (8-11:59 am), Afternoon (Noon-3:59 pm), Evening (4-7:59 pm), Night (8-11:59 pm), Late Night (midnight-3:59 am). This will make it impossible to connect a participant back to a specific assignment through the timestamp.</td>
</tr>
<tr>
<td>Lab Submission Time</td>
<td>All labs for this course will be handed in via svn (a version control system common in software development). SVN records the timestamp each time something is saved to the system. The most recent timestamp for labs will be retrieved from svn after the lab due date.</td>
<td>Rather than using exact times, the submission timestamp will be converted to a date and the nearest sixth of the day. Early Morning (4-7:59 am), Morning (8-11:59 am), Afternoon (Noon-3:59 pm), Evening (4-7:59 pm), Night (8-11:59 pm), Late Night (midnight-3:59 am). This will make it impossible to connect a participant back to a specific lab through the timestamp.</td>
</tr>
<tr>
<td>Lab and Assignment description reading</td>
<td>The time that students first read through an assignment or lab description can be obtained from the log files in Moodle (the courseware system in use for this course). The time that labs and assignments are first examined will be recorded as well as times and dates of additional readings.</td>
<td>Rather than using exact times, the time of reading will be converted to a date and the nearest sixth of the day. Early Morning (4-7:59 am), Morning (8-11:59 am), Afternoon (Noon-3:59 pm), Evening (4-7:59 pm), Night (8-11:59 pm), Late Night (midnight-3:59 am). This will make it impossible to connect a participant back to a specific lab through the timestamp.</td>
</tr>
</tbody>
</table>
### Course Section
CIS 2500 is a large course with several lab sections. The lab section for each participant will be recorded as a means of grouping participants at the end of the study.

Knowledge of the lab section of the student poses no risk to participant identity.

<table>
<thead>
<tr>
<th>Course Section</th>
<th>CIS 2500 is a large course with several lab sections. The lab section for each participant will be recorded as a means of grouping participants at the end of the study.</th>
<th>Knowledge of the lab section of the student poses no risk to participant identity.</th>
</tr>
</thead>
</table>

### Lab Grading Date
Students in CIS 2500 may be permitted to have their lab graded at a different time than their lab section. The date of lab grading will be recorded to allow experimenters to calculate the time between first reading a lab and having it completed.

Only a date, not a time will be recorded. Since many students will have their labs graded on the same date, it will not be possible to identify an individual from the date.

<table>
<thead>
<tr>
<th>Lab Grading Date</th>
<th>Students in CIS 2500 may be permitted to have their lab graded at a different time than their lab section. The date of lab grading will be recorded to allow experimenters to calculate the time between first reading a lab and having it completed.</th>
<th>Only a date, not a time will be recorded. Since many students will have their labs graded on the same date, it will not be possible to identify an individual from the date.</th>
</tr>
</thead>
</table>

### Lecture Attendance
Students in CIS 2500 will use clickers in class to promote student engagement. The presence or absence of clicker answers for experimental participants will be used to keep a record of lecture attendance.

Attendance records will be noted as present or absent, no records of the clicker interactions will be kept.

<table>
<thead>
<tr>
<th>Lecture Attendance</th>
<th>Students in CIS 2500 will use clickers in class to promote student engagement. The presence or absence of clicker answers for experimental participants will be used to keep a record of lecture attendance.</th>
<th>Attendance records will be noted as present or absent, no records of the clicker interactions will be kept.</th>
</tr>
</thead>
</table>

### Seating Choices
The seats in the lecture hall for CIS 2500 are numbered. Students will be asked to provide their seat number to allow experimenters to use information about seating choices. Seat numbers will either be captured using the clickers on a daily basis (if possible) or on the three in-class quizzes (if clickers do not permit numerical input).

The seats in the classroom will be grouped into sets of no fewer than 8 seats prior to the beginning of the semester. Student seating choices will be recorded as being in one of these groups which will reduce the chance that seating records will identify specific students.

<table>
<thead>
<tr>
<th>Seating Choices</th>
<th>The seats in the lecture hall for CIS 2500 are numbered. Students will be asked to provide their seat number to allow experimenters to use information about seating choices. Seat numbers will either be captured using the clickers on a daily basis (if possible) or on the three in-class quizzes (if clickers do not permit numerical input).</th>
<th>The seats in the classroom will be grouped into sets of no fewer than 8 seats prior to the beginning of the semester. Student seating choices will be recorded as being in one of these groups which will reduce the chance that seating records will identify specific students.</th>
</tr>
</thead>
</table>

### Time taken to complete quizzes and exams
When students hand in a quiz or exam, the time of handing in will be recorded. Since the start time of quizzes and exams is known, completion time can be calculated.

Times will be rounded to the nearest 5 minutes.

| Time taken to complete quizzes and exams | When students hand in a quiz or exam, the time of handing in will be recorded. Since the start time of quizzes and exams is known, completion time can be calculated. | Times will be rounded to the nearest 5 minutes. |
### Table 2- Expertise Measurements

<table>
<thead>
<tr>
<th>Short Name</th>
<th>Description</th>
<th>Identity Protection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekly Problem Set</td>
<td>Students in CIS 2500 will be given five multiple choice questions each week. The questions will be of the same difficulty as questions that will appear on the final exam. Answers to the weekly problem set do not affect student grades, but students will be required to finish the problem set before beginning the week’s lab. The student’s score on each problem set and the number of problems attempted will form one measure of student expertise with the material. (note. These problems are not part of the graded course material, they are practice problems.)</td>
<td>The problem set scores will be recorded as raw scores since they form a key measurement for this study. Since the scores do not affect grades in any way, and since the students will already know their own score, there is no reason for these scores to be posted anywhere. Individual scores on problem sets are not going to be useful in deducing the identity of students.</td>
</tr>
<tr>
<td>Self Evaluation</td>
<td>Along with the weekly problem set, each student will be asked to answer three self-assessment questions each week (the same questions, each week). The answers to these questions will form the second weekly measure of expertise.</td>
<td>The self assessment questions will be stored using only the experimental ID number for each participant.</td>
</tr>
</tbody>
</table>

### Table 3-Course Grades

<table>
<thead>
<tr>
<th>Short Name</th>
<th>Description</th>
<th>Identity Protection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab Marks</td>
<td>Lab marks will be used as part of the comparison between expertise ratings and final grades. Lab grades will be collected throughout the semester, but used only after the data collection period ends.</td>
<td>Labs are graded out of 2 points for the class (0=fail, 1=minimal pass, 2=meets expectations). Given that many students will have the same mark, lab grades are not useful in deducing identity and need not be transformed.</td>
</tr>
<tr>
<td>Assignment, quiz and exam marks</td>
<td>The other graded portions of this class are assignments, quizzes and the final exam. These items are graded on a percentage scale. The marks will be added to the data set only after the final grades are submitted to the university in April.</td>
<td>Grades will be transformed to the letter grade scale used by the University of Guelph prior to use in this study. (<a href="http://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-grds.shtml">http://www.uoguelph.ca/registrar/calendars/undergraduate/current/c08/c08-grds.shtml</a>)</td>
</tr>
</tbody>
</table>

(Any additional text or notes found in the image are included in the tables or context where applicable.)
Appendix B

Course Website

Moodle provided five categories of resources to the students. Course information as seen in Figure B.1 provided students with information about the format of the course including the course outline, topics for lectures and labs, advising hours of the professor and TAs, private tutor information, and a link to files used in the course.
The discussions, feedback and voting section, seen in Figure B.2 contained forums for course discussions, and off topic discussions as well as a forum in which instructor announcements could be found. Students were not required to visit the forums to read discussions, as they were provided the option to sign up for automated emails which contained posts as they happened or a daily digest of posts throughout the day. This section also included a link to the course wiki with notes from labs, tutorials, links and class specific directions about how to connect to the servers, IRC channel and SVN repository for the course.

Figure B.1: Course Information Section of Moodle

Figure B.2: Discussions, Feedback and Voting Section of Moodle
The assignments section of Moodle found in Figure B.3 contained descriptions of each assignment. Students did not submit their assignments through Moodle but were able to access their marks via the assignment links in Moodle. Assignments were submitted through the course SVN repository.

Figure B.3: Assignments Section of Moodle

Everything related to quizzes could be found in the quizzes section of Moodle as seen in Figure B.4. The three quizzes in the course were completed in class and the marks for each quiz were made available to students via Moodle. The clicker scores could also be found in this section. A bonus quiz that students took using Moodle was also available in this section.

Figure B.4: Quizzes Section of Moodle
Interactive lessons were available in the lessons section of Moodle as seen in Figure B.5. Lessons were segmented into the concepts taught in the course and allowed students to review or learn the material at their own pace.

![Lessons](image)

Figure B.5: Lessons Section of Moodle

Students were required to complete a weekly problem set in order to see their lab for the week. As such, problem sets were grouped together with their respective lab in the problem sets and labs section of Moodle, found in Figure B.6. Each week, students were asked to complete a problem set which consisted of both self-evaluation and skill testing questions. Once completed, students were then able to view the description of the assigned lab for the week.
Problem Sets and Labs

- Resources for Labs
- Problem Set 0
  - Restricted: “Available from 10 January 2011.”
- Lab Zero
  - Restricted: “Not available until you achieve a required score in Problem Set 0.”
- Problem Set 1
- Lab One
  - Restricted: “Not available until you achieve a required score in Problem Set 1.”
- Problem Set 2
- Lab Two
  - Restricted: “Not available until you achieve a required score in Problem Set 2.”
- Problem Set 3
- Lab Three
  - Restricted: “Not available until you achieve a required score in Problem Set 3.”
- Problem Set 4
  - Restricted: “Available from 6 February 2011.”
- Lab Four
  - Restricted: “Not available until you achieve a required score in Problem Set 4.”
- Problem Set 5
- Lab Five
  - Restricted: “Not available until you achieve a required score in Problem Set 5.”
- Problem Set 6
  - Restricted: “Available from 27 February 2011.”
- Lab Six
  - Restricted: “Not available until you achieve a required score in Problem Set 6.”
- Learning Styles Inventory (the other half of lab 6)
  - Problem Set 7
  - Restricted: “Available from 6 March 2011.”
- Lab Seven
  - Restricted: “Not available until you achieve a required score in Problem Set 7.”
- Problem Set 8
  - Restricted: “Available from 13 March 2011.”
- Lab Eight
  - Restricted: “Not available until you achieve a required score in Problem Set 8.”
- Problem Set 9
  - Restricted: “Available from 19 March 2011, 01:00 AM.”
- Lab Nine
  - Restricted: “Not available until you achieve a required score in Problem Set 9.”
- Problem Set 10
  - Restricted: “Available from 26 March 2011, 01:00 AM.”
- Lab Ten
  - Restricted: “Not available until you achieve a required score in Problem Set 10.”
- LSI results (lab 6 part 2)

Figure B.6: Problem Sets and Labs Section of Moodle