PATTERNS OF BEHAVIOR IN ONLINE HOMEWORK FOR INTRODUCTORY PHYSICS

A Dissertation Presented

by

Colin Fredericks

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May, 2007

Department of Physics

PATTERNS OF BEHAVIOR IN ONLINE HOMEWORK FOR INTRODUCTORY PHYSICS

A Dissertation Presented

by

COLIN FREDERICKS

Approved as to style and content by:

_____
William Gerace, Chair

_____
William Leonard, Member

_____
Alan Feldman, Member

_____
Guy Blaylock, Member

_____
Jonathan Machta, Department Head
Department of Physics

ACKNOWLEDGEMENTS

I would like to acknowledge the assistance of the following people in the preparation and execution of this thesis:

- Ian Beatty, for his encouragement, his many suggestions on data analysis and interpretation, and his willingness to act as a sounding board.
- Emma White, for her patience in listening to far more statistical jargon than anyone outside the field should be exposed to.
- My thesis committee, for their advice and encouragement.
- William Gerace, in particular, for his support in both my academic and real-life endeavours.

ABSTRACT

PATTERNS OF BEHAVIOR IN ONLINE HOMEWORK
FOR INTRODUCTORY PHYSICS

MAY 2007

COLIN FREDERICKS, B.A., RENSSELAER POLYTECHNIC INSTITUTE

M.A., RENSSELAER POLYTECHNIC INSTITUTE

Ph.D. UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor William Gerace

Student activity in online homework was obtained from courses in physics in 2003 and 2005. This data was analyzed through a variety of methods, including principal component analysis, Pearson's r correlation, and comparison to performance measures such as detailed exam scores. Through this analysis it was determined which measured homework behaviors were associated with high exam scores and course grades. It was also determined that homework problems requiring analysis can have an impact on certain types of exam problems where traditional homework does not. Suggestions are given for future research and possible use of these methods in other contexts.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This work was originally undertaken for one reason, and ended up being primarily concerned with something different. We initially wanted to answer the question of whether engaging with analysis-style questions leads to an improved ability to analyze situations later on. In the process of answering this question we created a significant amount of "machinery," especially in the area of quantifying student activity, and this slowly became the meat of the thesis. While the original question was eventually answered, it takes a back seat in this study to the various useful methods that were discovered and created along the way.

Answering the original question was deemed useful because it could inform the creation of future homework problems and classroom materials. While there is a certain amount of presumption involved — namely, that analysis is indeed a useful activity for a physicist or for other students taking physics — it seems almost certain to us that this presumption is correct. At the very least, this study has shown analysis-style problems to be no less valuable in general than traditional homework problems. More on testing the initial hypothesis can be found on page 106.

The measures that were created and examined in order to answer that question show a great amount of promise in several different areas. Given the proper approach in a course's homework and exams, they could be used to answer similar questions about the relation between homework questions and later performance. They can be used to gather information on courses and teachers, which may be useful to both those teachers and to

administrators. They could guide advice to students (and perhaps even feedback from teachers or automated systems) on what sorts of activity are most effective when it comes to doing homework. In short, the tools created to explore this thesis' original hypothesis may be more useful and powerful than the answer to that hypothesis, and it is for this reason that so much of this thesis is devoted to them.

This thesis, like most, uses a moderate amount of specialized language. Some of it is unique to the field of physics education, while other terminology was created specifically for this study. Those who encounter an unfamiliar term should refer to the glossary on page 134, or the literature review that begins with the next section.

CHAPTER 2

LITERATURE REVIEW

Full references for all items can be found in the Bibliography, starting on page 157.

2.1 The Validity of Online Homework

It would not be worth proceeding with this study if we were not first sure that electronic homework is at least as valid — that is, as effective and useful for the students, and as informative for the instructor — as traditional written homework.

This question has been examined by several prior studies, starting with Bonham and Beichner (2001), who concluded that web-based homework was no better or worse, overall, than traditional written homework. Their study found no significant differences between the performance (in laboratory, in exams, in visits to the tutoring center, and on the Force and Motion Conceptual Evaluation (Thornton and Sokoloff, 1998)) of classes using online homework and classes using written homework. Whether the class was algebra-based or calculus-based made no difference in this result. The only significant difference found was that students typically reported spending more time (approx. one hour more) on web-based homework than on written homework.

Cheng and Crouch (2004) showed that ungraded written homework was less effective than graded online homework when it came to increasing students' conceptual understanding as measured by the Force Concept Inventory (Hestenes, Wells, and Swackhamer, 1992).

Dufresne, Mestre, Hart, and Rath (2002) found an improvement in exam scores when students used web-based homework. The typical improvement was one third of a standard deviation, so most improvements were not statistically significant. They also found a cost savings of roughly $130,000 per year by switching their department's large-enrollment courses to online homework.

Bonham, Deardorff, and Beichner (2003) put their results and those of Dufresne in better terms than we could:

> "Thus, we conclude that we have established the null hypothesis that, in the case of introductory university-level physics with standard lecture sections using typical end of chapter problems, there is no significant difference in student course performance between Web-based homework with computer grading and homework written out on paper and graded by hand. Comparison with recently published work of (Dufresne et al., 2002) is instructive. In that study, the introduction of Web-based homework generally increased amount of homework that was graded and student time on-task, and gave students some assistance in solving problems when errors were made. This suggests that the medium of Web-based homework is not intrinsically more effective that traditional paper-based homework, but doing homework in general has pedagogical value, and that additional support and feedback enabled by the medium may be of real value." (Bonham, et. al., 2003, pg 1066)

Cole and Todd (2003) again noted no significant difference between the performance of students using written or online homework, despite using "multimedia homework with immediate rich feedback." They noted suspicions of bleed-through between their experimental and control sections, as students in pen-and-paper sections sometimes used the logins of the students in electronic homework sections, in order to receive the feedback the electronic homework provided. This, in itself, may be a noteworthy anecdote.

2.2 Student Views and Opinions of Online Homework

Jones and Kane (1994) examined students in an introductory mechanics course over 27 consecutive semesters, finding that students consistently rated a computer-based instruction system higher than their textbook, homework, and lecturer when it came to how much it helped them learn during the course. Notice the year this was published — this was a computer-based system, but not an online one.

Some studies (e.g., Johnston 2002) have shown that students have a preference for web-based over written homework, despite the minimal difference in outcomes. They found that students consider the benefits of online homework to outweigh its drawbacks.

Lust and Vuchetich (2004) compared student performance in and acceptance of an online pharmacy calculations course, finding that students rated the online course higher than the campus-based course. Students on campus performed better in the first year of the study, but the two groups performed equally in the following two years.

Hauk and Segalla (2005) surveyed 19 college-level algebra-based physics courses, finding that there was no significant performance difference between those that used WeBWorK (an online homework system) and those that assigned pen-and-paper homework. They also made a detailed examination of student and professor reactions to the system, finding that the opinions of students and professors tracked each other quite reliably. Students' complaints about the system arose primarily when the system's behavior contradicted Spangler's (1992) list of beliefs that college students hold regarding

mathematics. For example, WeBWorK allowed students to enter answers in symbolic form, such as entering (7-1)/3 instead of 2. Either answer would be marked correct, and some students found this to be a problem, holding to the idea that mathematics problems have only one right answer.

Clarke, Flaherty, and Mottner (2001) examined students' opinions of various different online components in a face-to-face Internet Marketing class. One of the components was online homework, which was found (in combination with other factors) to positively influence students' opinions of their overall learning, their ability to get a job, and their expected performance in that job.

Other studies can be found regarding student opinions and behaviors in online courses, but both of the courses we examined were face-to-face on-campus courses, so these references were not deemed relevant to this study.

Overall, students seem to appreciate and benefit from online homework at least as much as they "appreciate" and benefit from ordinary homework.

2.3 Categorizing Problems

During our study we have categorized problems that appear on homework and exams (see page 15 for details). Previous work in this area includes:

Chi, Feltovich, and Glaser (1981) did the first work (to our knowledge) on problem categorization by experts and by novices. They found that novices categorized problems based on easily-observed surface features — block-on-incline, merry-go-round, etc. — whereas experts put problems into categories based on the physical laws and properties relevant to the system, such as energy conservation, torque balance, etc.

Thomas and Hume (1997) looked at the effect of problem complexity on quiz scores. They found that assigning complex multiple-step problems had positive long-term effects on all students, especially those who would normally receive lower grades. The complex problems in this paper are the closest analog we have found in the literature to this study's analysis problems. Here is an example they give of such a problem:

"As a second example we take deflection of a moving charged particle in a magnetic field. A standard text book problem is to calculate the velocity of an ion which has a particular radius orbit ... Our replacement scenario is to take a plate with two holes separated by a distance, with a magnetic field on one side a distribution of ions on the other. The question is to determine which ion velocities permit the particle to move through one hole and deflect so that it emerges from the second. The problem is made complex by stating that the ions have all velocities and any approach angle to the first aperture. ... The student is driven to consider a few test scenarios to identify what kinds of trajectories are permitted (the apertures must lie on a chord of the circular orbit). This leads to a limiting case and identification of a minimum velocity which permits the trajectory." (Thomas and Hume, 1997, pg. 2)

2.4 Research on Homework Behaviors

Stewart (1996) used extensive interview data to create a detailed "procedural model" for students' behavior; specifically, their behavior while solving traditional textbook problems. The model given in their work is much more process-oriented than our work, which focuses more on outcomes, but they do present a method for "the conversion of raw process data into measures of the educational value of a physics course." (pg. 205)

Elby (1999) used surveys of introductory physics students epistemological beliefs to examine their attitudes about grades and understanding, finding that their perception of "trying to do well in the course" differed greatly from their perception of what it means to "try to understand physics well."

Kotas (2000) conducted a study somewhat similar to our own work, analyzing log files from online homework to extract various measurements and correlations. The strongest predictor was procrastination on homework (a negative indicator), followed by interacting with the material on a daily basis (positive). Working on homework immediately after lectures was also found to be well correlated with later performance, though this relationship may or may not be causal. All of Kotas' indicators were at what our study would call the "gauge" level. See also Minaei-Bidgoli, below, who analyzed data from the same homework system using genetic algorithms.

Cooper and Valentine (2001) conducted a meta-analysis of nine different studies relating achievement to time spent on homework, in high school and junior high, showing increasing achievement as homework done increased towards 10 hours per week. They

also gave a review of homework-related research and discussed some of the reasons that this research has not had widespread effect. They found that many studies had a poor experimental setup, making them both unreliable and unable to prove causal relations. Also cited were the context dependency of many educational findings, the frequently contradictory messages that different studies put forth, political motivations on the part of both lawmakers and teachers, and these two groups' lack of understanding of statistical results from prior studies.

Kotas and Finck (2002) used surveys, log data from an online homework system, and institutional data to show that homework collaboration between students is well correlated with performance as measured by final grade, despite the asynchronous nature of online homework. This is a particularly useful result, as collaborative effects are one of the blind spots of our thesis.

Warnakulasooriya and Pritchard (2004) examined the effects of completing tutorial problems delivered through the online homework system Mastering Physics. They concluded that not only did twice as many students complete these problems correctly (compared to equivalent back-of-the-book problems), but that students who did these tutorial problems were also able to complete later problems more quickly, with fewer hints required.

The same team of Warnakulasooriya and Pritchard (2005) used metrics similar to our gauges, gathered from the Mastering Physics homework system, to classify problems as to their difficulty level, with high reliability. They then used this difficulty level to con-

struct an “item discrimination measure” that allowed them to predict students’ final exam scores with moderate accuracy (r = 0.643).

2.5 Engagement

Fredericks (no relation), Blumenfeld, and Paris (2004) presents a comprehensive overview of school engagement, describing the various different kinds of engagement that had been examined at that time. Engagement is categorized as behavioral, emotional, or cognitive; different means of encouraging it and measuring it are examined; and outcomes from engagement studies are reviewed. This review article may be quite useful for those interested in examining various forms of engagement, and includes dozens of references for further reading. For this study, we used this work primarily as a terminology guide for describing the sort of engagement we were interested in (see, for example, page 87).

2.6 PCA and General Data Mining References

The primary reference for Principal Component Analysis during this study was Malinowski's "Factor Analysis in Chemistry" (2002), which we read and referred back to for a fundamental understanding of the method. Despite the difference in applied subject matter, the terminology was sufficiently familiar as to make this a more effective reference than many others that dealt with PCA in psychology, education, or political science.

Minaei-Bidgoli (2004) used clustering ensembles, genetic algorithms, and other state-of-the-art methods to obtain a highly accurate predictor for student performance based on online homework log files. The main focus of Minaei-Bidgoli's thesis is data mining rather than educational research, and the methods used are significantly more powerful and sophisticated than the ones we employed; nonetheless, some of the results there are comparable to ours. See Appendix 3 (on page 143) for more.

CHAPTER 3

DATA SETS

This chapter describes the data collected in the course of this study. The first section is the most detailed, with later sections referring back to it. Samples of the data in varying degrees of rawness can be found in Appendix #5, starting on page 147.

3.1 Physics 151, Fall 2003

This course provided data from homework, exams, and surveys. The surveys are discussed in their own brief section later on.

All homework in Physics 151 was online. Correct answers gave full credit if inputted before the due date, and half credit afterwards. We were provided with the full text of the questions asked in the course's online homework assignments, which was useful in categorizing the OWL log files.

The logs from OWL came in four separate files: a "sessions" file, a "modules" file, a "students" file, and a "question log". We were informed by the OWL development team that the first one contained somewhat unreliable information, so they were ignored. The second file contained due dates and descriptions for each assignment. The third file was useful in matching students' ID numbers in OWL to their ID numbers in the course — the two systems used different identifiers. It was the last file, the question log, which was the most fruitful in this study, and which provided the vast majority of our raw data.

OWL presents students with a web page including a written question, answer boxes, a "submit" button, and (typically but not always) a diagram or picture pertinent to the question. Each time a student submits an answer, the OWL system logs the following data into the "question log" file:

- A Student ID number, which is entirely internal to the OWL system and different from the ID number assigned by UMass.
- The number of the module the student is currently working on.
- The current session number.
- The number of the instructional unit the student is currently working on.
- The current question number.
- The score received on this particular attempt on this question.
- The date of the submission.
- The time of the submission.
- The number of seconds between the opening of the web page and the time the "submit" button was pressed. (Entitled "Seconds To Respond;" see comment for Physics 181, in which this value was calculated differently.)
- Any variables stored for this attempt. (OWL randomizes questions for each student, and can do so for each attempt as well. This piece of data can be used by the OWL development team to double-check the answers given by OWL.
- The correct answer to the current attempt of this problem.
- The response given by the student for the current attempt.
- The number of the current attempt.

Using information from the course syllabus and the homework problems, the following pieces of data were added to each line: the due date and time for the problem, the

type of problem being attempted (Analysis, Traditional, etc., see below), and the type of assignment being worked on (Homework, Lecture Prep, Tutorial, etc.). Some pieces of data, such as the instructional unit number or the response given for the current attempt, were never used in analysis, but none of this data was discarded. The data was then split into smaller files, since Microsoft Excel (used as the primary data analysis tool) does not handle more than 65,536 lines of data in a single worksheet. More advanced analysis tools were not easily available at the time of this analysis, and it was later decided that analysis through Excel was sufficient for our purposes.

From here the data was processed according to the needs of the study at the time.

As for "performance" data, there were four exams in this course: three "midterm" exams and a required final exam. All were composed entirely of multiple-choice questions, answered on bubble sheets and machine-graded. Each midterm contained 25 questions, with the final exam containing 33. On all exams a partial credit scheme was used, allowing students to "bubble in" more than one answer if they were unsure. Questions were worth four to six points, with the possibility of receiving negative points if multiple bubbles were filled and the right answer was not chosen.

We had access to individual scores on each problem for each student, as well as the text of the problems themselves. Actual student responses were available as well (indicating which answers were chosen), but were not used for this study. Names and student ID numbers were provided.

For all exams, problems were categorized as Analysis, Conceptual, Definition, or Traditional. Because so few questions were Conceptual or Definition questions (typically no more than two of each type on each exam), these question types were ignored. Each student's subscores for Analysis and Traditional questions were calculated and retained for later use along with their total exam grades.

3.1.1 On Problem Types

Four categories were used to sort problems on homework and exams: Analysis, Conceptual, Multiple-choice/Definition, and Traditional. This section describes how we categorized these problems and gives some examples for them. See page 80 (Section 5.2, "Evidence for Problem Types") for some results regarding the validity of sorting things this way.

Analysis questions are those that examine a physical situation without the intent of immediately solving it quantitatively. They often rely on multiple concepts, have multiple means of solution, and do not explicitly draw attention to the needed or critical concepts. Comparisons between physical quantities, determination of solubility, critical examination of an existing solution to a problem, matching a situation to its representation (or vice versa), all of these are analysis problems. All of these types of questions can appear both in homework and on exams.

Here is an analysis problem from Physics 151. It involves calculation in parts a, b, and c, and those would make a fine traditional question. Part d (which was worth as much

as the other parts together) provides the analysis segment: it *can* be done through calculation, but it is much faster and more efficient when done through analysis.

> Q: Two disks, the larger having a radius of 50 cm and the smaller having a radius of 25 cm, are attached to each other and are mounted on a fixed axle such that they can spin frictionlessly. A heavy object of mass 55 kg is attached to a rope that is wound around the smaller disk. A person is pulling on a second rope that is wound around the larger disk, causing the object to lift slowly off the ground.
>
> (a) What is the tension in the rope supporting the object?
> (b) What is the tension in the rope being pulled by the person?
> (c) How much work does the person do to lift the object by 80 cm?
> (d) If the smaller disk were even smaller, how would each of the preceding quantities change in order to lift the object by 80 cm?

Here is an example of a "pure analysis" question from Physics 151, with no calculation required. It is classified as an analysis problem because it relies on multiple physics concepts rather than a single item.

> Q: A wheel of radius 0.2 m is mounted on a frictionless axis. A massless cord is wrapped around the wheel and attached to a 9-kg block that slides on a frictionless surface inclined at an angle of 20° with the horizontal. The block is released from rest at the top of the incline.
>
> (a) How many forces are exerted on the wheel?
> (b) How many of these forces exert torques about the center of the wheel?
> (c) Is the tension in the string larger than, smaller than, or equal to the component of the weight of the block parallel to the incline?

Conceptual questions require understanding of a single concept (such as Gauss' Law or Newton's Third Law). The principles involved in a conceptual question are more complex than those seen in a definition question (which might involve the definition of

work, or the relation between linear and rotational variables). Conceptual problems are distinguished from analysis problems in that the former typically draw a student's attention to the needed principle, and are essentially single-step problems with only a single method of solution.

Here is a conceptual question from Physics 151 that relies specifically (and solely) on vector decomposition.

> Q: A sled is pulled up a shallow 10° incline, with a force of 19.0 lb, directed at 47° upwards from the surface of the incline.
>
> (a) What is the horizontal component of this force?
> (b) What is the vertical component of this force?
> (c) What is the magnitude of this force?
> (d) What is the angle that this force makes with the horizontal?

Here is another conceptual problem from the same course, using the idea of conservation of momentum with less emphasis on calculation:

> Q: A pair of stationary masses $M_1 = 6$ kg, $M_2 = 3$ kg have a compressed spring between them and are tied together by a string. The string breaks, the spring releases, and the masses fly apart. The velocity of the first mass is measured to be $v_1 = 3$ m/sec toward the left. What is the total momentum of the two masses after the string breaks

Multiple Choice and Definition questions were grouped together for the purpose of this study. These two seemingly different types of question are put in the same category because of how quickly they can (potentially) be answered. Online homework turns multiple choice questions into a guessing game for many students, and questions based

on a single definition are either answered immediately, or after a quick search through the textbook.

Here is a typical multiple-choice question from Physics 151, in which students can pick one or more answers:

> Q: If the acceleration of an object is zero, which of the following statements is correct?
>
> a. The velocity is decreasing.
> b. The velocity is constant.
> c. The object has no velocity.
> d. The velocity is increasing.
> e. None of the above

Definition questions are based on the application or knowledge of a single definition. Here is a sample definition question, relying solely on student's knowledge of the formula momentum = mass x velocity.

> Q: A car has a mass of 1000 kg and is moving with a speed of 10 m/s. What is its momentum?

Traditional questions are the variety one most typically sees in physics textbooks. Their primary feature is that they require numerical calculation. While they may require some element of conceptual knowledge or some amount of analysis, these features are ultimately secondary to the mathematical work needed. See also Analysis, Conceptual, and Multiple Choice problems.

Here is a typical Traditional question from Physics 151:

Q: The Zero Gravity Research Facility at the NASA Lewis Research Center includes a 146 m drop tower. This is an evacuated vertical tower through which, among other possibilities, a 1 m diameter sphere containing an experimental package can be dropped.

(a) How long is the sphere in free fall?
(b) What is its speed just before it reaches a catching device at the bottom of the tower?
(c) When caught, the sphere experiences an average deceleration of 23.2g as its speed is reduced to zero. Through what distance does it travel during the decceleration?

3.2 Physics 181, Fall 2005

Physics 181 provided us with data from homework and exams.

The OWL data for this course was essentially identical to that for Physics 151. There were fewer lines of data, but the only real difference was in the way the time taken for each attempt was calculated.

Because of a difference in preference settings for the courses, Physics 151 students were required to use different random numbers on each attempt, while Physics 181 allowed students to use the same random seed when they retried questions. We have anecdotal evidence from conversations with students that they notice this difference, but because we have only two classes to compare, we are unsure of what its effects are and whether students treat homework problems differently as a result.

This change in settings means that the Seconds To Respond value for each attempt in Physics 151 was measured from when the web page for the current attempt was first

opened to when the "submit" button was pressed for that attempt, while Physics 181 instead calculated this value from the first time the web page was opened for the first attempt on the problem to the time when the "submit" button was pressed for the current attempt. The timer was only reset when a student switched to a new problem, or logged in again for a new session. In other words, 181's STR measurement was cumulative, while 151's was not. Simple subtraction cast the data from 181 into the old format for better comparison.

As for tests, Physics 181 had two midterm exams and a required final exam. Unlike Physics 151's exams, these were hand-graded, with a different format. They began with a short multiple-choice section, followed by four or five complex, multiple-part problems. Some parts of the problems emphasized calculation, while others required more conceptual thinking and analysis. Because we could not classify entire problems as Analysis or Traditional, and did not have part-by-part scores available, our data for the Physics 181 exams remains broken down by problem number. Each question was worth the same number of points, with the multiple-choice section being worth as much as a single question.

The midterm exams in this course (though not the final) were given twice: once in a timed setting, and once as a take-home exam allowing any resources the student desired to use, including other students. The total grade on each exam was the average of the timed and untimed versions.

3.3 Supplemental Data

When data was taken from Physics 181, the Seconds To Respond data showed a (seemingly) anomalously high result — four times the response time that was seen in Physics 151. To make sure that this data was valid, another class was examined. Physics 152 in Spring 2003 was chosen because most of the students in it were somewhat more experienced than those in Physics 151, and it was thought that they might be a better model for the Physics 181 students. Histograms from Physics 152 indicated twice the response time seen in Physics 151, and Physics 181's time and date stamps agreed with its Seconds To Respond data. We therefore have assumed that there was in fact no anomaly in Physics 181's Seconds To Respond, and the difference in STR averages are correct.

In an unrelated item, Physics 151 gave fourteen course feedback surveys, one each week. Each survey contained a standard set of repeated questions, plus a set of topic-specific questions. Topics included electronic homework, the classroom response system, the grading format for the exams, the lecture prep assignments, and other matters pertinent to the course. These questions primarily addressed student attitudes, though there was occasionally some self-reporting of (for instance) time spent on homework. Those surveys that we examined in detail were primarily those that asked about online homework, and the results of this examination can be found starting on page 93.

3.4 Discarded Data

There were several pieces of data available to us that were discarded for the purposes of this study. This section gives a quick overview of each one and why it was not used.

Scores on fourteen weekly quizzes were available for Physics 151, with each one consisting of a small number of multiple-choice questions. However, completion of these quizzes was very spotty, making it difficult to compare students on this basis. There were also some quizzes for which the grade sheet and the quiz itself disagreed on the number of questions the quiz contained. Rather than introduce another level of unreliability into our analysis, we chose not to include data based on the quizzes.

Seven topic surveys were used during Physics 151. Five of them determined students' existing knowledge about and confidence with various topics: motion, forces, energy, momentum, and the mathematical prerequisites for the course. One was used as a post-test, and another as an attitude survey. Again, data here was sparse, and the occasional use of free-response format made analysis difficult, and this data was ignored.

Physics 181 assigned both online and written homework. The grades for the written homework were not broken down by problem or by problem type, so most of the analysis that was done for online homework could not be replicated here, and the written homework scores were not used for this study. An honors section of Physics 181 included major homework projects; however, most of the class took the course without honors, so this was ignored.

Eight surveys were given to students in Physics 181. Six of them were identical, given bi-weekly. They were designed to give the professor feedback on which problems students had the greatest amount of trouble with, and what resources the students used from week to week. The remaining two were a pre-post set that asked students about their backgrounds, what practices were most important to doing well in physics, what purpose they thought various components of the course served, and various other items. Because the responses to these surveys were somewhat sparse and highly idiosyncratic, and because similar survey information from Physics 151 was not included, this data was not used in our analysis.

Both courses used a classroom response system with clickers; specifically InterWrite PRS (from GCTO CalComp). Data from PRS was available from both courses as well, but Physics 151 did not have questions matched up to student responses, which makes analysis all but pointless. This data was ignored in our study.

CHAPTER 4

METHODS

This chapter discusses the methods that were used and created for the purposes of this thesis. We begin by discussing the limitations of our methods, go on to looking at the mathematical tools used, and end with a discussion of the procedures used to draw conclusions and create higher-level data from raw data.

4.1 Limitations of These Methods

This section describes the problems and limitations inherent in our methodology. These caveats should be kept in mind while reading the rest of the thesis.

4.1.1 Limitations Inherent to the Data Set

The majority of the data for this study was gathered "behind the scenes" from student's actions on their online homework. The advantage to obtaining data on the students' gauges from OWL is that there is no "measurement error" involved (unless one counts the fact that OWL rounds times to the second). Unlike most scientific experiments, there is no intrinsic uncertainty in, for instance, the number of attempts a student has taken on a particular problem. A student's Seconds To Respond gauge may not be a good or informative number to use for a particular purpose (as it does not measure time spent on homework while away from the computer), but it does not contain any inherent error.

The primary drawback is that only actions related directly to the submission of attempts are recorded. This data only lets us see a tiny slice of the students' actual activities, giving us no information about any studying, time spent working problems or discussing them in groups, cheating that may have taken place, and untold dozens of other outside factors that make the data noisy and occasionally unreliable.

A smaller amount of data was obtained through online surveys given throughout the courses. Performance ratings, such as the students' exam and course grades, are the only information that does not come directly from OWL. Even a student's test scores can be suspect — is a student's test score low because of the gauges we were able to measure, or because they were up until 4 AM studying for their other test that day? More on this topic is discussed in the "Implications for Future Research" section, starting on page 113.

Despite the impact of outside factors, there are still many significant correlations between our measurements and performance. We are by no means saying that we believe our data is made invalid by these outside circumstances. Using the guideline that the square of a correlation factor indicates the amount of variance accounted for, our best correlations account for roughly 65% of the variance between homework activity and course grades. Comparing eigenvalues obtained from Principal Component Analysis can also estimate the amount of variance one can reasonably account for within a particular set of data. Within the OWL homework, this method indicates that over 70% of the variance can be explained within the first four factors. There will necessarily be a significant amount of variance unaccounted for by even our best models, but we believe that this amount is low enough to make the study worthwhile.

4.1.2 Limitations Inherent to the Analysis Methods

The most severe limitation of our methodology is its restriction to purely linear analysis.

Pearson's "r" correlation coefficient detects only linear relationships between measurements (that is, between gauges and performance, or between different gauges). While scatter plots do indicate that most of these relationships are indeed linear, not all of them are. As an example, the Seconds to Respond (STR) gauge in Physics 181 shows a significant correlation to performance until it reaches about 600 seconds. If one examines only STR measurements greater than that time, performance and STR are essentially uncorrelated. The relationship is not a simple functional one (i.e. not clearly quadratic, cubic, gaussian, etc.), but more of a general scattering. As STR increases beyond 600, the average performance doesn't change, but the spread in it has increased to the extent that it overshadows any actual relationship that may or may not exist. Creating a "cutoff" value of 600 seconds creates a gauge with a larger number of significant correlations. This is not necessarily a more or less valid measurement than, say, putting the cutoff at 1200 seconds; it merely results in stronger correlations.

An argument can be made that setting cutoffs in general makes the data more robust, since it removes outliers from the data set. However, every time we change the cutoffs, we must remember that we are creating a new and different measurement. "Seconds to Respond under 20 minutes" is a different measurement from "Seconds to Respond, no cutoff."

Another constraint to linearity comes from Principal Component Analysis. PCA assumes — in fact, it would be more accurate to say it creates — a multilinear relationship between input and output factors. This is a completely different linearity assumption from the one just discussed, as Pearson's "r" *assumes* linear relationships between its two inputs and outputs a single number, while PCA outputs reduced factors which are, *by definition*, linear combinations of the input measurements. It is certainly possible that the real underlying factors behind student activities (if they can even be quantized) have nonlinear relationships to the students' measured gauges.

An attempt was made to find a systematic way to treat nonlinear relationships; the discussion of the Eta correlation ratio (starting on page 46) describes one approach and the reason it was rejected. No suitable, standardized method for dealing with nonlinear relationships was found, and so this limitation remains.

An additional limitation of PCA that was realized after some amount of consideration was the orthogonality of the factors returned. Student activities are not necessarily orthogonal — much of their conduct is interrelated, and one type of behavior can trigger another. The factors returned by PCA are still useful, but for this reason they may be difficult to interpret as "physically" (or emotionally or scholastically) meaningful behaviors.

The unavailability of median measurements in Excel's Pivot Table reports provides another limitation to our analysis. Pivot Tables were used to significantly speed the transformation of raw data into gauges and behaviors, allowing us to analyze a semester-long course in just a few days (and also to avoid the errors in mathematics and data handling that would have occurred had the author attempted to write his own code). Unfortunately,

it provides only averages, not medians. In a data set of this size and spread, medians are often more informative and robust than averages, being less influenced by outliers (see, for instance, Press et. al.'s Numerical Recipes in C, 1992). Some gauges have few outliers, so we believe this is not a major problem, but others have several rather large outliers that bring the average up significantly. The Elapsed Time gauge is a striking example: average times are on the order of hours, while median times are on the order of minutes. It is hoped that such outliers are present in many different students, so that correlation factors will be relatively unaffected by the change, but the possibility exists that gauges and behaviors with large numbers of outliers in a few students will show unreliable correlation factors.

4.1.3 Accumulation of Uncertainty

While the individual data points recorded are effectively without error, each of the correlation factors examined in this study has a non-zero chance of being a false positive (or false negative, as the case may be). Given that we have examined several thousand correlation factors, it is a near-certainty that some of them indicate a relationship that does not truly exist.

This problem is compounded by comparisons between correlation factors (and the possibility of false significance in such correlations), and by the occasional use of correlations between sets of other correlation factors and performance. At that stage it becomes nearly impossible to meaningfully rate the probability of significance for the resulting measurements.

Because of this, we have tried to keep to the most stringent requirements for reliability in correlation factors, using only those with $p = 0.01$ or lower, and preferring scores with $p \leq 0.001$. The large number of students available (211 in Physics 151, 55 more in Physics 181) means that it is not very difficult to find such highly reliable correlations.

4.1.4 Behaviorist Bias

Because the majority of the data relates to students' measured actions and performance, there is a natural behaviorist bias to this thesis. Any thought processes internal to the students are essentially invisible to this study: we can see only what they have done, and only in a limited manner. Because this dissertation does theorize the existence of inner states (and even attempts to quantify them to some extent), our approach would properly be called theoretical behaviorism. However, attempting to consider student thought processes solely on the basis of such data is difficult, and can easily be misleading.

The gauges measured in this study are data recorded directly from student activity, and are named literally. However, the "behaviors" that are combinations of those gauges have intentionally been given names evocative of student mental states. While examination of their correlations and interrelations supports these names, it is vital to remember that the same set of outward symptoms (i.e. the same combination of gauges) can represent more than one internal mental state. These are truly behaviors in the behaviorist sense, and not indisputable mental states.

This was even more true in the case of the student homework types that we attempted to generate. A full discussion of the various approaches attempted, and the reasons for their rejection, can be found starting on page 149.

4.1.5 Interpreting Behaviors

Generated from the gauges measured in this study are a set of "behaviors," which are intended to correspond to certain student mental processes. Whether they actually do correspond to said mental states is a matter for future research. While the results from combinations of behaviors does give us reason to believe that their names have some validity, they also require a certain amount of interpretation. See page 103 for more information on combinations of behaviors and their clarification.

It is worthwhile to take the names of these behaviors with a grain of salt. In the end, they are only linear combinations of gauges. Thinking of behaviors as tentative names, open to interpretation, will be more fruitful than locking one's mind into a single meaning for each word.

4.2 Mathematical Methods

There were two primary mathematical tools used during this study: Pearson's "r" correlation, and principal component analysis.

4.2.1 Principal Component Analysis

We begin with an overview of the method, and go on to look at characteristics of both random and non-random data as seen through PCA.

Principal Component Analysis (or PCA) is a multilinear data analysis method. It operates on a two-dimensional data matrix consisting of a number measurements made on a (typically different) number of objects. For the method to work properly, the number of objects must be equal to or greater than the number of measurements. The more objects in the data set, the more reliable PCA will be. The goal of PCA is to discover a minimal set of meaningful underlying factors, from which the original data set can be recreated with significant accuracy.

PCA differs from Factor Analysis (a similar procedure) in that it does not assume an underlying causal model — it is only a variable reduction scheme.

The process of PCA requires many steps, with each step having multiple options as to how to proceed. While the literature indicates that equivalent (or at the very least, valid) results can be obtained through all these different options, we have found that some options are more fruitful than others for our particular data sets.

The objects in our data set are students, and the measurements are the gauges for each student. Using the measurements we chose gave us an 18 x N matrix of data (with N the number of students in the class or a desired subset of it).

Once the initial data matrix has been created, the first step is normalization. PCA inherently favors larger numbers, and the gauges measured in this study vary widely in scale. For instance, Seconds to Respond ranges from 1 to 1200 (after removing the worst outliers), whereas Average Score per Attempt ranges from zero to one. Since it was our desire to give an equal weighting to all gauges at the beginning of the analysis, each measurement was normalized by subtracting the average and dividing by the standard deviation. All the different measurements (gauges) then had an average of zero, and a standard deviation of one. This was performed separately for each type of problem, in each group for which PCA was conducted.

After normalization, we had the option of creating a covariance matrix to work from. Our references indicated that this was not necessary, but we found that it aided in the later discovery of the proper number of factors. The covariance matrix is created by premultiplying the data matrix by its transpose, creating a square matrix.

The covariance matrix is then decomposed via Singular Value Decomposition (SVD) into abstract factors in a “row matrix” R, and a “column matrix” C. These abstract factors are linear combinations of the original measurements, but are not physical meaningful — they are merely one of a large number of different ways to recreate of the original

matrix. Each factor comes with an eigenvalue, indicating its relative importance, and various other pieces of information can be calculated for each.

Again, using SVD at this stage is merely one of a number of possible options for this step — the power method, the Jacobi method, the method of non-iterative partial least squares, and likely others as well, are also valid. We chose SVD because it was the most easily available method; however, Malinowski (2002) also indicates that it is "the most stable under the widest range of applications."

Once the abstract factors are obtained, one either applies a multidimensional "rotation" on the factors to make them physically meaningful, or employs "test vectors" and target transformation to check for the existence of a particular suspected factor. The latter is more useful for hypothesis testing and theory confirmation, while the former (our choice) is more exploratory in nature.

After having chosen to rotate the factors, one must choose a rotation method and a number of factors to act on. SVD will give as many factors as the original data set had measurements, which is not at all desirable. One wishes to create a minimal set to reproduce the majority of the original data, not a maximal set for perfect reproduction. The goal is to understand what's going on behind the scenes, and what gauges might spring from common causes.

To choose a number of factors, one seeks a minimum in the imbedded error function, a minimum in the factor indicator function, performs a "scree test" with a graph of the eigenvalues, and/or picks those factors with eigenvalues greater than unity. In a perfectly

orderly data set, all methods should agree; however, the data sets in this study were anything but perfectly orderly. Imbedded error graphs, especially, rarely if ever showed minima. In many data sets the graphs of these functions were not very far from those generated by a set of random data (see page 36 for how to identify random data). Where two methods for choosing the number of factors agreed, we used the number they indicated. Where there was no agreement, we typically did not continue to examine that data set, as PCA did not seem to be a fruitful method to use in such cases.

Rotating the chosen factors comes next. A plethora of rotational methods exists, each maximizing or minimizing a particular quantity. Some of them perform orthogonal rotations, which keep the axes of the factor space orthogonal to each other, while others are oblique, and do not. We chose the varimax rotation method, which maximizes the squared variance of the factor loadings and maintains orthogonality between the factors. It is one of the more common methods used in the social sciences, and should be suited to analysis of educational data. Note that rotations do not change the factors' eigenvalues, and thus their relative importance stays the same.

The rotated factors are, finally, meaningful linear combinations of the original measurements. Exactly what they mean, and whether they represent a cause or an effect, is subject to interpretation. It is vital to remember that these rotated factors are not necessarily related to performance on quizzes, examinations, or even the homework grade. They merely recreate the original data to some degree of accuracy. It is also occasionally useful to view them as orthogonal basis vectors in a new measurement space, and we will use that terminology on a regular basis.

The computer program Igor Pro (version 5.05b1) was used to perform PCA on our data. Igor is published by the Wavemetrics company. Our primary reference for PCA was Edmund Malinowski's "Factor Analysis in Chemistry", 3rd edition. ISBN 0-471-13479-1, © 2002, Wiley, NY. This was also the reference used by Igor.

4.2.1.1 Identifying Random Data

In the process of testing Igor's PCA routines, we performed (at the suggestion of Dr. Beatty) a Principal Component Analysis of a set of random data. A set of 1328 random data points was generated by Excel, created to follow a gaussian distribution with an average of zero and standard deviation of one. A histogram (below) confirmed this shape.

The histogram, imbedded error, factor indicator function, and eigenvalues are plotted below. These are characteristic shapes, having little variation from one random data set to another. The histogram includes a gaussian best-fit line. As one can see, there are no noticeable minima in the embedded error function or factor indicator function, nor would a scree test on the eigenvalues show any useful results.

In attempting to analyze certain of our data sets, we discovered that the shapes of their functions were quite similar to those of a random set. This implied that there was very little underlying order in these particular sets. Examining the cross-correlation tables

Fig. 4.1: Random Data Histogram

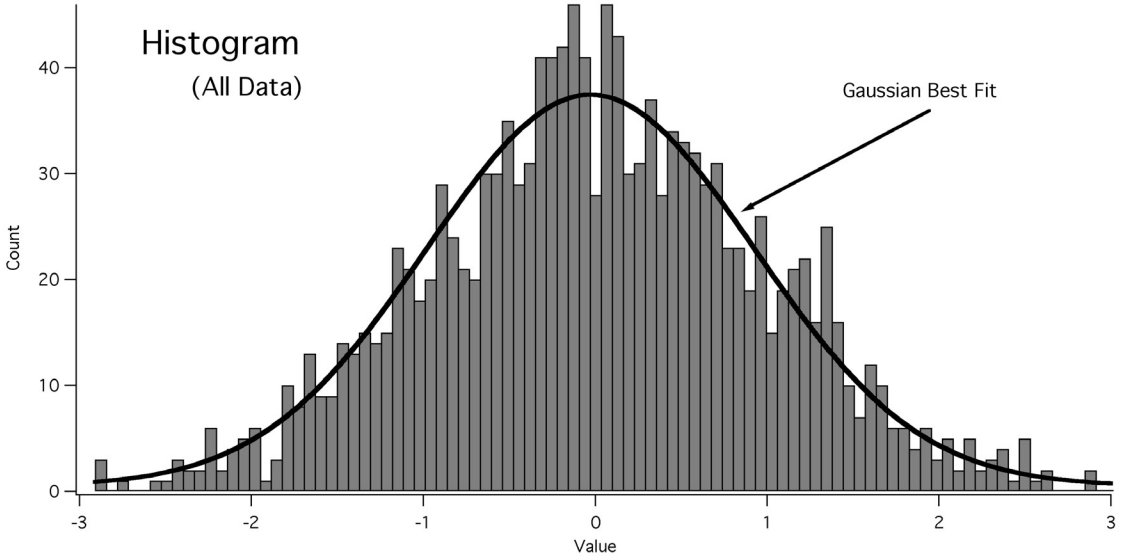
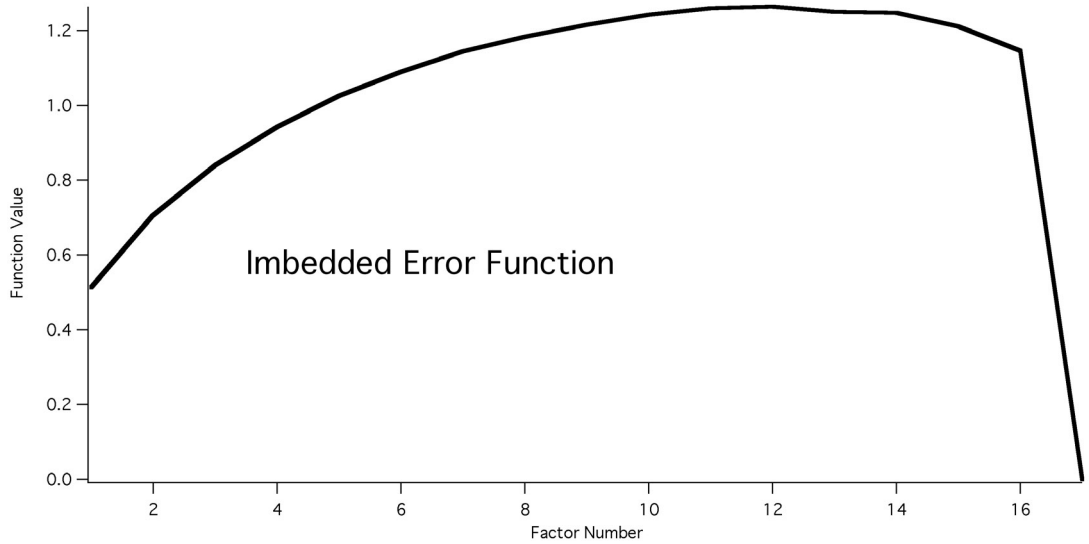
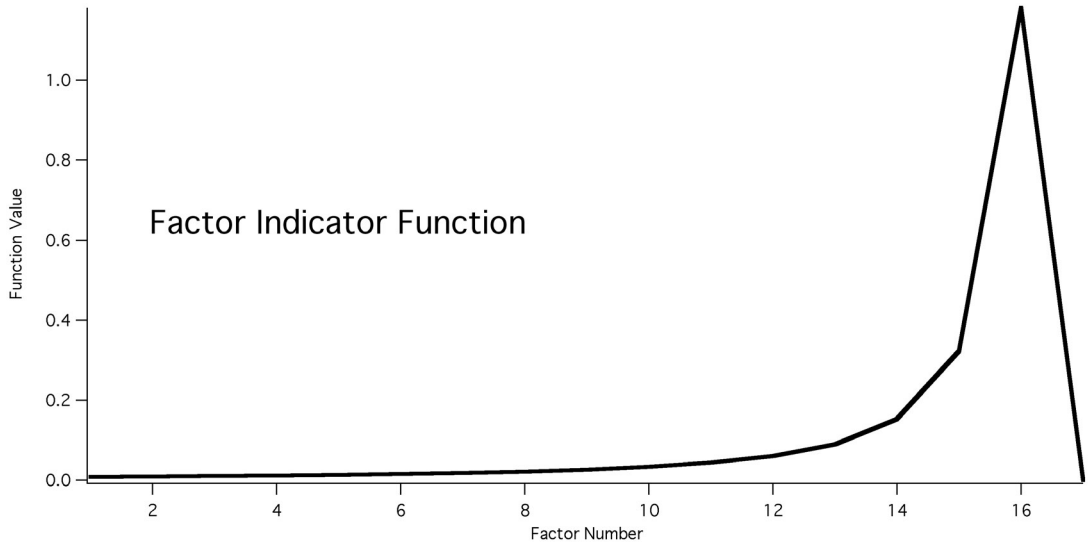
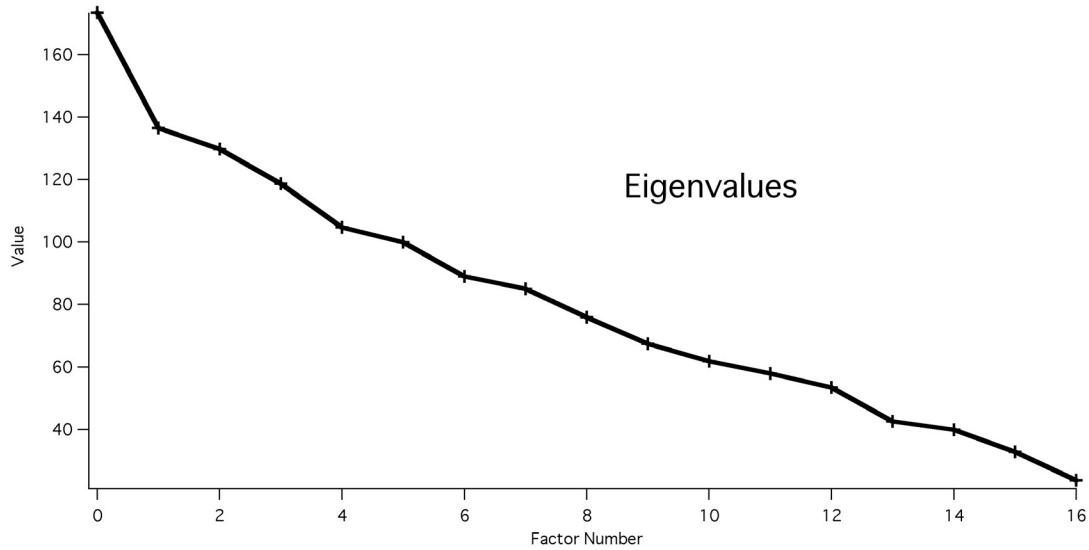


Fig. 4.2: Random Data Plots







for these data sets gave further confirmation: these data sets showed very few strong correlations between different measurements. Data sets that Igor showed little trouble with (such as our next test) had more and stronger correlations between measurements.

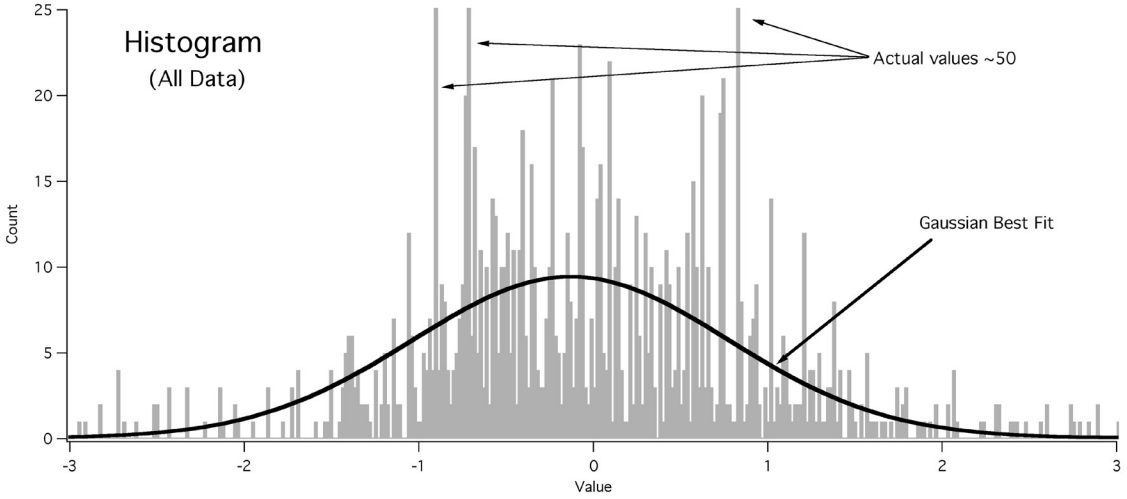
PCA relies on "hidden" connections between variables, and, as one might imagine, it works in a less-than-optimal manner when those connections are weak. In the case of pseudo-random data, where connections between the variables should be essentially nonexistent, PCA can still yield results, but they will have no actual meaning. This is why we avoid discussing PCA on gauges from certain problem types — random data is not the only thing that will yield these results. If one measures variables that are totally independent, there will be no underlying connection between the variables. For the purposes of PCA the data might as well be random. In educational data it is relatively rare to obtain measurements that are 100% uncorrelated, but in our study it was not difficult to find measurements with a very low amount of correlation. Examining the cross-correlation tables (see page 83) will show many gauge combinations with a correlation factor below the threshold for significance; PCA analysis done on a data set made entirely of these gauges would show essentially the same thing as PCA performed on random data.

Examining the eigenvalues, imbedded error, and/or the factor indicator function are the only way to check for random data. It is easy to mistakenly treat this very important step as an irrelevant intermediate stage, and prematurely move on to later portions of the analysis. If one continues on to calculate the rotated factors, they will appear to carry useful information, but one will actually be interpreting nonsense. Examining these factors will not lead one to the conclusion that the data is too random to analyze, but rather to unwarranted conclusions about the data set.

4.2.1.2 Identifying Non-Random Data

To ensure that at least some portion of our data was identifiable as having an underlying order, we took six specific gauges from Physics 151's traditional-style homework problems and performed PCA on them. The graphs below show the histogram (with best fit), imbedded error, factor indicator function, and eigenvalues for this analysis. There is a total of 1266 data points. The gauges chosen were: Start Time, Time Before Due, Problems with Full Credit, First Attempts, Late Attempts, and Late Problems. The first two, middle two, and last two gauges were known to be well-correlated with each other, and it was anticipated that PCA would show three underlying factors.

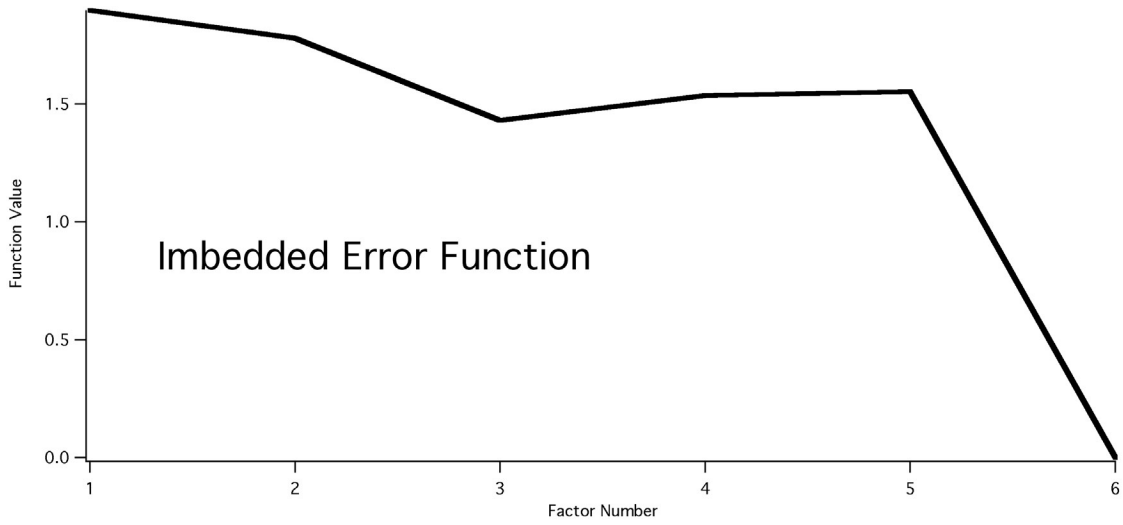
Fig. 4.3: Non-Random Data Histogram

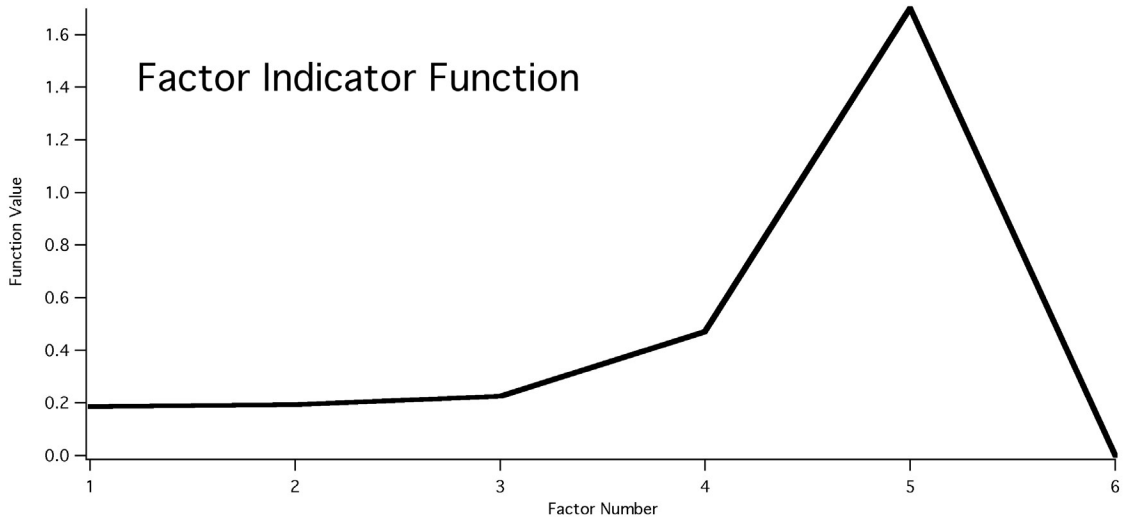


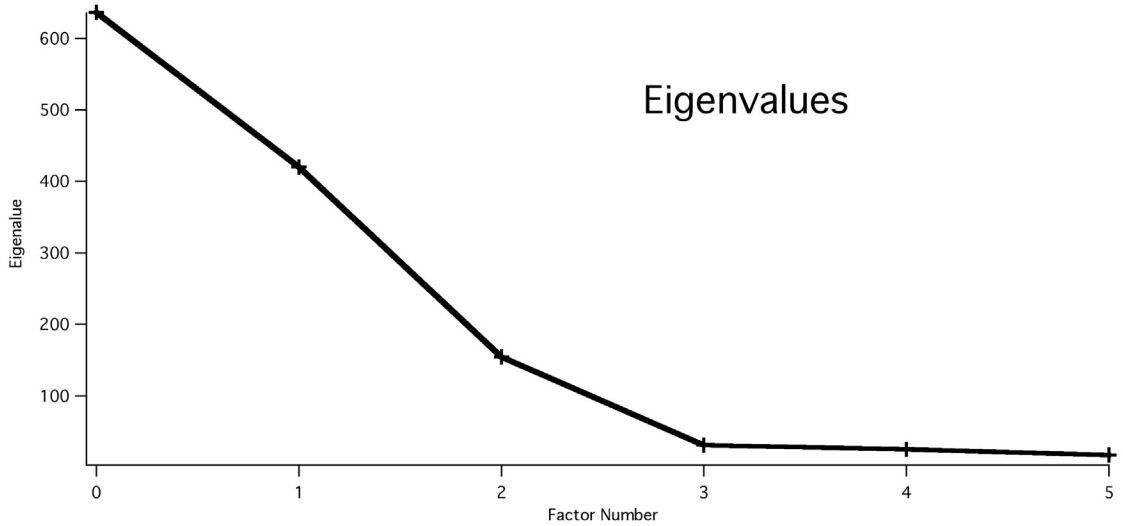
One can see that the gaussian fit to the histogram is not as good as in the random data set, which is primarily due to the increased likelihood of identical data points on the discrete measurements (First Attempts and Full-Credit Problems).

The Factor Indicator function still shows no minimum. However, the Imbedded Error

Fig. 4.4: Non-Random Data Plots







function shows a local minimum at three factors, and a scree test on the Eigenvalue graph also indicates three factors. It was not unusual in this study for us to rely on two out of three methods, or even a single method, when the remaining strategies proved unhelpful.

After rotation, the three factors did indeed show strong relationships between the groupings mentioned earlier. Factor 1 combines attempts and full-credit problems, factor 2 involves the time-related gauges, and factor 3 represents late homework. The table below shows these factors normalized to length 1:

Table 4.1: Non-Random PCA Results

Gauge	Factor 1	Factor 2	Factor 3
Start	0.020	0.708	0.001
TBD	0.053	0.690	-0.085
FC	0.673	-0.032	0.026
1st Att	0.728	0.111	-0.061
LAtt	0.083	-0.099	0.637
LProb	-0.082	-0.015	0.763

4.2.2 Correlation Coefficients

This section serves as a quick introduction to correlation coefficients for those who have not used them extensively before. All correlations used in this study are Pearson's "r" correlations, determined using the following formula, and typically calculated through Microsoft Excel:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\left(\sum (x - \bar{x})^2 \sum (y - \bar{y})^2\right)}}$$

Positive coefficients indicated a direct relationship, negative numbers an indirect relationship. The value of r is always between -1 and +1. The square of r shows how much of the variance of one measurement can be accounted for by the other measurement alone.

All correlation coefficients have a certain chance of being "false," that is, of indicating a relationship between two elements when it is only chance that makes them seem connected. One can measure this with a level of significance, denoted "p". The value of p is the chance that the correlation is *false*, so low p values are desirable. The higher the number of measurements taken and the higher the r value, the lower the chance of a false reading.

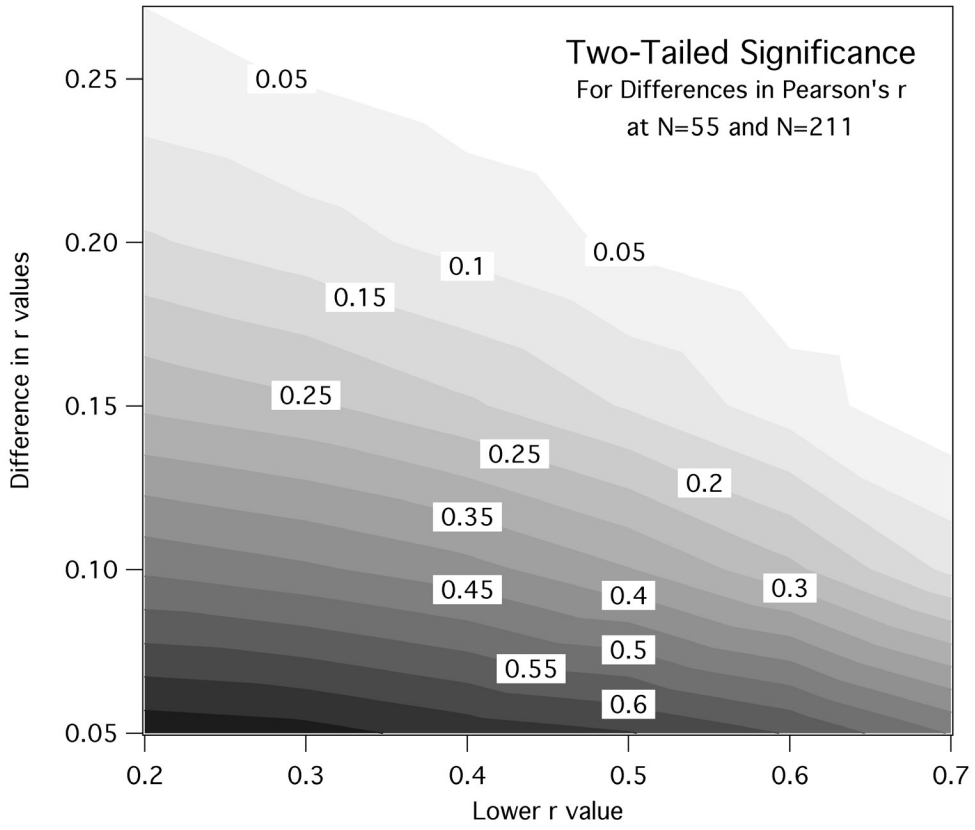
Minimum "r" values were assigned for correlation factor significance, using p = 0.01 and p = 0.001 as the cutoff values. It should be noted that because of the high number of

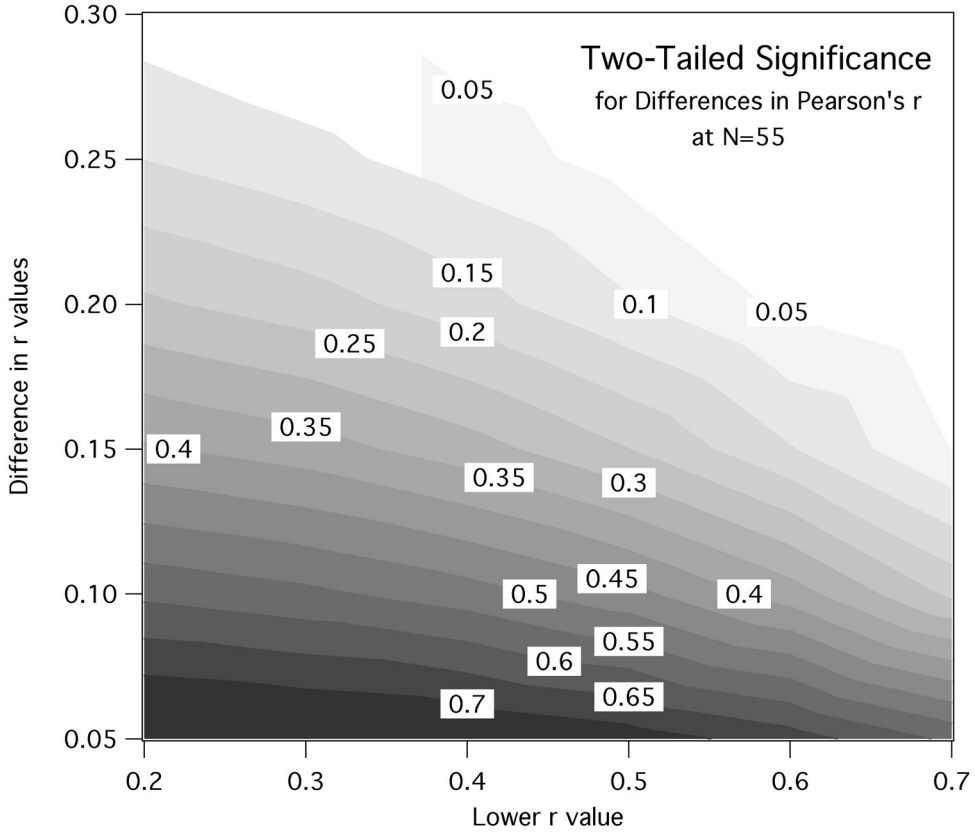
students involved in this study, small values of "r", sometimes as low as 0.181, could be considered statistically significant at the p = 0.01 level. To be consistent, we will use the word "significant" only in the statistical sense, and it should not be taken to mean that the effect size for a particular correlation is large. Correlations with a larger or smaller effect size will be referred to as "stronger" or "weaker," whether the effect is positive or negative. A "more positive" correlation is one where the result is closer to +1 than to -1.

One purely mathematical difficulty that arose in our study was the comparison of correlation factors from different sample sizes. We expected from the outset that Physics 151 would have a greater number of significant correlations simply because the threshold for significance is so much lower (r = 0.18 rather than r = 0.35, for the p = 0.01 level). A result of r = 0.30 would be a significant correlation for Physics 151, but not Physics 181, even though the level of correlation would be identical. For differing values of Pearson's r, the significance of the difference can be calculated using the Fisher r-to-z transformation. The function is anything but linear, and a difference of (for example) .2 between r values is much more significant for higher r values than for lower ones. The graphs on the next page show the significance of differences for N = 55 in both samples, or for N = 55 in one and N = 211 in the other. Physics 181 has 55 students, and Physics 151 has 211.

To ensure greater certainty in our results, we used the two-tailed values of p for all significance measurements. When a single correlation is larger or smaller than another, this test is reasonable and easy to interpret; when there are twenty correlation factors in (for example) Physics 181 that are marginally but uniformly higher than the same twenty in Physics 151, interpretation becomes somewhat more difficult. For more on comparisons of correlation factors between courses, see page 76.

Fig. 4.5: Correlation Coefficient Comparison





Cross-correlation tables were used from time to time to check the relationships between various measurements, especially between gauges and between behaviors. The primary disadvantage inherent in cross-correlation tables is that of cumulative unreliability. Pearson's "r" correlation can sometimes be "fooled" by data that appears correlated, but is actually random. The smaller the sample size, the larger this chance becomes for any particular correlation. While individual correlation factors are unlikely to be wrong, especially if they are very strong, the probability for error increases as more correlations are calculated. A cross-correlation table for the eighteen gauges in our study contains 171 distinct correlation factors. Even if each one is reliable at the $p = .01$ level, the chance that at least one of them is in error is 82.1%.

Principal Component Analysis does not suffer from this drawback, and can even become more accurate as a greater number of gauges are used. In addition, PCA allows the creation of reduced factors that are linear combinations of the original gauges, something that cannot be achieved with correlation tables. These two reasons alone are sufficient to recommend the use of PCA for data sets with many measurements.

Cross-correlation tables are still useful to assist PCA (or Factor Analysis, ANOVA, or other methods), as PCA should show a close relation between gauges with a high degree of correlation. If it does not, it may be time to examine one's methods and data preparation, in case an error has been made. In the end, however, cross-correlation tables cannot substitute for a more sophisticated method. PCA, despite all the different possible steps and choices, is a better procedure to use.

4.2.3 Attempts at Nonlinear Analysis

Our analysis of this data set made extensive use of the Pearson's r correlation coefficient, which can only effectively measure linear relationships. In an attempt to detect nonlinear effects that Pearson's r would not uncover, we employed two techniques: scatter plots and the Eta correlation ratio.

When examining scatter plots, we compared students' exam and final grades to their scores in several different gauges. We chose a range of gauges that included gauges well-correlated to performance (such as average score per attempt) and some completely uncorrelated gauges (such as elapsed time). We were unable to "eyeball" any significant nonlinearity in the graphs.

We also noted yet again that the data tended to be very "noisy." We were occasionally unable to discern the presence of small but statistically significant linear relations, even when we knew they were present. Because of this, we could not say for certain that there were no nonlinear trends in the data. This, along with an unwillingness to examine several hundred scatter plots by hand, led us to search for a more formulaic mathematical approach.

The Correlation Ratio, Eta, is a method of nonlinear correlation whose results bear some similarity to Pearson's r. Eta's value ranges from zero to one, with higher values indicating stronger correlation. Where its value is greater than that of Pearson's r, it indicates the presence of a nonlinear relationship. While the discovery of Eta was initially encouraging, two major drawbacks led us to discard it.

First, Eta requires multiple measurements at each X value. The more measurements each point has, the more reliable Eta will be. Single measurements, such as those found in our data set, cannot be compared, and reorganizing the data for analysis with Eta would require substantial effort.

Second, Eta is supposedly a ratio of variances: the variance of the average data divided by the variance of all the data as a whole. (We were actually unable to verify this assertion. The ratio of variances seemed a good estimate for Eta in some cases, but never yielded exactly the same result.) This is problematic, because pseudo-random data will be interpreted as having a strong relationship, and Eta will be near one. This was our largest complaint against Eta — that it detects not only nonlinear relationships, but occasionally nonexistent ones as well.

The large amount of work required to fit our data into Eta's mold, coupled with its unreliability, led us to reject it as a useful method of analysis in this study.

4.3 Constructing Higher-Level Data

This section discusses the methods we used to create more useful or meaningful higher-level data from the raw data that OWL provided. It starts with a brief discussion of the different levels of data and comparisons thereof that you will find in this thesis.

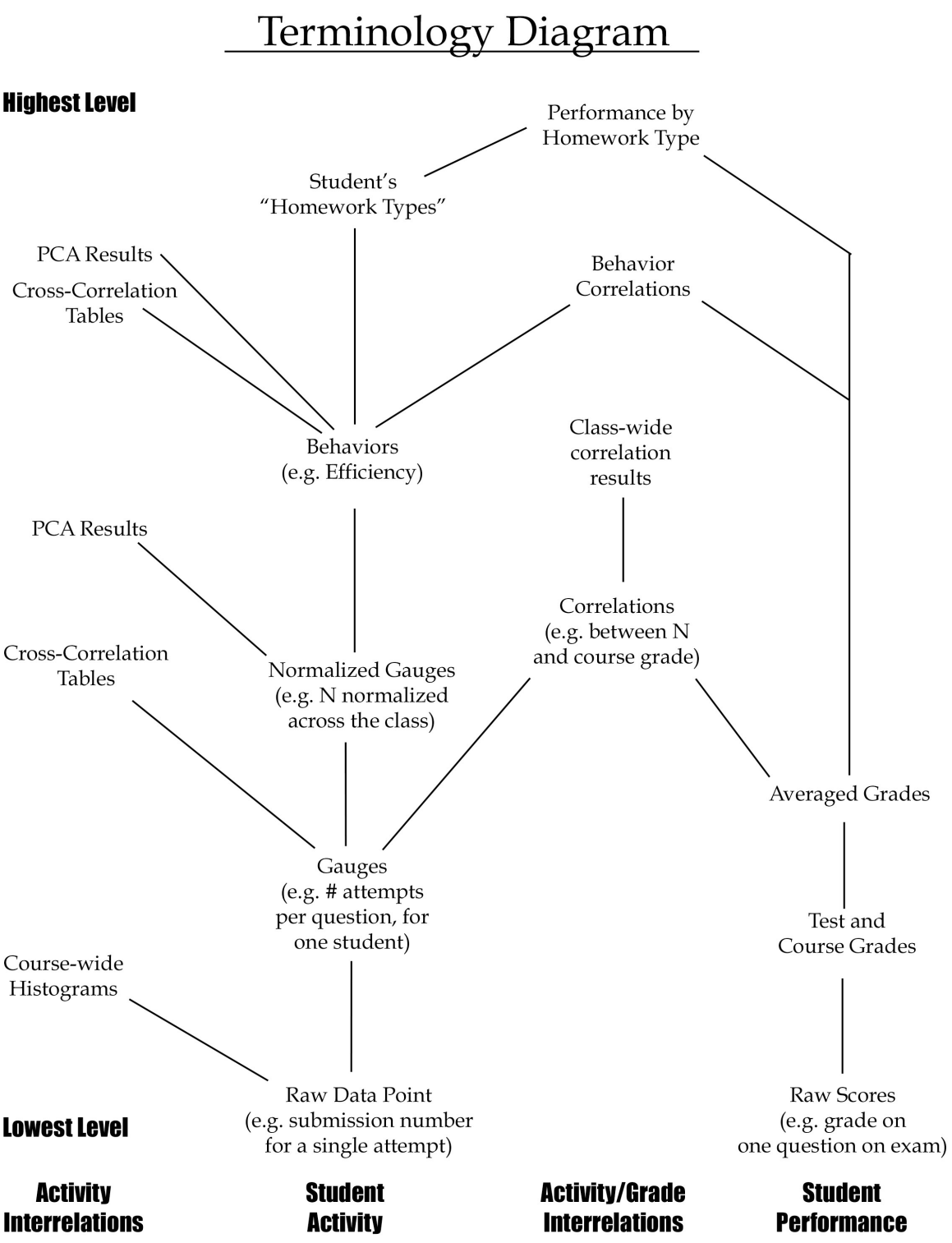
4.3.1 Terminology Diagram

The diagram shown on the following page is a conceptual layout of our data, from the most basic and "raw" level at the bottom to highly processed compound measurements at the top. The leftmost column shows relationships between different measurements of student homework activity. The activity measurements themselves are found in the column second from the left, while the column to the right of that shows relationships between homework activities and performance (as measured by various exam grades). The right-hand side of the diagram shows purely performance-related measurements.

At the bottom level of the diagram are the two sets of initial data available to this study: raw data taken from the OWL online homework system, and students' scores on exam and quiz questions. All of the information we have about students and their activity is derived from this data in one way or another.

Histograms can be constructed from the raw OWL data, both for individual students and on a course-wide basis. We can also create "gauges" that measure a single fact for a single student: for instance, how many homework problems a particular student finished with full credit, or how long it took a particular student to respond to homework ques-

Fig. 4.6: Terminology Diagram



tions on average. These gauges were broken down by problem type, so that (for instance) conceptual questions and multiple-choice questions were examined separately. Gauges are discussed in much greater detail below.

Exam questions were categorized by type of question (Analysis, Conceptual, etc.), and average scores were found for each question type on each exam for each student. These were used to construct the next level "up" on the chart; specifically, correlations between gauges and exam performance. These are the first connections we made between the two types of data Also found on this level are normalized gauges and cross-correlation tables (examining correlations between different gauges).

Of slightly greater complexity are the results from Principal Component Analysis. PCA in this study must be performed on normalized measurements if it is to be of any use, which necessarily puts it on a higher level of complexity. Also, while PCA shows information similar to that which can be found in the cross-correlation tables (one level down from here), it uses a more sophisticated and elaborate method to obtain it and avoids some of the drawbacks of correlation factors.

The next level contains student "behaviors," which are linear combinations of various gauges. The combinations were chosen "by hand" to fit a particular model of what effects a certain behavior or mental state would show. Behaviors are discussed at length in the next section. This level also holds class-wide correlation results, which are averages taken from the correlations on the previous level.

One level up from this we find cross-correlation tables between behaviors, and correlations between behaviors and performance. Again, slightly higher than this, we have PCA results, which were built from normalized behavior data (not shown).

The next two levels have just one item each. Student homework types are constructed from behaviors in various different manners. Using these and students' average grades we can look at the effects of students' homework types on their performance.

The first set of gauges and correlations were created in a test run using only a small portion of Physics 151. This run was successful in that the numbers we extracted from OWL were reasonable enough to pass a "reality check" — the number of problems attempted did not exceed the number of problems assigned, the time between starting and finishing a problem did not exceed the time between when it was assigned and the due date, and so forth. We believed that there was enough promise in this early test to continue with a course-wide analysis.

4.3.2 Behaviors and Gauges

In this section we discuss our methods for creating gauges and behaviors from lower-level data, as well as the reasons we chose these specific measures.

4.3.2.1 Gauges

Each gauge is the average or sum of a set of raw data points. Some gauges, such as the average number of attempts per problem and average credit per attempt, come directly from a single piece of raw data. Others, such as the number of late attempts and the average time before due, require comparison or calculation between several elements.

Because our gauges are named descriptively rather than evocatively, they do not involve any level of inference. It is important to remember this while examining them, as some of them seem to lend themselves to inference all too easily — guessing at the reasons why students might have a particular gauge rated high or low can be useful for hypothesis testing, but without the actual test we find that it is very easy to assume one sort of relationship between gauges and performance (or between multiple gauges) when a different one actually exists.

Here is what each gauge measures, our reason for including it in the set, and the abbreviation we used for it. There are also some comments on how some gauges were created where we believe it is not necessarily obvious.

- Seconds to Respond (STR): OWL reports this number for each attempt a student makes. Numbers above 1200 (which indicate times longer than 20 minutes) were stripped out, and the rest were averaged by attempt. This gauge (and this one alone) was calculated differently for Physics 151 and Physics 181; see pages 19-20 for details. Included to see whether time spent in front of the computer made a difference.
- Late Attempts (Latt): The total number of submissions a student makes with an answer date after the due date. Some few of these are only a couple minutes late, a result of accident rather than intent. Included to see whether students who worked on problems after they were due saw any benefit from it.
- Late Problems (Lprob): The total number of problems a student attempted for which the final attempt was submitted after the due date. Some of these problems were started before the due date, some after. In Physics 151, late homework was worth partial credit; the instructor for Physics 181 stated that late homework would be taken into account, but had no official policy regarding it. Included for the same reasons as Latt.
- Full Credit (FC): The total number of problems that a student completed with 100% credit. Included to compare one measure of homework performance with exam performance.
- Final Score (Fscore): A student's average final score on all homework problems. OWL records the score for each individual attempt; for this measure only those attempts that were the last attempt for a particular problem were taken into account. Included as another measure of homework performance.
- Average Score (AvgScore): The average score by attempt. Differs from the previous gauge in that the score was averaged for all attempts, not just final attempts. Included as yet another measurement of homework performance.

- First Attempts (1st): OWL records the "attempt number" for each submission, telling whether it is the first, second, etc. attempt at a particular problem. This measure is the total number of first attempts for a particular student, and thus gives us the number of problems said student attempted. Included to examine whether simply doing (or even just seeing) more problems was helpful.
- Number of Attempts per Problem (N): The final attempt number for each problem was determined, and the result was averaged for each student. Included to see whether taking many attempts was helpful or not.
- Short, wrong, first attempts (sw1): This counts the number of first attempts that were incorrect and had a Seconds to Respond of under 10 seconds. Included in an attempt to determine whether guessing could be measured, and what sort of effect it might have. We sometimes refer to this measure and the next one as the "guessing" gauges. This may also be indicative of students attempting to see the "feedback" provided by OWL or Homework Central before doing the problem.
- Short wrong attempts (sw): Total attempts (with any attempt number) that were incorrect and had STR under 10 seconds. This gauge is most prevalent in multiple-choice problems, whereas the previous gauge (sw1) has roughly equal prevalence among all problem types. We believe it indicates a guess-and-check approach to the problems. On non-MC questions, it may also occasionally be indicative of a student forgetting to put units on their answer, missing a negative sign, or forgetting to perform some very basic mathematical operation, and quickly "fixing" the mistake. Included for the same reason as sw1.
- Elapsed Time (Elapsed): The time from the beginning of the first attempt (attempt date & time minus STR) to the end of the last attempt (final attempt date & time). This measure was created because of a hypothesis that students' unconscious minds work on problems even when their conscious mind is focused elsewhere.

- Start Time (Start Time): The time from the beginning of the first attempt to the due date. Averaged by problem. Included to determine the effects on exam score from starting homework early.
- Time Before Due (TBD): The time from the beginning of each attempt to the due date. Averaged over all submissions. Very similar to Start Time (above). The difference is that a problem that is started early and finished early will have a large TBD, while one started early and finished late will have a lower TBD. However, both would have the same Start Time. Included for essentially the same reason as Start Time, and created because subtracting one from the other might show whether starting early has an effect even for students who did not begin serious work on the problem until later.
- Abandoned Questions (Abandon): The number of final attempts with less than 100% score. This differs from (first attempts minus full credit problems) in that there are some problems that just weren't attempted, and those don't count as abandoned. This behavior and the following three (QChange, Sessions, and Breaks) were included for essentially the same reasons: the hypothesis was that students who are stuck get more benefit from switching to a new problem than from struggling with the original one.
- Question Changes (QChange): Tallied when a student changes from one problem to another before their final attempt.
- Sessions (Sessions): Students who log into OWL start a "session." When they log out, log back in, or after a certain amount of time passes since their last submission, their old session ends. This gauge measures the total number of sessions a student used over the course of the term.
- Breaks Taken (Breaks): Tallied when a student changes session but does not change problems. Theoretically indicative of a student going away from the computer to do something else, hypothetically because the problem confused

them, made their head hurt, or just plain pissed them off. Also possibly due to the need to eat dinner, go home for the weekend, or leave the computer lab before it closes. Regardless, it's a break from the problem.

- Highest-Credit Attempt (highAtt): The attempt number that gave the student the greatest number of total points over the whole term. For instance, if a student typically received partial credit on their first attempts, full credit on their second attempts, and rarely used their third or higher attempts, this number would be 2 for that student. Included for the same reasons as Number of Attempts.

Later sections will discuss details on the correlations between these gauges and performance, correlations between different gauges, and various patterns that can be found in them. For right now, it should be noted that not all of the hypotheses mentioned above proved testable; they were simply the initial reasons for including these gauges in our data set.

Several preparatory steps were taken before examining correlations between students' gauges and their performance. In Physics 151, exams were broken down by problem type, and those that were Analysis or Traditional problems were summed for comparison (there were too few Multiple-Choice and Conceptual problems). In Physics 181 only the score by part was available, so those scores were used for comparisons instead (see page 20 for more details).

Each gauge was individually normalized through an affine transformation. The lowest rating in that gauge became a zero, and the highest a one. This was not necessary for the purposes of individual correlations (since correlation factors are invariant under affine transformations), but was critical for examining correlations with multiple gauges. For

example, if one wanted to make a combination of the "number of problems completed" gauge (typical order 100) and the "average final score" gauge (typical order 1), addition leaves the latter gauge utterly swamped by the former unless normalization is performed. With the gauges normalized, no gauge in a sum became inherently more "important" just because it had a larger scale.

For the majority of the Results & Findings chapter data was extracted from the courses as a whole, rather than any particular subgroups of them. For data on only engaged or only disengaged students, see the section starting on page 87.

In Physics 151, significant correlations had $|r| \geq 0.18$ for $p = 0.01$, or $|r| \geq 0.27$ for the $p = 0.0001$ level. In Physics 181, significant correlations had $|r| = 0.35$ or higher for $p = 0.01$, or $|r| \geq 0.44$ for the $p = 0.001$ level, since Physics 181 was much smaller than Physics 151.

4.3.2.2 Discarded Gauges

Some gauges were calculated, but were not used in later analyses for various reasons. These include measurements of short correct attempts, time-of-day measurements, and a "clumping factor." There was also a failed attempt to calculate a "group work" gauge.

As counterpoints to the "short incorrect" gauges used to check for guessing, a pair of "short, correctly answered" gauges were created, one measuring the number of attempts and one the number of questions. The abbreviations were SR and SR1, and they still appear on some graphs and tables in this thesis. As before, any attempt less than ten seconds long was considered "short." Only full-credit answers counted. The intent was to see whether evidence of cheating could be found, but it was decided (see below) that this method could not reliably reveal such activity.

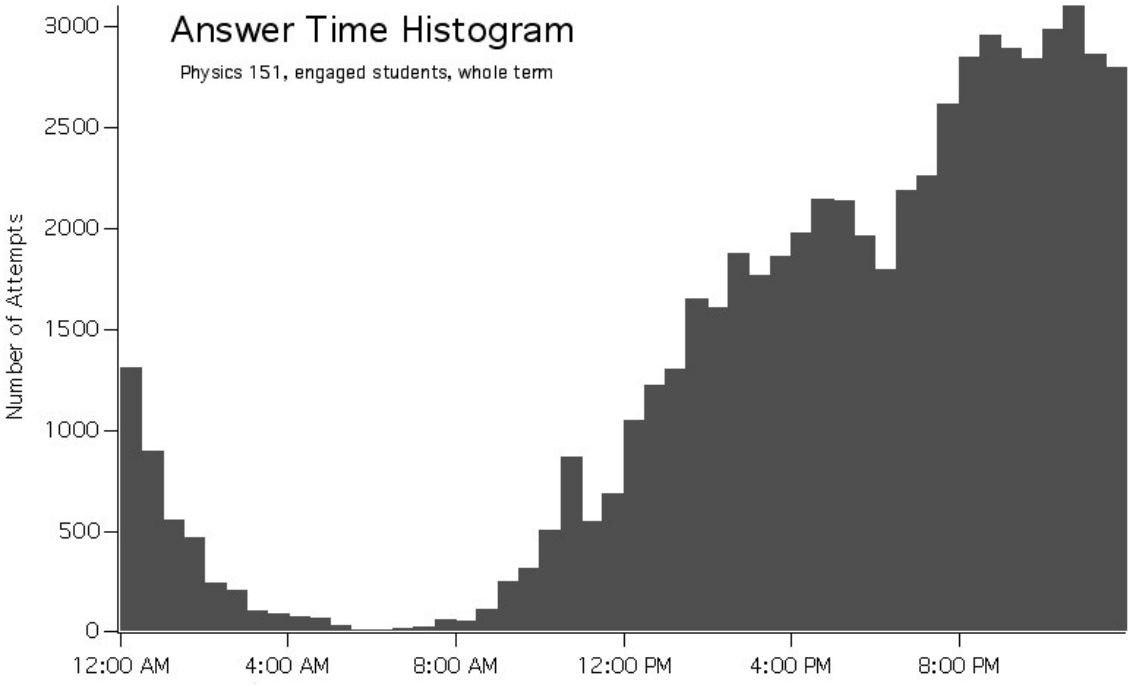
Three hypotheses were formed as to how a student with many short, correct answers could arrive at them. First, he or she could simply be a good or lucky guesser, with this gauge being highest for multiple-choice questions and practically nonexistent elsewhere. Second, he or she could be downloading the problem, printing it out, working it out on paper, and inputting the answer as soon as the page was opened when he or she returned for a second (or later) "attempt." Third, he or she could be using another student's help, either legitimately or in a dishonest fashion, to obtain answers. Since several of the possibilities seemed legitimate, and our data cannot disambiguate the latter two, these gauges were discarded as being comparatively information-free. If there had been more widespread correlation from either one, we believe that closer examination would have been in order.

Since there was little significant correlation to performance from this activity (and none whatsoever in Physics 181), it was discarded in later analysis. However, there was one relatively clear set of correlations to be found in Physics 151: short, correct answers to multiple-choice problems were negatively correlated (averaging $r = -0.34$) with performance on exams, thought not in the course as a whole. It may be that students who had many correct guesses on multiple-choice problems either leaned towards guessing on exams, or drew incorrect or weak conclusions from their answers and thereby hurt their exam performance. However, short, correct, *first attempts* did not have this negative correlation, but instead a positive one (averaging $r = 0.31$) between traditional problems and exam and course performance. We are unsure what this result signifies. Though the correlation is statistically significant, it is still low, and can likely be ignored for most purposes. It may warrant closer examination in the future.

Time-of-day measurements were created to check both when a student submitted responses (on average) and how reliable his or her daily homework time was (as measured by standard deviation). Both showed little to no correlation to performance. The only result that seemed worth reporting from these gauges was the histogram of student response times. Peaks can be seen before and after dinnertime, as well as a small but definite peak around 10:00 AM — just before classtime.

Though the gauges were not calculated for Physics 181, the histogram looked similar for that course, as well as for the disengaged students in Physics 151.

Fig. 4.7: Answer Time Histogram



The "clumping factor" gauge turned out to be difficult to create, low on significant correlations, and of dubious quality. While the intent was to see whether students did their homework all at once, or spread out across the term, we were unable to formulate a worthwhile method for determining this. The general noisiness of the data also reduced the reliability of such a measurement, as we suspect that "environmental" factors (e.g. other courses, exams, holidays) had greater impact than a student's desire to work in a particular manner. In the end, this gauge was simply too unreliable to keep. In any event, the "Sessions" gauge should record similar data: the greater the number of sessions used, the more spread out a students' homework attempts would be.

Finally, when we started this research, it was thought that it might be possible to gauge, to some degree, whether people were working in a group or alone. Unfortunately, this was stymied by two different problems.

First, the only location-oriented data available (which could indicate people using computers in the same room or apartment) was the IP address used in a particular session. After examining the sessions file for some time, using its start and end times to look for students on the same computer, or in the same lab, we were informed that the times and dates in the session file were unreliable, making that approach fruitless.

Second, through conversations with students, we discovered that many of them who did work in groups did so outside of a computer lab, or with only one or two computers available for six students. One or two people would try the solutions on OWL, and others might not log in until the next night.

Given these two factors, we concluded that any measure of group work that we might uncover would be suspect at best, and more likely completely invalid. We discarded all references to group work and concentrated only on individual habits.

For those interested in the effects of collaboration, Kotas (2002) has made a study of student collaboration and its outcomes, drawing on both survey data and online homework log file data. In short, his paper revealed that "students who regularly worked with others completed their assignments earlier and did well in class."

4.3.2.3 Behaviors

Behaviors are evocatively-named linear combination of gauges. There are seven different behaviors examined in this study: Tenacity, Efficiency, Inactivity, Uncertainty, Slow and Steady, Frustration, and Grade-Consciousness. When we talk about what these behaviors are, we mean our current understanding of what they reveal about the students. Much more so than gauges, they are open to interpretation.

- Tenacity is the desire to "get" a problem completely; at least for the purposes of credit, if not necessarily for the purposes of understanding (though we hope it indicates some desire for that as well).
- Efficiency is the tendency to make good use of each homework attempt. Efficient students use few attempts, don't waste them on guesses, and are at least partially successful in their efficiency.
- Inactivity combines starting homework late, abandoning questions partway through, guessing, and using as few attempts as possible. Inactive students may do well or poorly on their homework, but it seems clear that they do not put in much effort.
- Uncertainty is indicated by a large amount of time, and a large number of attempts, spent on each problem. It is also characterized by switching problems, guessing, and taking breaks.
- The Slow and Steady behavior characterizes those students who are not necessarily the brightest in the class, but who work steadily at the problems until they are successful. They take a long time, and many submissions, but do achieve full credit eventually.

- Frustrated students find themselves stumped often, and though they may succeed, the route they take is not smooth. They take many attempts, guess often, abandon questions, and switch between questions.
- Grade-Conscious students, obviously, care about their grade. They get full credit on many problems, work on problems after the due date for partial credit, and do every problem. They may or may not guess, be efficient, start early, switch problems, etc., but they work for the best credit they can get.

The process of creating a behavior was one of careful examination, comparison, and instinct. We often asked, "What sort of activity might we be able to see with this data?" and then attempted to create a suitable behavior by combining our gauges in various ways. Behaviors were not derived via any sort of multilinear modeling process or optimization procedure, and they are neither orthogonal to each other nor uncorrelated. It is not unreasonable to believe that significant overlap can exist between behaviors (e.g. students who are efficient may be less frustrated), or that not every behavior should be weighted so as to optimize correlations with performance.

There are some behaviors that we wanted to create, but could not. Some were unsuitable because they relied on information that our data set did not contain (a "likely cheater" behavior and a "collaborative" behavior were both discarded). Others were discarded because they turned out to be nothing more than a combination of the highest-correlated gauges, and showed little real meaning (a "good student" behavior was rejected for this reason).

On the next page are the behaviors in terms of the gauges measured earlier.

Table 4.2: Tenacity

Gauge	Weight
FC	+2
Fscore	-1
1st	+1
Abandoned	-1

Table 4.3: Efficiency

Gauge	Weight
N	-1
sw	-0.5
sw1	-0.5
AvgScore	+1

Table 4.4: Slow & Steady

Gauge	Weight
STR	+1
FC	+1
N	+1
Elapsed	+1
HighAtt	+1

Table 4.5: Grade-Consciousness

Gauge	Weight
Latt	+1
Lprob	+1
Fscore	+1
1st	+1

Table 4.6: Inactivity

Gauge	Weight
N	-1
sw	+1
sw1	+1
Start Time	-1
TBD	-1
Abandoned	+1

Table 4.7: Frustration

Gauge	Weight
N	+1
sw	+1
Abandoned	+1
Sessions	+1
Qchange	+1

Table 4.8: Uncertainty

Gauge	Weight
STR	+1
N	+1
sw	+1
Qchange	+1
Sessions	+1
Breaks	+1

As one can see, there was no “splitting hairs” in the creation of these behaviors. Gauge weightings were kept to integers or half-integers, and only those gauges which we believed to be conceptually related to the chosen behavior were included.

Not every linear combination of gauges creates a worthwhile behavior. To understand the meaning and import of a behavior, one must examine its correlations with performance and its interactions with other behaviors. Such interactions can be seen in cross-correlations between behaviors, in the correlations from multiple behaviors at once, and in the principal component analysis of groups of behaviors. We present some of this work in the Results and Findings chapter.

Behaviors were calculated for individual problem types, in much the same way as the gauges on which the behaviors are based. When overall measurements of behavior were used, they were weighted by problem type. This means that "overall" behaviors tend to be most similar to Traditional and Analysis behaviors, as there are far more of these types of problems than there are Conceptual or Multiple-Choice problems.

Further work would have to be done to check for causal relationships and nuances of meaning in behaviors. This work is beyond the scope of this thesis; we discuss it briefly in the Summary and Closing Remarks chapter.

CHAPTER 5

RESULTS AND FINDINGS

This chapter concentrates on the major conclusions that can be drawn from our work.

The most important finding we made was that it was indeed possible to characterize a course and its students solely through the use of online homework data. That analyzing such data could yield reliable predictions about student activity and its relationship to performance was by no means certain at the beginning of our work. In the course of examining the thousands of correlation coefficients between gauges, performance, and behaviors, a multitude of smaller facts were discovered, the majority of which will not be related here. Instead, we will concentrate on some of the more interesting and important kinds of discoveries that were made.

Many kinds of student activities made for excellent predictors in one way or another, and we start by examining those kinds of data. Next we look at evidence for validity to the division of homework problems into analysis, conceptual, multiple-choice / definition, and traditional categories. We also explored differences between different groups of students, such as comparisons between courses, and between engaged and disengaged students, and examined some mini-hypotheses with the data we obtained.

Finally, we were able to obtain some insight into our original research question, the effectiveness of practicing analysis questions. The final section of this chapter is devoted to that question.

5.1 Predictors and Links to Problem Types

Many of the activities measured in this study act as predictors of exam and final grade performance, with varying amounts of predictive power (correlation strengths). Using the gauges and behaviors below one can estimate what sort of scores a particular student is likely to have. In addition to their utility as exam and course grade predictors, some activities illustrate interesting points about the various different types of question and their relationship to student work on exams.

Below we have tabulated results from each gauge in each class. We use a notation of "+" for positive correlations and "-" for negative ones, with more symbols indicating a greater strength. Because the two courses have different numbers of students, the same correlation factor will have different levels of significance depending on which course it is measured in. This is noted in the table below. The first two "r" values were chosen because they are the $p = 0.01$ and $p = 0.0001$ cutoffs in Physics 151.

Table 5.1: Significance of Correlation Factors

Value	Correlation	Significance (151)	Significance (181)
+/-	$r \geq 0.18$	$p < 0.01$	$p < 0.19$
++/--	$r \geq 0.27$	$p < 0.0001$	$p < 0.046$
+++/---	$r \geq 0.4$	$p < 0.0001$	$p < 0.002$
++++/----	$r \geq 0.6$	$p < 0.0001$	$p < 0.0001$

Keep in mind that the values of p in Physics 181, for the first row, are so high as to make a single + or — sign very unreliable indeed.

Table 5.2: Physics 151, Gauges vs. Exams

Gauge	Analysis	Conceptual	MC/Def	Traditional
STR	++		+	++
Latt		-	-	-
Lprob				
FC	++	+	++	++
Fscore	++		+	+++
1st	+		++	++
N	-	--	---	-
sw1			-	
sw	-		--	-
elapsed				
start	+	+	++	+
abandon	-	-	-	--
qchange		-	-	
sessions				
break				
TBD	+	+	+	++
highAtt		n/a		
AvgScore	+++	+++	+++	+++

Here's an example of how to read this table. Looking in the "Fscore" row, in the "Analysis" column, shows "++". This means that having a high final score on analysis-type homework problems yields a positive correlation factor with performance on exams that averages $0.27 \le r \le 0.4$. In Physics 151, the chart on the previous page lets us know that this has a significance of $p \le 0.0001$.

The strongest predictors of exam performance are easy to see — average score per attempt (AvgScore), final score (Fscore), problems completed with full credit (FC), and average number of attempts (N) all have many moderate-to-high correlations. Other factors are weaker predictors, and some are negligible.

Some gauges show particularly interesting behavior, such as the fact that students who take many attempts on multiple-choice problems do worse than those who take many attempts on conceptual problems. Similarly, students whose final score per problem is high on traditional problems do very well, but when it comes to conceptual questions, there is no significant impact from this gauge.

Here's the same chart from Physics 181:

Table 5.3: Physics 181, Gauges vs. Exams

Gauge	Analysis	Conceptual	MC/Def	Traditional
STR	+++	++	++	+
Latt	-			
Lprob	-			
FC	+++	+++	++	+++
Fscore	+++	+++	+	+++
1st	++	++	+	++
N	--	-	-	--
sw1		-	-	
sw	-			
elapsed		++		
start	+++	+++	++	+++
abandon	---	---	-	---
qchange	-		-	
sessions		+		
break	-			
TBD	+++	+++	+++	+++
highAtt		--	-	
AvgScore	+++	++++	+++	+++

The importance of some homework types has shifted as: for example, conceptual problems now have just as many strong predictors as do traditional problems. However,

not every detail can be captured with any data reduction scheme such as this one. For instance, in Physics 181, the number of full-credit problems completed showed no relationship to certain exam problems, even though it correlated well with exams as a whole.

We can also see that Physics 181's correlations are typically stronger than those from Physics 151. A greater number of gauges show correlation, but remember that correlations with only a single + or — are much more likely to be false in Physics 181, and should probably be ignored.

Here are tables that show the impact on final course grade instead:

Table 5.4: Physics 151, Gauges vs. Course Grade

Gauge	Analysis	Conceptual	MC/Def	Traditional
STR	++		+	++
Latt		-	-	-
Lprob				
FC	+++	++	++	+++
Fscore	+++	+	+	+++
1st	++	+	+	++
N	-	--	---	-
sw1			--	
sw	-		--	-
elapsed				
start	++	++	+	++
abandon	-	-	-	--
qchange		-	-	
sessions				
break				
TBD	++	++	+	++
highAtt		n/a		
AvgScore	+++	++	+++	+++

Table 5.5: Physics 181, Gauges vs. Course Grade

Gauge	Analysis	Conceptual	MC/Def	Traditional
STR	+++	+	++	+
Latt				
Lprob	-			
FC	++++	++++	++	++++
Fscore	+++	+++	++	+++
1st	+++	+++	++	+++
N	-		-	--
sw1		-	-	
sw	-			
elapsed		++		
start	+++	+++	++	+++
abandon	---	---	-	---
qchange			-	
sessions		++		
break	-	+		
TBD	+++	+++	++	+++
highAtt		--		
AvgScore	+++	++++	+++	+++

Again, Physics 181's correlations are stronger overall.

Those gauges that acted as predictors for both courses seem to us to be somewhat more robust and reliable, and we believe that they could easily be used as predictors in other courses as well. Other gauges and even behaviors showed very little correlation to performance on their own. Those with weak or scattered correlations, on the other hand, should instead be reserved for other purposes. We do not suggest that ignoring any of these gauges is a useful thing to do — even if they have little impact on their own, many of them are useful in combination, for the construction of various behaviors. They may also be worth measuring for other reasons, such as assessing progress towards a particular

teacher's goals for the course. We merely state here that correlation factors alone give little to no indication that these activities are important.

Next we move on to the behaviors:

Table 5.6: Physics 151, Behaviors vs. Exams

Behavior	Analysis	Conceptual	MC/Def	Traditional
Inactivity	--	-	--	--
Uncertainty		--	--	
Tenacity	++	+	+	++
Efficiency	++	++	+++	++
Frustration	-	--	---	--
Grade-Conscious				
Slow and Steady				+

Table 5.7: Physics 181, Behaviors vs. Exams

Behavior	Analysis	Conceptual	MC/Def	Traditional
Inactivity	---	----	--	---
Uncertainty				
Tenacity	+++	+++	++	+++
Efficiency	+++	+++	++	+++
Frustration	---	--	-	---
Grade-Conscious				+
Slow and Steady	++	+++		++

Table 5.8: Physics 151, Behaviors vs. Course Grade

Behavior	Analysis	Conceptual	MC/Def	Traditional
Inactivity	---	---	---	---
Uncertainty	+	-	-	+
Tenacity	++++	+++	++++	++++
Efficiency	++	+	++	++
Frustration		--	--	
Grade-Conscious	++			+++
Slow and Steady	+++	+	+	+++

Table 5.9: Physics 181, Behaviors vs. Course Grade

Behavior	Analysis	Conceptual	MC/Def	Traditional
Inactivity	---	---	--	---
Uncertainty	+	++		+
Tenacity	++++	++++	+++	++++
Efficiency	++	+	++	++
Frustration				
Grade-Conscious	+++	+++	+	+++
Slow and Steady	+++	+++	+	+++

Tenacity and efficiency are some of the strongest positive predictors, and the combination of the two is particularly impressive when it comes to both course grade and exams (see page 104). Uncertainty and grade-consciousness have relatively weak effects on their own, but show definite effects when combined with other behaviors (more on this on page 103). Again, these tables don't show everything — for instance, there are differences in the correlations from Efficiency when it comes to timed vs. untimed exams in Physics 181 — but they do provide a reasonable amount of detail on how each behavior interacts with the various problem types.

5.1.1 Special Cases

Some activities are worth examining individually, either because of the complexity of their correlations or because they showed results that were very different from what we had hypothesized.

- Number of Sessions: In Physics 181 this gauge shows few significant correlations. In Physics 151 all of its correlations with exams are all negative; however, those with final grade are all positive. The more sessions a student has that end with a multiple-choice question, the worse their performance on exams. The more they have that end with an analysis or traditional problem, the better their final grade. It is interesting to note that multiple-choice questions typically come at the beginning of a homework assignment; students who end their session on one are either working backwards or returning to questions they missed or skipped before.

- Question Changes: All of this gauge's correlations are negative in Physics 151, which was surprising and somewhat disappointing. One of our hypotheses coming into this analysis was that switching questions when you get stuck was a beneficial thing to do. This might still be true — the effect here might be indicative of students who are less capable, who switch problems because they have less ability, rather than more capable students who need a break — but we cannot find support for it here.

- Number of First Attempts: In Physics 181, all of this gauge's correlations are positive. An excellent predictor of final exam performance, with $r = 0.74$ overall.

A mild predictor of average exam performance, typically having no significant correlations to any single exam question. The exception is the third question on the final exam, for which the overall correlation is r = 0.57 and all individual correlations are significant. This problem is of particular interest when it comes to our examination of analysis problems; see page 107.

- Start Time: In Physics 181, All of this gauge's correlations are positive. Most are in the r = 0.4 to 0.5 range. Multiple-choice correlations are typically weaker. Significant correlations come from the timed portion and average scores of the midterm exams (but not the untimed portion), the final question on the midterm exams, and the overall course grade. That none of the untimed exams showed more than one or two weak correlations was very unexpected; we had anticipated that this gauge might indicate a general tendency to start early, which might help students on their untimed exam performance. Instead, the timed portion of the exam shows far stronger correlations.

In a slightly different line, we attempted to look at the relationship between gauges and "exam improvement" from students' midterm and final scores in Physics 151, to see whether any gauges had particularly strong correlations. Unfortunately, effect sizes tended to be small to insignificant, and we did not pursue this line of inquiry in Physics 181. We would note that, in general, gauges that showed significant positive correlations to exam score improvement were those that showed negative correlations to performance. Abandoned questions and average score per attempt were good examples of this.

5.1.2 Comparisons Between Courses

The homework activities measured in our study are not only useful as predictors for student performance or for analyzing the impact of various different problem types; they can also be used to make comparisons between courses. While not every gauge or behavior is useful in this way, we have found several that helped to distinguish Physics 151 from Physics 181.

The majority of gauges show no significant differences between the two courses. The greatest differences between courses were significant to the following levels:

Table 5.10: Gauge Comparisons Between Courses

Gauge	p (two-tailed)	Stronger Overall
Seconds to Respond	0.46	Physics 181 (positive)
Late Attempts	0.28	Physics 151 (negative)
Late Problems	0.43	varies†
Full Credit Problems	0.047	see below
Avg. Score/Question	0.11	Physics 181 (positive)
Avg. Score/Attempt	0.48	varies†
First Attempts	0.39	Physics 181 (positive)
Number of Attempts	0.60	Physics 151 (negative)
Short Wrong 1st	0.54	Physics 181 (negative)
Short Wrong Attempts	0.52	Physics 151 (negative)
Elapsed Time	††	††
Start Time	0.10	see below
Abandoned Questions	0.089	see below
Question Changes	0.014	see below
Number of Sessions	0.32	Physics 181 (positive)
Breaks	0.67	varies†
Time before Due	0.11	see below
Highest-Scoring Att.	††	††

† Gauges marked "varies" have a stronger correlation with exams in one course, and a stronger correlation with exams in the other course. In these cases Physics 181's exam correlations are typically the stronger, while Physics 151 shows better correlations to course grade.

†† Gauges marked with a double asterisk show so few significant correlations that examining the differences between them seems counterproductive.

Certain gauges had more interesting or significant patterns of correlation. The first of these is the number of problems completed with full credit. The strongest correlations from this gauge were with traditional for 151, but with analysis for 181. The weakest in both cases came from multiple choice questions. Correlations are as follows:

Table 5.11: Correlations from Full Credit

Course	Course Grade	Exam Grades
Physics 151	0.70	0.29
Physics 181	0.74	0.54

Average Score per Question also had an interesting pattern. While the final course correlations were approximately equal at about 0.5, the average exam correlations were 0.504 for Physics 181, but 0.30 for Physics 151. The strongest correlations were with traditional questions for Physics 151, but analysis questions for Physics 181. The weakest correlations came from conceptual questions for Physics 151, but MC for Physics 181.

The two "short wrong" gauges were interesting in that their effects were reversed on the two classes: stronger in Physics 151 for one gauge, and stronger in Physics 181 for the other. The SW gauge on multiple-choice problems in particular sends test per-

formance plummeting in Physics 151. However, in Physics 181 these gauges show no significant correlations whatsoever in multiple choice. In both courses, the overall correlation factors are relatively weak, and often at the edge of significance, so it is difficult to say how important these differences are.

Start Time and Time Before Due show similar patterns: the course grade correlations are very similar, with Physics 151 slightly ahead, but Physics 181 comes out significantly ahead (with $p = 0.1$) in the exam correlations. An average of the two gauges is shown below.

Table 5.12: Correlations from Starting Early

Course	Course Grade	Exam Grades
Physics 151	0.50	0.24
Physics 181	0.44	0.45

The Abandoned Questions gauge shows an insignificant difference between Physics 151 and Physics 181 in overall course grade (-0.29 vs. -0.32), but a significant difference when it comes to exam grade (-0.25 vs. -0.47). The strongest correlations come from analysis problems for Physics 181, but traditional problems for Physics 151. The weakest come from conceptual questions for Physics 151, but multiple choice questions for Physics 181. Note that this matches up with the pattern discovered in the Average Score per Question gauge.

Finally, the Question Changes gauge showed that one course can show positive (if weak) correlations from a gauge, while another course can show negative correlations. On overall course grade, this gauge showed a -0.22 correlation in Physics 151, but a

+0.155 correlation in Physics 181. While the actual correlation itself is not significant in Physics 181, the disparity between them is large enough to make the difference significant.

A pattern that emerged while comparing the two courses was a connection between problem types. The pages above mention only the strongest and weakest correlations, but we also noticed that certain types of problem tended to "naturally" group together in each course. In Physics 151, if we judge by correlation factors, we would put analysis and traditional questions together, assuming that where one was high the other would likely be high as well. Likewise, multiple-choice and conceptual questions tended to group together. In Physics 181, on the other hand, traditional questions and multiple-choice questions tended to form one group, while analysis and conceptual questions tended to form the other.

In our opinion this says something important about the two courses. One possible reason is that Physics 151 is an engineer's course, oriented towards calculation and utility, while Physics 181 is intended for physicists and astronomers, and tends to be more conceptually oriented. A calculation-intensive course might lead more students to categorize problems on a "short vs. long" axis, in which case traditional and analysis questions do indeed belong together, both typically being longer than conceptual or MC questions. However, a course that explicitly focuses on conceptual understanding and the analysis of questions might give students a greater appreciation for the connection between physics concepts and the solution of problems. In that case, conceptual questions become the building blocks for analysis questions, and the two are naturally grouped together, with traditional and multiple-choice forming the other group.

5.2 Evidence for Problem Types

In our work we divided the problems in the homework, quizzes, and exams into four basic types: Analysis, Conceptual, Multiple Choice / Definition, and Traditional. However, simply because experts can detect and articulate the difference between certain problems doesn't mean that novices can. See, for instance, the work of Chi, Feltovich, and Glaser (1981). For our problem types to be meaningful in our analysis, there must also be differences in the way that students see and react to these problems.

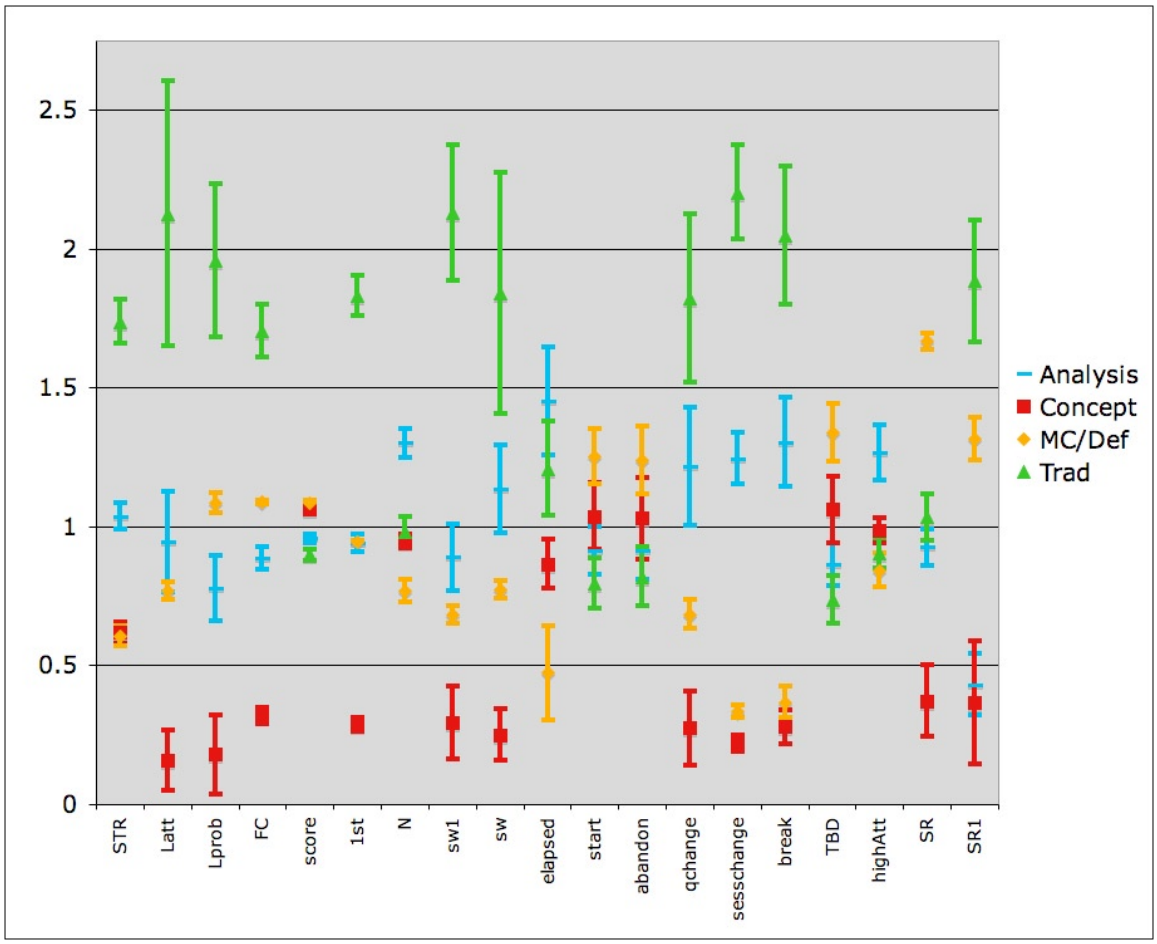
Evidence for problem type separation was discovered early on, on a student-by-student basis, in the first trial run. (The students themselves were also sometimes distinguishable by their gauges, showing differences between their gauge ratings beyond one standard error.) When the first major analysis of Physics 151 was undertaken, it seemed worthwhile to reexamine this idea. Data for the entire course was used to calculate averages and standard errors for each gauge and each problem type. Each gauge had a set of four separate averages and their standard errors, which have been plotted on the figure above. The bottom axis lists gauges. Different gauges have been rescaled differently to fit on this graph, with their standard errors rescaled properly with them — the vertical scale does not have the same meaning for all items.

As can be seen from the figure, almost all gauges show good evidence for problem type separation. The error bars are set at twice standard error. While there are some gauges incapable of distinguishing between problem types, other gauges can be used to disambiguate them with better than 95% accuracy. It seems possible to construct a com-

bination of gauges that could be used to determine a homework problem's type from how students react to it.

It is interesting to note that, while problem types are definitely a "reality" in this study, how students treat them makes little to no difference when it comes to their performance. Students who are very consistent in their behavior towards the various problem types see no benefit over students who treat different problem types in different ways.

Fig. 5.1: Problem Type Separation



5.3 Relationships Between Activities

Cross-correlation tables were created for both Physics 151 and Physics 181, showing the degree of correlation between different gauges.

The figure on the next page shows cross-correlation tables from the Analysis gauges, with Physics 181 on top and Physics 151 on bottom. Cells are color-coded to indicate degree of correlation: darker orange for $r > 0.85$, brighter orange for $0.85 \geq r > 0.65$, and pastel for $0.65 \geq r > 0.45$. Note that the tables are symmetric (as they should be) and that Physics 181 has a greater number of stronger correlations.

The gauges from traditional homework in Physics 181 were put into a form similar to a block diagonal, where off-diagonal elements are minimized as much as possible, and strong correlations are moved towards the central line. The second figure on the next page shows this. This arrangement helps one see connections between gauges, identifying some "clusters" of gauges with clearer similarities. Coloring on this graph is looser to allow for easier identification of blocks; the ranges are $r > 0.75$, $0.75 \geq r > 0.5$, and $0.5 \geq r > 0.25$. Though it may seem to have stronger correlations overall, in actuality, the cross-correlation tables for traditional and analysis work share a similar number of strong correlations. Their block-diagonal forms are also similar.

There are three major well-correlated groups in the block-diagonal representation: one composed of time measures (Elapsed Time, Start Time, and TBD), one composed primarily of performance measures (Full Credit, Final Score, Average Score, and Abandoned Questions), and one related to the amount of work done (First Attempts, Breaks

Fig. 5.2: Gauge Cross-Correlations

	STR	Latt	Lprob	FC	score	1st	N	sw1	sw	elapsed	start	abandon	qchange	sessions	break	TBD	highAtt	AvgScore
STR	1.00	-0.33	-0.39	0.26	0.34	0.06	-0.43	-0.23	-0.27	-0.16	0.09	-0.34	-0.33	-0.23	-0.17	0.18	-0.25	0.44
Latt	-0.33	1.00	0.84	0.11	0.03	0.18	0.75	0.44	0.46	-0.05	-0.18	0.04	0.56	0.48	0.58	-0.17	0.16	-0.33
Lprob	-0.39	0.84	1.00	0.02	-0.07	0.19	0.64	0.46	0.59	-0.06	-0.17	0.19	0.67	0.58	0.56	-0.17	0.15	-0.28
FC	0.26	0.11	0.02	1.00	0.77	0.76	-0.09	0.18	0.09	-0.06	0.28	-0.66	0.19	0.34	0.20	0.27	0.06	0.45
score	0.34	0.03	-0.07	0.77	1.00	0.27	-0.17	0.12	0.01	-0.04	0.21	-0.88	-0.13	0.02	0.05	0.19	0.11	0.70
1st	0.06	0.18	0.19	0.76	0.27	1.00	0.04	0.19	0.11	-0.03	0.18	-0.02	0.55	0.51	0.23	0.20	-0.02	0.08
N	-0.43	0.75	0.64	-0.09	-0.17	0.04	1.00	0.40	0.56	0.09	-0.20	0.18	0.64	0.51	0.57	-0.24	0.48	-0.66
sw1	-0.23	0.44	0.46	0.18	0.12	0.19	0.40	1.00	0.53	0.17	-0.09	-0.07	0.54	0.50	0.42	-0.10	0.48	-0.17
sw	-0.27	0.46	0.59	0.09	0.01	0.11	0.56	0.53	1.00	0.07	-0.10	0.00	0.50	0.66	0.72	-0.13	0.29	-0.33
elapsed	-0.16	-0.05	-0.06	-0.06	-0.04	-0.03	0.09	0.17	0.07	1.00	0.50	0.05	0.21	0.23	0.13	0.42	0.25	-0.18
start	0.09	-0.18	-0.17	0.28	0.21	0.18	-0.20	-0.09	-0.10	0.50	1.00	-0.23	0.00	0.17	-0.08	0.95	-0.13	0.12
abandon	-0.34	0.04	0.19	-0.66	-0.88	-0.02	0.18	-0.07	0.00	0.05	-0.23	1.00	0.34	0.06	-0.04	-0.19	-0.11	-0.60
qchange	-0.33	0.56	0.67	0.19	-0.13	0.55	0.64	0.54	0.50	0.21	0.00	0.34	1.00	0.74	0.54	-0.01	0.35	-0.45
sessions	-0.23	0.48	0.58	0.34	0.02	0.51	0.51	0.50	0.66	0.23	0.17	0.06	0.74	1.00	0.71	0.17	0.24	-0.33
break	-0.17	0.58	0.56	0.20	0.05	0.23	0.57	0.42	0.72	0.13	-0.08	-0.04	0.54	0.71	1.00	-0.09	0.23	-0.37
TBD	0.18	-0.17	-0.17	0.27	0.19	0.20	-0.24	-0.10	-0.13	0.42	0.95	-0.19	-0.01	0.17	-0.09	1.00	-0.18	0.16
highAtt	-0.25	0.16	0.15	0.06	0.11	-0.02	0.48	0.48	0.29	0.25	-0.13	-0.11	0.35	0.24	0.23	-0.18	1.00	-0.24
AvgScore	0.44	-0.33	-0.28	0.45	0.70	0.08	-0.66	-0.17	-0.33	-0.18	0.12	-0.60	-0.45	-0.33	-0.37	0.16	-0.24	1.00

	STR	Latt	Lprob	FC	score	1st	N	sw1	sw	elapsed	start	abandon	qchange	sessions	break	TBD	highAtt	AvgScore
STR	1.00	-0.39	-0.34	0.31	0.40	0.17	-0.37	-0.37	-0.52	-0.03	0.18	0.03	-0.13	-0.02	0.17	0.18	-0.14	0.56
Latt	-0.39	1.00	0.85	-0.03	-0.22	0.15	0.63	0.36	0.73	0.02	-0.27	0.08	0.37	0.28	0.01	-0.21	0.12	-0.42
Lprob	-0.34	0.85	1.00	-0.08	-0.27	0.14	0.45	0.31	0.50	-0.02	-0.35	0.17	0.38	0.28	-0.02	-0.28	0.09	-0.31
FC	0.31	-0.03	-0.08	1.00	0.76	0.86	0.01	0.29	0.07	0.28	0.60	-0.06	-0.07	0.42	0.44	0.48	0.01	0.65
score	0.40	-0.22	-0.27	0.76	1.00	0.42	-0.16	-0.05	-0.23	0.15	0.44	-0.04	-0.13	0.13	0.33	0.41	-0.04	0.74
1st	0.17	0.15	0.14	0.86	0.42	1.00	0.15	0.45	0.28	0.27	0.49	-0.05	0.04	0.56	0.39	0.34	0.06	0.41
N	-0.37	0.63	0.45	0.01	-0.16	0.15	1.00	0.29	0.80	0.41	0.03	-0.04	0.36	0.50	0.20	0.00	0.40	-0.59
sw1	-0.37	0.36	0.31	0.29	-0.05	0.45	0.29	1.00	0.57	0.17	0.17	-0.04	0.04	0.31	0.11	0.06	0.09	-0.14
sw	-0.52	0.73	0.50	0.07	-0.23	0.28	0.80	0.57	1.00	0.26	0.03	-0.06	0.23	0.43	0.08	-0.01	0.26	-0.54
elapsed	-0.03	0.02	-0.02	0.28	0.15	0.27	0.41	0.17	0.26	1.00	0.46	-0.05	0.15	0.63	0.34	0.45	0.12	-0.06
start	0.18	-0.27	-0.35	0.60	0.44	0.49	0.03	0.17	0.03	0.46	1.00	-0.07	-0.23	0.34	0.31	0.88	0.08	0.35
abandon	0.03	0.08	0.17	-0.06	-0.04	-0.05	-0.04	-0.04	-0.06	-0.05	-0.07	1.00	-0.22	-0.05	-0.05	0.02	0.01	-0.01
qchange	-0.13	0.37	0.38	-0.07	-0.13	0.04	0.36	0.04	0.23	0.15	-0.23	-0.22	1.00	0.16	0.04	-0.10	-0.07	-0.14
sessions	-0.02	0.28	0.28	0.42	0.13	0.56	0.50	0.31	0.43	0.63	0.34	-0.05	0.16	1.00	0.61	0.24	0.14	-0.06
break	0.17	0.01	-0.02	0.44	0.33	0.39	0.20	0.11	0.08	0.34	0.31	-0.05	0.04	0.61	1.00	0.29	0.10	0.17
TBD	0.18	-0.21	-0.28	0.48	0.41	0.34	0.00	0.06	-0.01	0.45	0.88	0.02	-0.10	0.24	0.29	1.00	0.03	0.36
highAtt	-0.14	0.12	0.09	0.01	-0.04	0.06	0.40	0.09	0.26	0.12	0.08	0.01	-0.07	0.14	0.10	0.03	1.00	-0.25
AvgScore	0.56	-0.42	-0.31	0.65	0.74	0.41	-0.59	-0.14	-0.54	-0.06	0.35	-0.01	-0.14	-0.06	0.17	0.36	-0.25	1.00

Above: Cross-correlations between gauges from analysis homework problems in Physics 181 (top) and Physics 151 (bottom)

Fig. 5.3: Gauge Cross-Correlations, Block-Diagonal

	STR	N	Latt	Lprob	qchar	sw1	sw	highA	sessi	break	1st	FC	Fscor	aban	AvgSc	elaps	start	TBD
STR	1	-0.4	-0.4	-0.4	-0.3	-0.5	-0.5	-0.3	-0.1	-0	0.1	0.14	0.21	-0.1	0.3	0.03	0.07	0.08
N	-0.4	1	0.77	0.68	0.7	0.71	0.68	0.7	0.51	0.39	0.37	0.35	0.14	-0.1	-0.4	0.1	-0	-0
Latt	-0.4	0.77	1	0.94	0.71	0.74	0.71	0.61	0.46	0.21	0.29	0.33	0.17	-0.1	-0.2	-0	-0.2	-0.2
Lprob	-0.4	0.68	0.94	1	0.73	0.64	0.67	0.54	0.49	0.25	0.31	0.31	0.14	-0.1	-0.2	0	-0.2	-0.2
qchange	-0.3	0.7	0.71	0.73	1	0.53	0.54	0.39	0.68	0.36	0.68	0.47	0.03	0.16	-0.3	0.16	-0	-0
sw1	-0.5	0.71	0.74	0.64	0.53	1	0.94	0.6	0.34	0.11	0.17	0.15	0.04	-0	-0.4	0	-0.1	-0.2
sw	-0.5	0.68	0.71	0.67	0.54	0.94	1	0.55	0.4	0.16	0.18	0.21	0.11	-0.1	-0.3	0.08	-0.1	-0.1
highAtt	-0.3	0.7	0.61	0.54	0.39	0.6	0.55	1	0.2	0.15	0.08	0.1	0.04	-0.1	-0.4	-0.1	-0.2	-0.2
sessions	-0.1	0.51	0.46	0.49	0.68	0.34	0.4	0.2	1	0.82	0.65	0.61	0.27	-0.1	0.05	0.28	0.27	0.22
break	-0	0.39	0.21	0.25	0.36	0.11	0.16	0.15	0.82	1	0.43	0.46	0.25	-0.2	0.05	0.31	0.33	0.27
1st	0.1	0.37	0.29	0.31	0.68	0.17	0.18	0.08	0.65	0.43	1	0.82	0.27	0.01	0.03	0.11	0.2	0.19
FC	0.14	0.35	0.33	0.31	0.47	0.15	0.21	0.1	0.61	0.46	0.82	1	0.74	-0.6	0.41	0.21	0.28	0.23
Fscore	0.21	0.14	0.17	0.14	0.03	0.04	0.11	0.04	0.27	0.25	0.27	0.74	1	-0.9	0.72	0.25	0.25	0.2
abandon	-0.1	-0.1	-0.1	-0.1	0.16	-0	-0.1	-0.1	-0.1	-0.2	0.01	-0.6	-0.9	1	-0.7	-0.2	-0.2	-0.1
AvgScore	0.3	-0.4	-0.2	-0.2	-0.3	-0.4	-0.3	-0.4	0.05	0.05	0.03	0.41	0.72	-0.7	1	0.1	0.22	0.21
elapsed	0.03	0.1	-0	0	0.16	0	0.08	-0.1	0.28	0.31	0.11	0.21	0.25	-0.2	0.1	1	0.69	0.67
start	0.07	-0	-0.2	-0.2	-0	-0.1	-0.1	-0.2	0.27	0.33	0.2	0.28	0.25	-0.2	0.22	0.69	1	0.98
TBD	0.08	-0	-0.2	-0.2	-0	-0.2	-0.1	-0.2	0.22	0.27	0.19	0.23	0.2	-0.1	0.21	0.67	0.98	1

Above: Cross-correlation of behaviors on traditional homework in Physics 151, arranged in a form similar to a block diagonal.

Taken, and Sessions). The latter two tend to “bleed” into each other; it is difficult to see which block Full Credit “really” belongs to.

Seconds to Respond, rather than being included in the time-related group, has few correlations with them and sits on its own in the top left. The rest of the gauges (N, Latt, Lprob, qchange, sw, and sw1) are all somewhat intercorrelated, and difficult to disentangle. They form the large block in the upper left.

While occasionally useful (and pretty), this table proved somewhat disappointing. Most of the connections revealed by it were common-sense functional sorts of relationships, rather than any sort of unexpected underlying structure. Even the fact that the Abandoned Questions gauge was in with the other performance measures is easily understood — a large number of questions abandoned means a large number with low credit; thus the strong negative correlation with Final Score. It is the way the gauges were calculated that connects them, not any deeper truth.

The largest block of factors seem to be related only in that they are all “bad” activities, with primarily negative correlations and/or negative impacts when combined with other correlations. It seems that if a student has one of these habits, he or she is likely to have (or develop) the others as well.

We were also able to construct cross-correlation tables for behaviors in both courses. On the next page is the cross-correlation table for the behaviors in Physics 151 and Physics 181 combined. The total number of students is 266, so a correlation of $r = 0.2$ is significant to the $p = 0.001$ level. Shading indicates breakpoints of $0.4 < |r| \leq 0.6$ (cyan),

$0.6 < |r| \leq 0.8$ (light blue), and $|r| > 0.8$ (medium blue). Cross-correlation tables for the two courses individually look relatively similar. The behaviors on this table were created from combinations of overall gauges, rather than problem-specific gauges, so "inactivity" here refers to that behavior overall, rather than on a particular type of problem.

Fig. 5.4: Behavior Cross-Correlations

	Inactive	Uncertain	Tenacious	Efficient	Frustrated	Grade-Conscio	Slow & Steady
Inactive	1.00	0.13	-0.48	-0.41	0.41	0.14	-0.36
Uncertain	0.13	1.00	0.30	-0.59	0.87	0.64	0.70
Tenacious	-0.48	0.30	1.00	0.09	-0.01	0.55	0.68
Efficient	-0.41	-0.59	0.09	1.00	-0.78	-0.40	-0.28
Frustrated	0.41	0.87	-0.01	-0.78	1.00	0.57	0.43
Grade-Conscious	0.14	0.64	0.55	-0.40	0.57	1.00	0.57
Slow & Steady	-0.36	0.70	0.68	-0.28	0.43	0.57	1.00

The strongest connections seem to be between frustration and uncertainty ($r = 0.87$) and frustration and efficiency ($r = -0.78$). Other major connections involve the slow and steady behavior connecting to uncertainty ($r = 0.70$) and tenacity ($r = 0.68$), and uncertain students being grade-conscious ($r = 0.64$) but not efficient ($r = -0.59$). Non-correlations can also be informative, such as that between inactivity and grade-consciousness, which might indicate that it is possible to care about one's grade, but not work particularly hard to improve it.

The table was also useful in making sense of the behaviors, and checking to see whether our names for them were sensible. For example, if cross-correlations indicated that inactive students tended to be highly tenacious it would be time to carefully reexamine our names for those behaviors.

One thing that makes this table interesting is that not all of its connections (or lack thereof) were immediately obvious in the way that most gauge cross-correlations were. The connection between frustration and inefficiency, for example, suggests a causal

relationship in one direction or the other. Does frustration lead to inefficient behavior? Is inefficient behavior a cause of frustration? Is there a feedback loop between the two? The relationship makes sense when examined, but speaks to an underlying order. Such connections may be worth examining in later studies.

5.4 Findings Related to Engagement

Engagement is an overused word in educational research. If it is to have any meaning in this thesis, it must be carefully defined, lest a reader familiar with one of the dozens of other definitions of engagement mistake our meaning for one of theirs.

In this study, students are often spoken of as being "engaged" or "disengaged." Physics 151 was initially split this way for purely technological reasons: the class as a whole provides of over 140,000 rows of data, and Excel (our primary data analysis program) handles less than half of that. Dividing the students into three groups, one "engaged" and two "disengaged," made it possible to analyze the data.

The decision as to whether or not students were engaged was based on the many different activities their class required. Physics 151 in 2003 included homework assignments, lecture preparation assignments, PRS problems, course feedback surveys, quizzes, exams, and pre/post test pairs. The last item turned out to be entirely unreliable (even otherwise good or engaged students often skipped these), so they were not part of the decision. Those students we counted as engaged were those who attempted 85% of the other items, and attended all four exams. Physics 181 used the same cutoff point on its assignments: online homework, written homework, and PRS problems

The score that students received on these activities was considered unimportant – what mattered was that they attempted the assignment.

The measure of engagement we use is a purely functional or operational one. Many other definitions of engagement have been discussed in the literature; an overview can be seen in Fredericks (no relation), Blumenfield, and Paris (2004). Many sources have used the word "engagement" to indicate an emotional attachment to one's work or to school in general, and such engagement has been shown to improve school performance and satisfaction. We use engagement to mean only that a student has done — or at least attempted — the majority of the work that the class requires. While this may imply the existence of emotional or cognitive engagements, it by no means requires them.

Table 5.13: Physics 151 Grade Distribution by Engagement

Grade	Total	Engaged	Disengaged
A	37	25	12
AB	39	23	16
B	58	29	29
BC	48	22	26
C	30	5	25
CD	11	1	10
D	14	0	14
F	6	0	6

There were no D, F, or Incomplete grades in the engaged group. Twelve students with unmatched IDs were removed from this group, as well as sixteen students with grades of "Incomplete." After removing them, the group was still too large to fit into a single Excel worksheet, and it was further trimmed by removing students without even a single pre/post test pair. The disengaged students (with N=126) consisted of everyone else.

Physics 181 was small enough to be analyzed as a whole, but was split into engaged and disengaged groups for the purposes of comparison with Physics 151. There were 55 students overall, 22 of which were counted as engaged and 33 of which were not. All students were kept, as there were no incompletes and no mismatched student IDs.

Table 5.14: Physics 181 Grade Distribution by Engagement

Grade	Total	Engaged	Disengaged
A	9	7	2
A-	5	5	0
B+	8	4	4
B	5	1	4
B-	10	4	6
C+	4	1	3
C	4	0	4
C-	3	0	3
D+	2	0	2
F	5	0	5

As can be seen from both classes, engagement is no guarantee of success, nor does disengagement guarantee failure. However, engagement certainly helped.

On the next pages are correlations gathered from engaged and disengaged students in Physics 151 and 181. They are displayed in the same "+/-" notation used earlier, and with the same r-value cutoffs given in the table on page 67.

Table 5.15: Correlations by Engagement in Physics 151

	Engaged		Disengaged	
Gauge	Course	Exams	Course	Exams
STR	++	+	++	++
Latt	-			-
Lprob	-			
FC	+++	++	++++	++
Fscore	+++	++	+++	++
1st	++		++++	+
N	---	---		--
sw1	-	-		--
sw	---	--		--
elapsed		-		-
start	+++	++	++	
abandon	---	--	--	--
qchange	--	--		
sessions	-	--	+	
break		-	+	
TBD	+++	++	++	
highAtt	-	-		-
AvgScore	++++	+++	++++	+++

One pattern that emerges in these differences is that disengaged students often show greater correlation. Whether this is a causal relationship, or simply a natural feature of a group with a greater range of responses, is an open question. Since Physics 181 does not show this feature, and in fact reverses it on exam scores, we suspect there may be some deeper reason for this pattern.

Table 5.16: Correlations by Engagement in Physics 181

	Engaged		Disengaged	
Gauge	Course	Exams	Course	Exams
STR			++	+++
Latt				
Lprob				
FC	++++	++++	++++	+++
Fscore	++++	++++	+++	+++
1st	++	++	++++	++
N	-	-		--
sw1				
sw				
elapsed	++	+		
start	+++	+++	+	++
abandon	----	----	-	---
qchange	---	--	++	-
sessions			++	
break			+	
TBD	+++	+++	+	++
highAtt				--
AvgScore	+++	++	+++	++++

On this table we can see some changes from Physics 151, as well as some stark differences in some cases between engaged and disengaged students — look at the effects of seconds to respond, abandoned questions, and changing questions often. These behaviors are significantly better (or at least “less worse”) for disengaged students. Meanwhile starting early becomes even more important for the engaged students. This implies that it may be worth treating the two categories differently when it comes to advice on homework.

Here are a few interesting items that do not appear on these tables due to their low “resolution”:

- Getting full credit on many problems helps engaged students more earlier in the term, but disengaged students more later on.

- Question #3 on the final exam is helped by this for the disengaged, but not the engaged. See page 107 for more on this interesting question.

- In Physics 151, engaged students show strong negative correlations to final grade from abandoning traditional and analysis questions, while disengaged students see weak negative correlations from MC and conceptual questions instead.

- Average Number of Attempts: In Physics 181, this gauge shows more negative correlations for the disengaged than for the engaged. In Physics 151, the reverse is true. Correlations from multiple choice questions are particularly strong.

- Time Before Due: In Physics 151, this is a moderate positive predictor for exams and course grade (though not the third exam) for engaged students. Disengaged students see almost no effect from this, only a weak positive on the course grade. For Physics 181, the effects were often at the threshold of significance, but in general the engaged students saw a greater benefit from this on course grade, and the disengaged saw it more on exam grades. Certain problems showed correlation with one group and not the other; the most impressive difference was in exam #1's question #5, a rotation/tension problem. Engaged students saw an average correlation above $r = 0.7$, while disengaged students saw an insignificant correlation around $r = 0.3$.

5.4.1 Physics 151 Survey Data

Much of the Physics 151 survey data was unsuitable for comparison with student activity. However, some items proved interesting to examine.

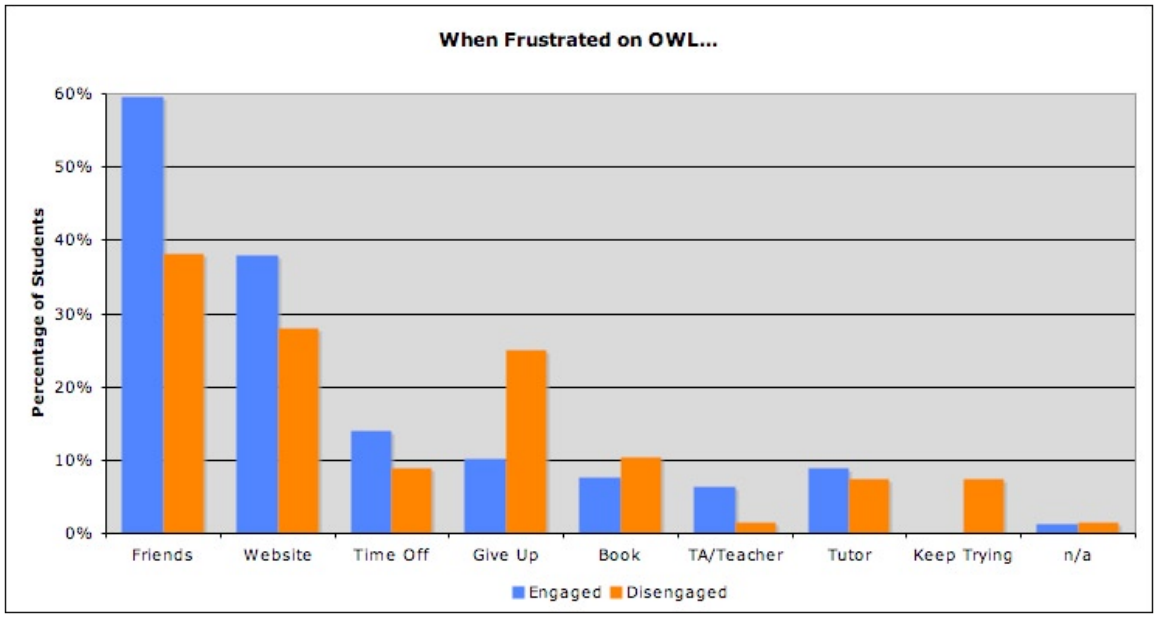
The final survey in the term included the question, "When working on electronic homework, what did you do when you were stuck, frustrated, or had a question?" This was a free-response question, allowing students to type basically as much as they wanted in response. All in all, 68 disengaged and 79 engaged students responded to this survey. The responses were separated into engaged and disengaged students, and then binned into the following categories based on the resources they used:

Table 5.17: Resources Used When Stuck On Homework, Physics 151

Resource Used	Engaged	Disengaged
Friends	59.5%	38.2%
Website	38.0%	27.9%
Time Off	13.9%	8.8%
Give Up	10.1%	25.0%
Book	7.6%	10.3%
TA/Teacher	6.3%	1.5%
Tutor	8.9%	7.4%
Keep Trying	0.0%	7.4%
n/a	1.2%	1.5%
Overage	45.6%	27.9%

Most resources are self-explanatory. The "n/a" row is for students who wrote "NA" or "I didn't get stuck." The "overage" row is included because some students used multiple resources — the sum of the other items in the column is not 100%. Blank responses were not counted.

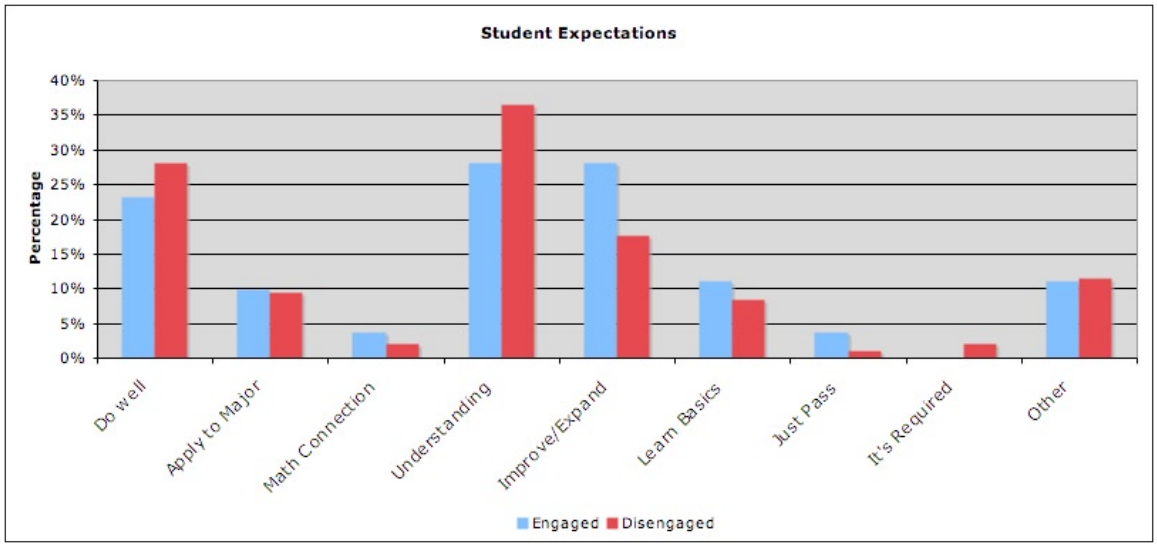
Fig. 5.5: When Frustrated on OWL



As can be seen in the graph, disengaged students were more likely to give up when confronted with frustration on OWL (this was statistically significant). They were also much more likely to simply "keep trying," rather than seek outside assistance.

Engaged students, on the other hand, were much more likely to seek multiple resources when in trouble. Only about a quarter of disengaged students did this, while nearly half of the engaged students did. This might be the result of a greater level of commitment to the course, or a better understanding of what it takes to succeed. Alternatively, some engaged students may have become more frustrated with the resources that were most easily at hand, and sought out others, while disengaged students were more content with what they had.

Fig. 5.6: Student Expectations



Also interesting is a comparison between engaged and disengaged students on their expectations for the course. The question was phrased, "What are your personal expectations for Physics 151 this semester?" This was in the first survey given. 82 engaged and 96 disengaged students responded.

Many students seemed to interpret the question as asking, "Why are you taking this class?" instead, which helps to explain some of the categories below.

Students who said they wanted a grade of B or higher were placed in the "Do well" category, while those who aspired to a C were placed in "Pass." The "Improve/Expand" category is for those responses that indicated some existing physics knowledge and a desire to increase it. The "Basics" category is for students who explicitly mentioned "learning the basics of (physics, mechanics, etc)." As before, the Overage category exists because some students gave more than one response to the question.

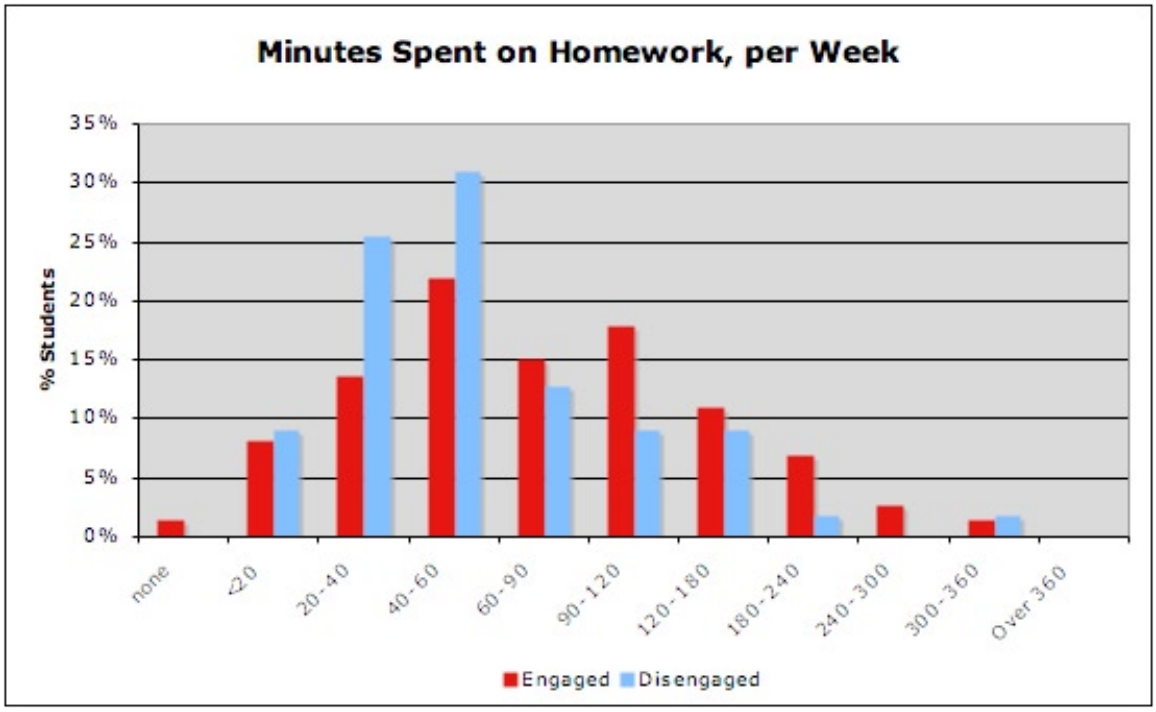
Table 5.18: Personal Expectations in Physics 151

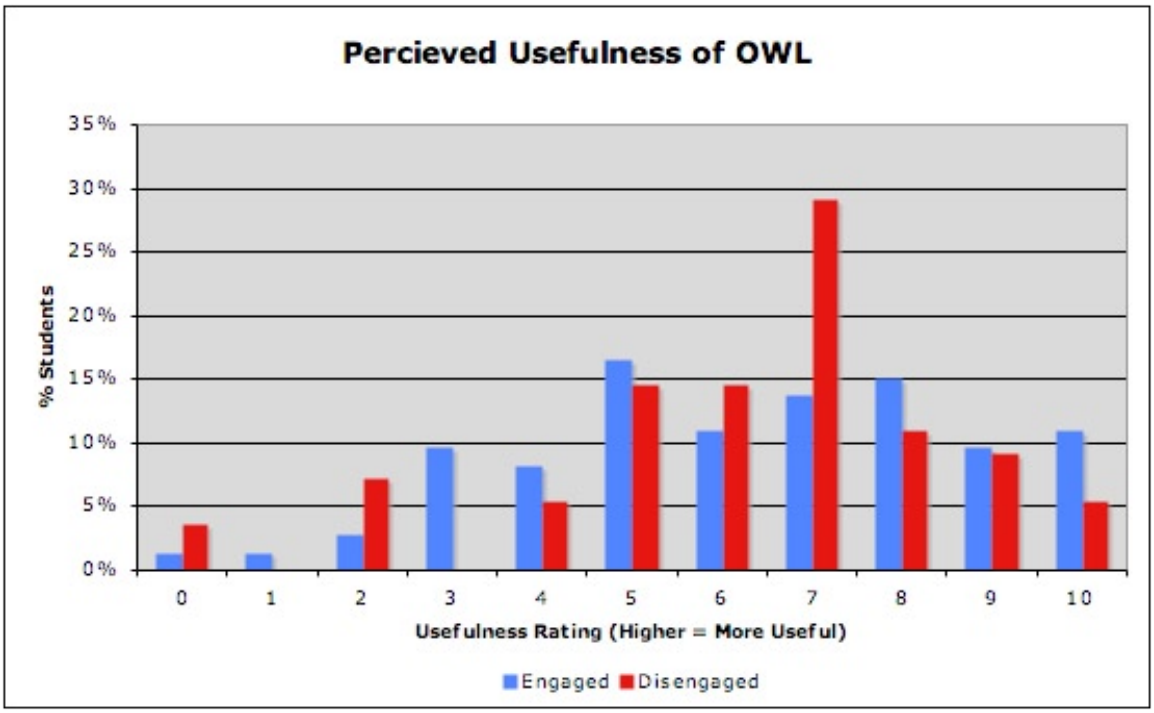
Response	Engaged	Disengaged
Do well	23.2%	28.1%
Apply to Major	9.8%	9.4%
Math Connection	3.7%	2.1%
Understanding	28.0%	36.5%
Improve/Expand	28.0%	17.7%
Learn Basics	11.0%	8.3%
Just Pass	3.7%	1.0%
It's Required	0.0	2.1%
Other	11.0%	11.5%
Overage	18.3%	16.7%

The most interesting difference we see in these numbers is that disengaged students more often indicated a desire to understand the material, while engaged students were more often interested in improving on existing knowledge. These two results are on the edge of statistical significance, so a deeper investigation of this dichotomy might be very revealing for student motivations.

Further comparison between engaged and disengaged students can be found in the third survey, which related directly to OWL. Students were asked how much time they spent per week on OWL ("On average, how much time have you spent on each Electronic Homework assignment? Please include any time spent reading or otherwise preparing to complete the assignment."), and how useful they found it ("Please rate the usefulness of the Electronic Homework assignments for bringing out your sources of confusion and resolving them.") The former was free-response, but easy to treat numerically. The latter was a 1-10 scale, with 10 being very useful. 73 engaged students and 55 disengaged students responded to this survey.

Fig. 5.7: Survey Data Comparison A





As one might expect, engaged students reported spending more time on homework. The two groups found OWL to be about equally useful, though with noticeably different distributions. Graphs are on the next page.

It is interesting to note that students' reports of the time they spent generally does not correlate well with their actual time spent as measured by various gauges. The table below shows correlation factors between reported time and gauges:

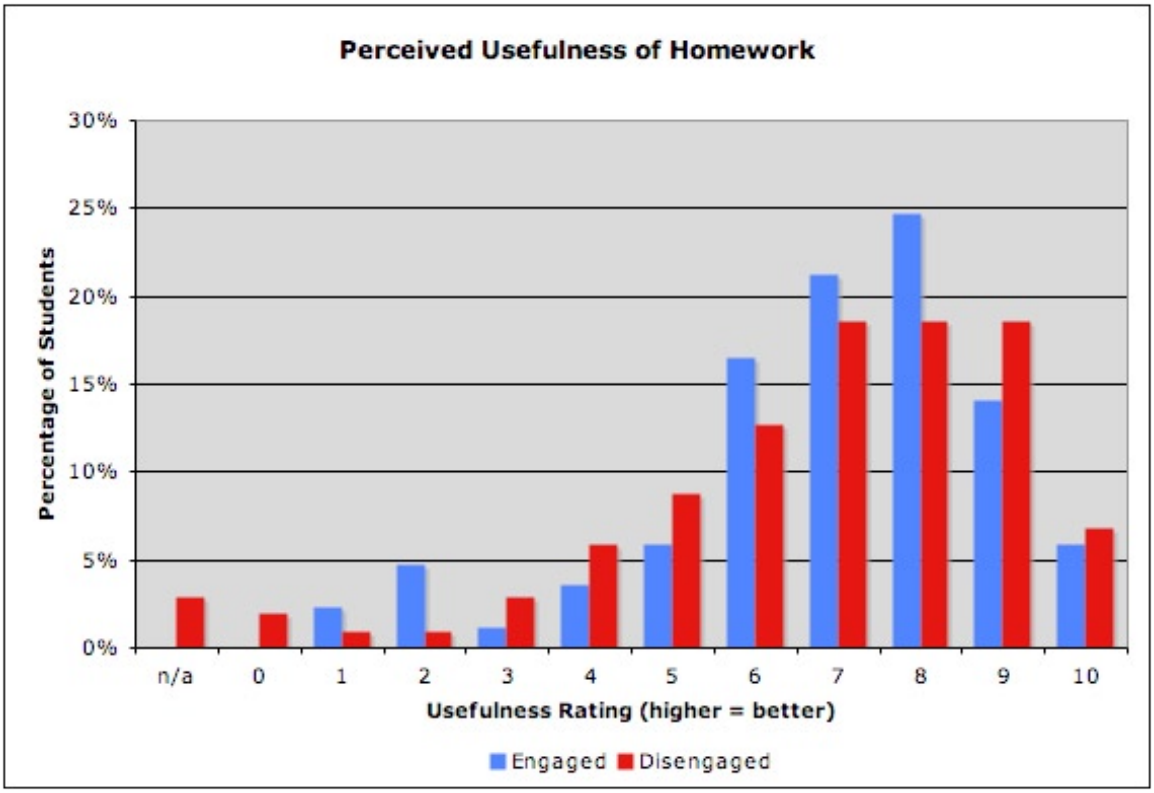
Table 5.19: Reported Time vs. Various Gauges, Physics 181

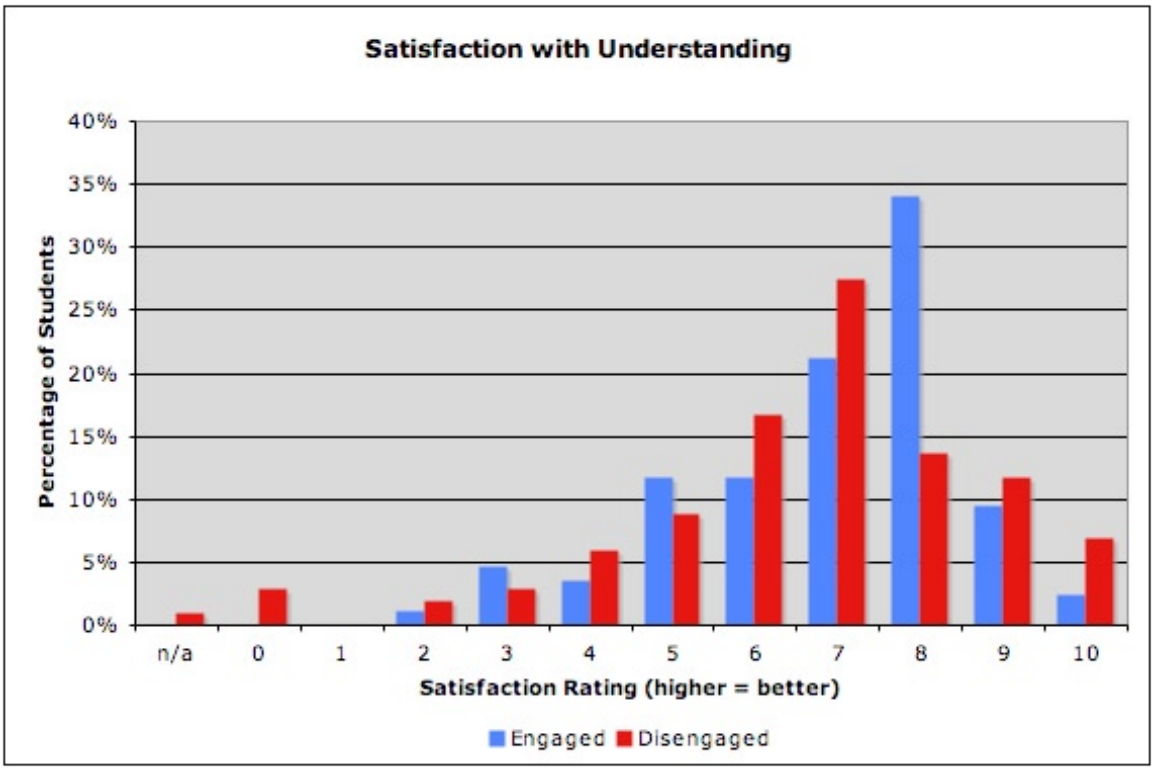
Gauge in question	Engaged	Disengaged
Elapsed Time	0.31	0.17
Time Before Due	-0.04	-0.18
Seconds to Respond	0.14	0.38

Only two factors (top left and bottom right on the table) were strong enough to be significant in this table, and they are still relatively weak. It is interesting to note that "time spent on homework" is closer to Elapsed Time for engaged students, but Seconds to Respond for the disengaged, leading us to speculate that engaged students do more work while not in front of the computer screen. This is particularly interesting because of the phrasing of the question in the survey ("Please include any time spent reading or otherwise preparing to complete the assignment.")

Finally, survey #10 asked students to "rate the weekly homework assignments in terms of helping you to understand course material and prepare for course exams" and, later on, "On a scale of 0 (not satisfied at all) to 10 (extremely satisfied), rate your under-

Fig. 5.8: Survey Data Comparison B





standing of the course material." Note the wording: satisfaction with understanding was the key, not perceived improvement or level of understanding. 85 engaged students and 102 disengaged students answered this survey.

There was no statistical difference between the average answers for either question. Distributions are shown on the previous page. One can compare the "perceived usefulness" on that page to a similar question (page 97) for some insight into how students' opinions changed over the course of seven weeks.

5.5 Other Hypotheses

While most of the study was spent figuring out how to quantize student activity and behavior, a small set of other hypotheses were created and tested during the course of this work.

The first hypothesis was that the correlations found in the course of investigating Physics 181's gauges would be very similar (in pattern if not in exact numbers) to those found in the engaged portion of Physics 151. This was hypothesized because Physics 181 is intended for physics majors and is advertised as such; it seemed likely that it would have more dedicated students. As it turned out, this was basically incorrect. The difference between Physics 181 and Physics 151 students is not just that Physics 181 students are equivalent to the highly-engaged subset of Physics 151 students. The difference runs deeper than this, and it seems likely that correlations are dependent on the instructor, the methods used to teach the course, and the general disposition and mental makeup of students who choose these courses.

Our second hypothesis was that the difference in behaviors between engaged and disengaged students in Physics 181 will be similar to the difference found in Physics 151 students. While we expected that Physics 181 students might all be considered "engaged" if they took Physics 151, we believed that the relative differences between engaged and disengaged students will be the same. When tested, this hypothesis turned out to be true in certain areas. In both courses, engaged students are less inactive, more tenacious, slower and steadier workers. Uncertainty is equally prevalent in both sides. However, there are differences as well: Engaged students are more tenacious in Physics 181, but not

Physics 151. Disengaged students are more frustrated and more grade-conscious in Physics 151, but not Physics 181. Again, we suspect the nature of the course and the students it attracts to be the root of these differences.

Our third hypothesis was that Physics 181 would, in general, have stronger (farther from zero) correlations than Physics 151. The course was smaller, and was a more controlled environment with a smaller number of student types. It seemed that spread in extracurricular influences might also be smaller, which would lead to more "certain" relationships and less noise in the data. In fact, this seems to be borne out not only by the correlation factors, but also by PCA results. Correlations show stronger relationships, but not necessarily a greater number of significant relationships (because the smaller class size pushes up the minimum significant value of "r"). In PCA analysis, Physics 151 showed some problem types that had no definite number of factors — that is, the methods of determining the number of factors showed differing or inconclusive results. Physics 151 also had no more than two factors in any problem type. However, Physics 181 showed at least two definite factors in all types, and four factors present in traditional problems. This is indicative of greater underlying order, lending credibility to this hypothesis.

5.5.1 On Combinations of Behaviors

When examining correlations from behaviors to performance, there were several occasions when our predictions of what a behavior's effects did not match the results we observed. Luckily, many behaviors become easier to understand when they are examined in combination with each other. Making and testing the hypotheses below proved useful in understanding the behaviors in greater depth.

As an example, it was initially predicted that uncertainty would have a moderate significant negative effect, because of the well-known positive effects of confidence on performance. When the effect was less negative than anticipated, it became fruitful to examine combinations of uncertainty with other things. For instance, inactivity in combination with uncertainty is not as bad as inactivity alone, perhaps indicating that uncertainty drives people to seek help while those who are inactive and "sure" that they can't do the problem will just give up.

When uncertainty and tenacity were combined, however, correlation coefficients plummeted. In these cases, rather than encouraging students who were tenacious but stuck to seek help, uncertainty may have pushed them to seek credit in unproductive ways. Uncertainty in general seems to "tone down" other behaviors, making them either less effective or less counterproductive.

Certain behavior combinations seem like they wouldn't make sense; for instance, inactivity and tenacity seem antithetical. The correlations from this particular combina-

tion are weak to nonexistent, which supports the idea that the two are, to some extent, conceptual opposites.

When combining inactivity and efficiency, it was initially assumed that correlations would be positive, as students who wanted to make the most of their attempts and get some free time afterwards might be even more effective with their time. This turned out to be incorrect: inactivity was the more powerful force by far, and correlations were entirely negative ($r \sim -0.35$ overall with course grade).

After examining many possible combinations, we found that the most "effective" combination by far was that of tenacity and efficiency. When added together, these behavior combinations provide outstanding positive correlations between student activity and exam scores. No combination or single behavior provided stronger correlations, either positive or negative. Pearson's r values are as follows:

5.20: Correlations from Tenacity + Efficiency

Component	Physics 181	Physics 151
Exam Average	0.71	0.49
Course Grade	0.80	0.83

This combination of behaviors provides correlations stronger than any discovered through principal component analysis. They are comparable to values obtained through much more intensive and complex optimization methods (see, for example, the thesis by Minaei-Bidgoli, 2004). Some possible uses of this will be discussed in more detail in the Classroom and Program Evaluation section, on page 125.

A short summary follows of other combinations that we examined:

- Grade-consciousness worsens the effects of tenacity, especially on multiple-choice questions.
- Uncertainty worsens the effects of being slow and steady, turning some correlations negative (especially on MC and conceptual questions).
- Slow and Steady combined with a small amount of "certainty" (negative 1/2 times Uncertainty) provided good correlations overall: $r \sim 0.58$ with Physics 181, 0.3 with Physics 151 exams, and 0.65 with the Physics 151 course grades.
- Inactivity and Frustration together provide many negative correlations, which are worse for exams than for the final course grade (presumably because of the time limitation).

5.6 Findings Related to Analysis

The original purpose of this thesis was to determine whether work on analysis-style homework problems has an impact on the ability to analyze problems later on. Our hypothesis was that exposure to and work on analysis-style homework would noticeably improve students' analysis skills. Without a control group we cannot show causality, but we can show a great deal of correlation.

5.6.1 The Effects of Behaviors and Gauges on Analysis

Many problems in physics can be solved in multiple ways. Not only are there many physical laws one can apply, but there are different cognitive approaches as well: analysis, guess-and-check, problem type memorization, and so forth. To show that analysis problems are effective, we would want to have a source that we can point to as being definitively analytical, one which utterly requires analysis.

We see this most clearly in the Physics 181 class, in the third section on the final exam (hereafter referred to as FE3). This section asks students which of the major concepts covered in the course can be used to solve a problem most efficiently. There are five parts to this section, each one giving a choice of Newton's 2nd Law, linear momentum, angular momentum, the Work-Energy Theorem (or conservation of energy), or both momentum and energy. Justifications are required as well.

While there are other problems on the exam that have components of analysis, this problem is the only one that can be solved solely through analysis or experience: there is

simply not enough time to try each attempt and see which is the most efficient. If there is a single case study that will reveal the importance (or lack thereof) of analysis problems, FE3 is it.

5.6.2 Behavioral Correlations with FE3

By examining the correlations each of the seven behaviors have with FE3, we can tell something about their effect on students' abilities to analyze problems. Below we will use these correlations to build an argument that struggling with the problems in a fruitful way seems to be the most useful thing for students to do when it comes to analysis.

The table below summarizes the correlations with this question. Remember that in Physics 181 (the course from which this question comes), correlations of $r = 0.354$ are significant to the $p = 0.01$ level. Significant correlations are marked in boldface.

Table 5.21: Correlations with FE3

	Inact.	Uncert.	Tenacity	Eff.	Frust.	G-C	S&S
Overall	-0.21	0.28	**0.56**	0.02	0.08	**0.43**	**0.52**
Analysis	-0.22	0.18	**0.54**	0.11	0.00	**0.39**	**0.44**
Conceptual	-0.27	**0.37**	**0.60**	0.04	0.10	**0.47**	**0.54**
MC	-0.10	0.14	**0.43**	-0.05	0.05	0.30	0.30
Traditional	-0.12	0.22	**0.37**	-0.01	0.14	**0.35**	**0.39**

Both inactivity and efficiency have no significant correlations with FE3. Neither of these activities involves a great amount of work; in fact, both of them minimize (albeit in very different ways) the amount of time students spend on problems.

Frustration likewise has no significant correlations. Students who spend more time frustrated on a problem are no doubt struggling with it, but they are doing so in an ineffective manner. The guess-and-check method that accompanies and signifies frustration seems unlikely to aid students in developing understanding. Not all ways of struggling with a problem are useful.

Uncertainty on conceptual questions has a moderate good correlation with FE3, though uncertainty on other problem types yields no significant correlations. Comparing uncertainty to other behaviors has shown that it can sometimes push students to find answers. As conceptual problems form a sort of "toolkit" for analysis questions, it makes sense that being uncertain on them can help those students who seek answers to better understand concepts, and thus be better prepared for analysis.

Being grade-conscious (that is, seeking a good score, trying every problem, and working problems after the due date) on analysis and conceptual questions provides a positive correlation with FE3. We know from an examination of individual gauges (see below) that simple exposure to problems is beneficial, so this is no surprise. However, grade-conscious students are not always interested in uncovering meaning and understanding, so we can guess as to why grade-consciousness in multiple choice and traditional questions do not help when it comes to FE3: students may work only for the grade, and garner little of the understanding that would help them effectively analyze problems.

The two behaviors with the strongest FE3 correlations are tenacity and "slow and steady." Both show positive correlations, stronger from analysis and conceptual questions than from MC and traditional questions. "Slow and steady" explicitly includes time-relat-

ed gauges, while tenacity refers to a student's unwillingness to abandon questions. Both are indications of hard workers who receive full credit, either immediately or after time.

Looking over these behaviors, it seems that students who struggle with analysis and multiple choice problems — that is, those who work hard, not those who work most efficiently or even most effectively — are the ones who get the most benefit when it comes to analyzing problems in the future.

5.6.3 Sheer Exposure to Problems

One can put forth the idea that the questions in FE3 can also be answered through experience or "physical intuition," i.e., that sheer exposure to multiple problem types is enough to show a student the right method. We do not claim that sufficient exposure to physics problems will not, in the long run, teach students to perform some form of analysis. What we claim is that intentionally practicing analysis speeds this process of acquiring "physical intuition" and analytical capabilities.

To illuminate this issue, we can examine one particular gauge: the Number of First Attempts. Like all gauges and behaviors in this study, it is broken down by problem type. If we examine the correlations between this gauge and the FE3, we find that traditional homework problems have a positive correlation of $r = 0.427$, while analysis problems yield a correlation of 0.636. These are significantly different at the $p = 0.066$ level (one-tailed). We believe that this shows the practice of analysis problems to be much more beneficial for later analysis than the practice of traditional problems. Despite the fact that

there were more than twice as many traditional problems as analysis problems, attempting analysis problems was much more important.

It is worthy of attention that neither the amount of time taken, nor the number of attempts made, nor whether one starts early or late, none of these gauges make the least amount of difference on FE3. Even the average score per attempt, which is the single best overall predictor out of all the gauges, provides no insight as to whether analysis or traditional problems are more effective at teaching students to analyze as FE3 requires. In fact, most gauges have no significant correlations with this section at all. Merely attempting the problems at all seems to be a useful thing to do, and more so for analysis problems.

5.6.4 Doing Well is Still Important

Only two other gauges have significant correlations with FE3: Average Credit per Problem and Number of Problems with Full Credit. In this case they tell us that there is value (as one would expect) not only to starting the problems, but to finishing them well. The correlation with analysis problems is once again higher than with traditional problems, adding to the argument that analysis problems improve the ability to analyze more effectively than traditional problems do.

5.6.5 Other Evidence

We have a veritable sea of evidence that analysis problems are, at the very least, not harmful to the students. Many gauges and behaviors show positive correlations between activity on analysis problems and performance later in the course, on par with those seen from traditional problems. Unfortunately, it has proven difficult to show whether work on analysis problems is more or less effective in general.

One attempt was made to isolate the effectiveness of certain problem types on the whole of students' homework practice in Physics 151. Each student's score for each gauge was scaled by the number of problems of a particular type they attempted. For instance, a student whose average Final Score gauge was 0.8, and who attempted 70% of the analysis problems in the semester, would have his Final Score gauge scaled to 0.56. This was done for each problem type, and the correlations for certain gauges were compared. This method effectively treats the students who took fewer attempts as a somewhat faulty control group.

It was found that, when scaled in this manner, most gauges showed a significant improvement (more positive) in correlation factors over unscaled gauges. Traditional-based scaling provided the greatest increase, followed by analysis-based, multiple-choice-based, and conceptual-based scaling. The changes were not always large, but appeared in most gauges, and in all correlation factors from those gauges. When compared to scaling by total number of attempts, traditional and analysis problems came out above that, with MC- and conceptual-based scaling below.

Our interpretation of this is as an overall "ranking" as to which problem types are the most effective when it comes to improving students' final grades. Under this interpretation, analysis problems were not as effective as traditional problems, but were more effective than MC and conceptual problems when it came to performance.

It should be mentioned that we do not recommend removing any problem types entirely, as all seem to have at least some benefit. Directed research on this topic would be necessary before such recommendations could be made.

CHAPTER 6

SUMMARY AND CLOSING REMARKS

This final chapter discusses possible directions for future research in this area, and the use of this sort of analysis in the evaluation (both formative and summative) of courses and teachers.

6.1 Implications For Future Research

The first area we would choose for future research is, unfortunately, the cry of experimentalists everywhere: more data! More data, on more topics, gathered more reliably, with less noise. While the last item seems unlikely in the field of educational research, the first three seem possible. There are also numerous items from this study that could stand to be revisited in future works, both for experimental verification of our findings and for the purposes of deeper exploration.

We start this section with a discussion of what other gauges one might attempt to form, continue into longitudinal studies and the possibility of generalization, speak to the need for verification of various behaviors, and end with a few miscellaneous unanswered questions.

6.1.1 Gauges from Other Data

There were several types of data that could have been used to create new gauges, if they were available. Such gauges might be useful for disentangling student activities and understanding certain behavior patterns.

6.1.1.1 Cooperation

Cooperation with other students is often cited as an important factor in improving understanding and performance. Unfortunately, the nature of this study's data made it nearly impossible to collect any useful information on group work or collaboration. Data on group work would almost necessarily be collected through self-reported means, which creates an additional source of error. Diligent recording of every instance of cooperative work seems almost impossible. Would a student who discusses a problem over dinner consider that important enough to report? Would they even remember such an event if it didn't help them solve the problem?

On the other hand, information on weekly study groups, tutoring sessions, help room use and the like should be relatively reliable, and could be used to create fairly robust gauges on a week-by-week or term-long basis. Possible gauges include "number of group meetings," "number of tutoring sessions," "visits to professor," and an overall "number of collaborative events" gauge. We suspect that the last gauge may be most effective, allowing us to discriminate between students who work alone, those who occasionally work with others or seek help, and those who use group work as their primary mode of operation.

Kotas and Finck (2002) used surveys, log data from the LON-CAPA homework system, and institutional data to examine the effects of homework collaboration between students on performance. Unfortunately for us, their homework system reports somewhat different data from ours, and we could not use OWL to detect collaboration without a significant system-wide overhaul.

6.1.1.2 Outside Influence

We have mentioned several times that the data collected through OWL is very "noisy," which is an odd thing to say about data without any experimental uncertainty. After all, if a student started 187 out of 250 problems, there's no uncertainty about it. There's no chance that he or she "might have started a problem" - OWL is quite definite on this point. The noise comes, instead, from the great variation in student activity from day to day and week to week.

We suspect that nearly all of this "noise" is due to factors external to the courses we examined. Besides the obvious requirement of studying for this course, students have other courses, familial and religious commitments, club meetings, sport practices and matches, social commitments, emotional entanglements, scuffles with the school administration, relatives in the hospital, birthdays, parental nagging, part-time jobs, sleepless nights, internships and co-ops, transfer or graduate school applications, work study, and uncountable other drains on their time. Frankly, it seems a miracle that we see any homework completed at all.

The difficulty in collecting data on out-of-class activity is that much of it is personal. One could certainly create a questionnaire (on OWL or hardcopy) that asked students to check boxes for various things that had happened to them that week, but we suspect that many students would be inclined to simply skip over it, and would rightly see it as an invasion of their privacy. The reliability of self-reported data can only be worsened when concerns over confidentiality come into play. A better approach might be asking specifically about those items that students are more willing to discuss (read: complain about). A weekly survey asking how many assignments a student currently has pending for other courses would be a good start.

6.1.1.3 Previous Knowledge

A third area that could significantly impact students' behavior and activity online is that of previous knowledge. Students who have seen the material before are probably more likely to spend less time on homework, with fewer attempts, and receive a higher grade. Luckily, a great deal of attention has been paid to this in the physics education literature (see, for example, Halloun and Hestenes (1985)). Tests such as the Force Concepts Inventory could be used to examine students' understanding in specific content areas, perhaps with the intent of calculating a weighting for gauges obtained from questions in that content area. Alternatively, a 1-5 scale could be added to each question, forcing students to rate questions on how familiar they are. Correlations between familiarity and various other gauges, especially as the term progresses, would be interesting to see.

6.1.1.4 Motivation and Emotional Engagement

Many studies have been done regarding student motivation and emotional engagement, both in a particular course and in their school as a whole (see, for example, Kuh (2003) regarding the National Survey of Student Engagement). One could apply any of a number of existing surveys designed to measure motivation, at key points during the semester (perhaps at the beginning and after each midterm), and use information from these to create motivation-oriented gauges. It would be interesting to look at different varieties of motivation and see their relationship to certain behaviors - such comparisons could aid in verifying the accuracy of the behaviors' names and their construction from the gauges. See the section "Are Some Gauges Misleading?" later in this chapter.

6.1.1.5 In-Classroom Work

The impact of teacher and course layout aside, we should pay some attention to what students do in the classroom. The advent of electronic classroom response systems provides an opportunity to add gauges based on data collected in the classroom. The questions used in the classroom were not reliably matched with response system logs in our study, which led us to discard the idea of a detailed analysis of these logs in favor of examining the more detailed homework logs. Given better record-keeping, this might be a fruitful area for future research. It would be interesting to see correlations between time spent on problems online and in the class, and also between performance in the classroom versus on homework.

6.1.1.6 The Dangers of Oversampling

It seems to us that one must engage in a balancing act when collecting extra data. The information obtained through OWL was unobtrusively gathered; it was generated behind the scenes as an integral part of students' homework activities. When collecting other data, it seems that students will necessarily be aware of certain data-gathering efforts, which may affect their behavior. Data-gathering will also take time away from other classroom activities, and one must balance data collection against the other needs of the course and the attention span and patience of the students. It seems to us that the best method for studying students' existing activities without altering them is to gather our information surreptitiously.

6.1.2 Longitudinal Studies

One item that we had hoped, but were unable, to examine was the change in students' activity over time. Large amounts of "static" from outside sources often made it impossible to say anything definite about the changes themselves, let alone their impact on performance. It is possible that this difficulty might be overcome through longer-scale studies, either with the same students or the same course.

Examining the same course, in the same term, can do more than simply improve the significance value of our data. By repeatedly tracking data for the same course over several years we can see certain "rhythms" of the course, such as the impact of yearly holidays, of the beginning and end of the academic term, and of certain days of the week. Treating this as background noise, we might be able to see whether certain courses have an overall increase or decline in their gauges, and thus be able to say something about a particular "crop" of students or a particular teacher. Individual student data is far too noisy for this approach, but data averaged or summed for an entire course should have a sufficiently low standard error for us to apply this method.

A more ambitious and difficult project would be the examination of particular students over the course of several years. Currently the largest barrier to attempting this is the very low prevalence of upper-level courses that use online homework systems, but this may change in future years. The goal would be to attempt to quantify changes in student behavior as they progress through their school career, so that teachers might anticipate the activity patterns of higher-level students. One could also see whether behaviors typical in upper-level students are helpful to students in freshman-level courses,

and, using control groups, whether encouraging such behavior improves performance. The sample size of this study would have to be relatively large, to smooth out external influences. Finally, one could examine student behavior in freshman year and compare it to performance in later years, to see whether there is significant agreement between those correlations and the more short-term ones we have examined in this study.

More comments on these ideas can be found in the "Classroom and Program Evaluation" section.

6.1.3 Generalization

The use of online homework has been spreading to other disciplines, and it would be helpful to see whether the results in this study could be generalized to other departments' courses. We do anticipate that science and mathematics courses, such as biology, chemistry, geology, and calculus, could use the gauges, behaviors, and findings from this study with minimal alteration, but it is certainly worth checking.

Disciplines that are conceptually farther from physics may find different mindsets and activities in their students. While there are still relatively few disciplines outside of the sciences that use online homework systems, the OWL system at UMass shows courses available for departments such as music & dance, theater, political science, art history, and communication. The number of non-science courses using online homework systems will no doubt increase as time goes on. Performing this sort of study in other areas will almost certainly turn up new question types, different gauge correlation patterns, and

different effects from our measured behaviors, and could be a useful tool for professors in those disciplines.

Generalization to other homework systems is important as well. While OWL shares many characteristics with other online homework systems, there are some that have their own unique characteristics. As an example, Pearson Education's "Mastering Physics" system has options to provide students with hints as they request them, and allows instructors to reduce the credit received for a correct answer on later attempts or after hints have been given. Studying this system with our methods could provide insight into the effects such differences have on later performance. Anecdotal evidence from instructors who use OWL suggests that seemingly small changes in the way an online homework system provides questions, feedback, and/or scores can have surprisingly large effects on student activity.

6.1.4 Are Some Gauges Misleading?

When we initially created the Late Attempts and Late Questions gauges it was expected that they would have positive correlations to performance - the idea behind allowing late attempts on questions was to encourage students to study homework after it was due. When we saw that the correlations were negative, we initially assumed that it was because the better students did not work on problems late. Later we realized that students who do late problems on a regular basis might feel encouraged not to try for full credit the first time around, since partial credit was available for late work, and that this might harm them on exams. In the end, we can speculate about the reasons behind a particular

gauge's correlations for hours on end, but solid data is needed in order to understand what's actually going on.

Research to establish the existence or nonexistence of causal relationships between gauges/behaviors and student performance is of paramount importance. If we as teachers intend to tell our students, "do X and your grades will improve, do Y and you will suffer" we had better be certain that X and Y are causes and not effects. Determining this will require a control group, and for some gauges and behaviors both groups would have to be willing to follow rather specific and well-worded directions.

Similarly, research that includes students' personality traits, along with interviews of students, may help to uncover underlying causes behind certain behaviors. This, in turn, may show us whether (for example) providing students the opportunity to do late work is actually beneficial, or is as detrimental as the correlations seem to say.

6.1.5 Miscellaneous Items

A control class, one without a single analysis question in the homework, would be invaluable for many purposes. Likewise, a control that completely lacks traditional problems would be very interesting. The exams would still contain both varieties of problem, but students would only have seen one type previously. It would be very interesting to see which course provided more benefit to the students, both in terms of short-term performance and long-term retention. Is a mixture of traditional and analysis problems best, or are analysis problems alone sufficient?

We have seen that students' responses to different question types shows that they react to them in different ways, at least unconsciously. Seeing whether students are able to consciously differentiate between analysis and traditional problems, both when prompted and when unprompted, would give even stronger evidence for a conceptual separation between these types of problem.

The difference in expectations and motivations for Physics 151 students, as shown by survey responses (see page 95), would be very interesting if it turned out to be statistically significant. One possible interpretation of this result is that students who are more interested in conceptual understanding are easily turned off by Physics 151 and similar courses, and thus become disengaged in the class, while those who are more interested in reinforcing and expanding on existing knowledge tend to find the course more engaging. The interaction between course style and student aspirations could be a major factor in "pushing" a student towards engagement or lack thereof, and it seems that such interactions are a ripe area for study.

Studying the homework habits of engaged and disengaged students may also prove useful, as our survey data hinted at a different way of viewing "time spent on homework" (see page 98). If one of these views is more fruitful, it may be worth encouraging.

Connecting gauges and behaviors to conceptual understanding, measured in whatever ways are possible, would be invaluable. All that we currently have is a comparison to performance. We can hope that performance and understanding are better connected now than they were in the days of early physics education and the "disaster studies," but hope is by no means a guarantee.

It would be interesting to see whether students who see their own gauges as the semester goes on will do better or worse. Personality issues no doubt conflate the issue, but the idea of examining the effect of such feedback loops is appealing to us.

Finally, I am forced to agree with this excerpt from Kotas (2003):

"For future study it would be fruitful to examine such behaviors in relation to other factors associated with learning, such as learning styles, attitudes, environmental variables, and learning strategies and methods. Additional contextual factors should be considered, to include students' extra-curricular activities. Such inquiries could prove to be beneficial for research involved with learning effectiveness."

In short, more data, on more topics, gathered more reliably.

6.2 Use In Classroom Or Program Evaluation

While the original intent of this thesis was to examine the effects of analysis, it would be a waste to discard the wealth of information gained on how to extract and categorize student behavior. There is much here that might be used by teachers and administrators for both formative and summative purposes.

6.2.1 Probing Students

The gauges and behaviors extracted from OWL can be used, with sufficient data, to review individual students for various purposes. This section suggests some possible ways for teachers and administrators to use the methods and data obtained in this study to examine their students. While reading this section remember that individual student data can be rather "noisy." A student whose gauges indicate trouble during a particular week may simply have three exams in other courses that week.

By examining an "incomplete" set of gauges drawn from the beginning of the semester, teachers might be able to identify students with difficulties, or those whose work habits need improvement. To see whether we could identify low-scoring students early on in the semester, gauges were calculated solely from the first half of the term and compared to performance as usual. While correlations from these first-half gauges tend to be weaker than those from the entire semester, they still give a good number of statistically significant results.

The best-correlated items from the first half of the term are Problems with Full Credit (positive), Average Score by Attempt (positive), Time Before Due (positive), and Attempts per Problem (negative, mostly from multiple choice and conceptual). Thus, if one has a student who is starting late, taking many attempts on multiple-choice problems, scoring badly on most attempts, and finishing problems without full credit, that student is more likely to do poorly on exams. It might be worthwhile to contact such a student for extra help, mentoring, tutoring, or simply a discussion of how to improve their chances. Note that Average Score per Problem in the first half of the semester is not a significant predictor of later performance, and so homework scores alone will not pinpoint the same students.

Because gauges cover very specific areas of student activity, it should be possible, assuming causal gauges, to tell students exactly what sorts of things they should do to improve their exam performance. Helping them implement such changes is even more important, but is outside the scope of this thesis. Whether the gauges involved are causal is an important topic for future research.

Behaviors collected from an entire term can be useful as well. If teachers are willing to share data with each other, it would be possible for a teacher in a second-semester course to know some details about the students who took the first-semester course. Specific behaviors that appeared in the first semester could allow the second-term teacher to place students in heterogeneous groups, direct teaching assistants or tutors towards inactive or uncertain students, and guess at which ones might do well in an honors section of the course.

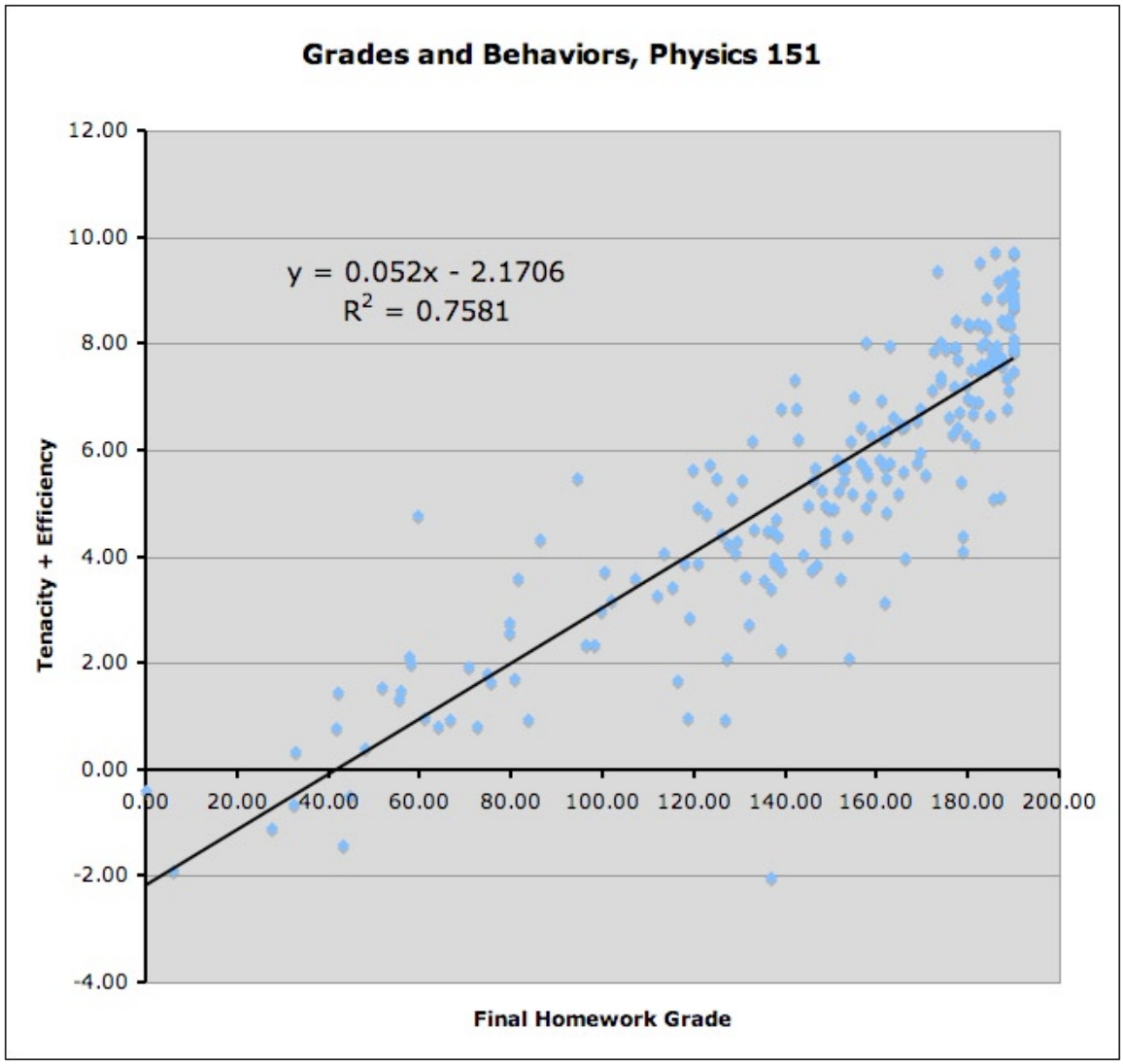
School administrators might find this information useful as well. If confidentiality is not an issue, students' advisors could be given access to records of their behavior. They could make themselves aware of students with bad habits by looking for patterns of inactivity or a lack of tenacity. They might also consider nominating students for awards based on measured behaviors rather than sheer performance. Face-to-face meetings are still important, of course, but this might provide an early warning of difficulty for those students who don't check in often. This can be especially useful for distance-learning programs in which true face-to-face meetings are difficult to impossible.

On a higher administrative level, a dean might be implementing a series of special-interest dormitories and be interested to see what effects such communities have on homework behavior. High-school principals might want to compare graduation or college entrance rates with homework tendencies, to better motivate their students. Such exploratory uses will be more time-consuming than merely looking over a single student's or course's gauges, but they also open the door for more interesting studies.

6.2.2 Evaluating Students

Because of the high correlation factors that can be obtained between student behaviors and their final performance, it may be worth considering grading students based on how well they display these behaviors. One can see from the graph below that the combination of the tenacity + efficiency behaviors does not precisely match the final homework grade for the term. This mismatch indicates to us that the two measures carry different information. Because of this, there may be some benefit in using a combination of the two

Fig. 6.1: Tenacity + Efficiency vs. Grade



scores, rather than just grade alone, to encourage good homework behavior. If efficiency is something we want our students to have, and the behavior with that name actually does measure what it claims to, why not grade students using that behavior? Naturally, this sort of grading scheme would have to either wait for a confirmation of the causal nature of certain gauges and the accuracy of the behaviors' names, or be part of such investigations. We do not suggest that such scores replace the entire semester's grade — quizzes and exams show relatively low correlation to any behaviors, and in any case there is too much conceptual difference between the two measures — but it seems worthwhile to use any insights we may have into students' activity for evaluation.

It is interesting to note that behaviors from the first half of the term did not, in our trial, show as much correlation with the third exam as they did with the course grade. The exceptions were as follows: Engaged students did well on the third exam if they had a habit of starting early in the first half of the term. Disengaged students did well on exam #3 if their average score per attempt was high. We are uncertain as to the reason for this; again, more research will likely be necessary.

Before any of these suggestions are implemented, it would be necessary to automate the process of extracting gauges and behaviors, preferably in a way that is built into the online homework system itself.

On a slightly different track, students might be interested to see their own level and ranking in the class when it comes to gauges and behaviors. Contests for efficiency, or some other combination of gauges that the professor wants to encourage, might drive

more competitive students to improve their standing (and hopefully their comprehension as well).

It is important in our minds that this be done in a way that promotes conceptual understanding, and further studies will be necessary to see what sorts of feedback and/or incentives will accomplish this. Some degree of morale control may also be necessary, to ensure that students who are rated low do not lose hope, and it may not be appropriate to rank students as the "5th most inactive in their class." However, when students understand the gauges and behaviors, and what methods are most effective in general and for themselves in specific, we feel that it cannot help but improve their performance.

Alternatively, gauges and behaviors could be used behind the scenes, to guide a more sophisticated or user-friendly automated feedback system. Displaying numerical values for calculated behaviors could be off-putting or incomprehensible to many students. Giving them practical suggestions as to what they might do to improve seems like it could be more effective. Some teachers might prefer that the system work by informing instructors rather than students, allowing them to direct comments and suggestions at the students themselves. Teachers' knowledge of individual students' personalities and proclivities could make this a more effective route than an automated system.

6.2.3 Probing Teachers and Courses

Despite the "noise" in our data, the sample size makes standard errors quite low. These gauges, taken from traditional homework questions in Physics 151 (top) and Physics 181 (bottom), are good examples. Time before Due is measured in days.

Table 6.1: Sample Gauges from Physics 151

Gauge	Average	Std. Error	StdErr / Avg.
Seconds to Respond	191	0.30	0.16%
Number of Attempts	2.53	0.0046	0.18%
Time Before Due	1.57	0.0063	0.40%
Avg. Score per Attempt	0.47	0.00063	0.13%

Table 6.2: Sample Gauges from Physics 181

Gauge	Average	Std. Error	StdErr / Avg.
Seconds to Respond	173	0.87	0.51%
Number of Attempts	3.11	0.019	0.61%
Time Before Due	1.59	0.025	1.58%
Avg. Score per Attempt	0.507	0.00214	0.43%

Because the standard errors are so low, it is possible to detect relatively small changes in a course's average gauges from year to year. While the correlations between gauges and performance may or may not change (especially since exams and exam scores can vary so much from one term to another), the raw gauges themselves can be quite useful.

As mentioned before, it will require an examination of gauges' and behaviors' causal power (or lack thereof) to tell which behaviors are worth encouraging, but the ability to measure average classroom behavior is useful even without that. Teachers with specific pedagogical goals and techniques can use the methods in this study to track students'

responses to their interventions, measuring how a particular teaching technique increases or decreases certain student activities.

For instance, if a teacher desired to have his or her students start on homework earlier, that teacher might implement certain changes in teaching methodology, or in the way students are supported or evaluated, to try to encourage such a change. Not all online homework systems present their teachers with data similar to the Time Before Due gauge, and those that do typically break the information down on a problem-by-problem basis. Individual variations in people and problems often swamp small changes in this data. By using the gauge for the entire term (or even a few weeks of the semester), teachers can tell whether their changes are effective.

Administrators might also be interested in using gauges to provide feedback or assessment for instructors and programs. If teachers use the same homework questions from year to year (as many college courses do), it becomes possible to track the dependence of various gauges on administrative factors. How does an increased class size change students' average score per attempt, or the average number of abandoned questions?

There is an opportunity here to fit specific teachers to specific courses, or to give teachers facts about their own abilities that they may not have known. Which teachers get the best performance from large or small classes? From students with certain majors? Do certain teachers seem to encourage specific behaviors in their students? Teachers' personal preferences must be taken into account as well, but just because a particular lecturer enjoys (for example) large lecture courses doesn't mean that he or she is necessarily good at teaching them.

6.3 Final Words

We feel that this study has been successful in its original goals: measuring and characterizing student homework activity, and determining whether practice of and exposure to analysis problems is beneficial to students. Our research shows that the ability to analyze problems on exams is significantly improved through prior exposure to and practice of analysis problems, and that such exposure does not reduce students' ability to handle other kinds of problem.

The "machinery" we created in order to answer this question and examine our students more carefully has proven very useful, and may be just as powerful an outcome as the answer itself. Examining students' homework activity yields a goldmine of different evaluative tools, as well as suggestions for students as to how they can maximize their performance in introductory physics courses (a question that plagues many majors and non-majors alike). We believe this may be the first serious examination of how college-level students do their homework.

We hope that this groundwork will be useful for future researchers in this area, either as inspiration for new electronic homework systems and educational materials, or as a starting point for further investigation. There has been very little research thus far in this area, and we hope that in the future there will be more attention paid to how students do their homework.

APPENDICES

Appendix 1: Glossary

This appendix is a listing of useful definitions and terminology pertinent to this thesis. It is arranged in alphabetical order.

Analysis Questions: See pages 15-16 for descriptions and examples of analysis homework questions.

Behavior: A higher-level measure of student activity, constructed from two or more Gauges (q.v.). See page 62 for descriptions and explanations of the various behaviors measured in this study.

Conceptual Questions: See pages 16-17 for descriptions and examples of conceptual homework questions.

Correlation: When the term "correlation" is used in this thesis, it is a standard Pearson's "r" correlation. One section (page 46) discusses the Eta correlation ratio, which will be distinguished from Pearson's "r" at all times. No other correlation measures were used.

Course Feedback Surveys: Both Physics 151 and Physics 181 administered certain surveys to the students, primarily to obtain feedback about the previous week's lectures and homework. Physics 181 gave bi-weekly surveys, while Physics 151 gave them on a

weekly basis, and in greater detail, with additional questions covering a different topic each week.

Definition Questions: See page 18 for descriptions and examples of definition homework questions. These were grouped with multiple choice questions (see below) for this study.

Engagement: Some parts of this study refer to engagement; by this we typically mean a purely functional form of engagement, rather than the emotional or cerebral interpretations that have also been studied elsewhere (see page 10). By our definition, students who attempted 85% or more of the class' homework assignments, lecture prep assignments, PRS problems, course feedback surveys, and quizzes were counted as "engaged." Engaged students must also have attended all of the major exams (midterms and the final). Performance on these items was deemed irrelevant to our definition of engagement; the only thing that mattered was that the student had a nonzero score on the vast majority of items. In Physics 181 the surveys were not emphasized as greatly as in Physics 151, so they were not used to define engagement in that course.

Exams: Both courses examined in this thesis involved multiple midterm exams, as well as a final exam. The exams in Physics 151 used machine-graded multiple-choice questions, with partial credit available. Physics 181 used a hand-graded format that included multiple-choice questions, short-response questions, and longer exercises that required calculation.

Gauge: A measurement gathered from a student's aggregated online homework data. See page 52 for descriptions and explanations of the various gauges. Not all gauges indicate student conduct — some are performance measures — but all are gathered solely from electronic homework data.

Homework Type: A category in which students can be placed based on their scores in multiple different behaviors (q.v.). This is the highest level description of student activity that was generated during this study.

Lecture Prep: Assignments due just before a lecture begins, which cover topics to be seen in the upcoming lecture. Lecture prep assignments are intended to increase familiarity with the topic, to raise questions in the students' minds before lecture, and to help ensure that students read their textbook. These assignments were typically shorter and easier than the regular homework, and were counted separately. They were also much more plentiful, being due before each lecture rather than once a week. Only Physics 151 (Fall 2003) assigned lecture prep; Physics 181 did not.

Multiple Choice Questions: See pages 17-18 for descriptions and examples of multiple-choice homework questions.

OWL: "Online Web-based Learning," an electronic homework system developed at UMass Amherst, and used in many academic departments there. Students log on with a web browser, view homework, enter answers, get feedback, and can take multiple attempts. OWL was also used to deliver lecture prep assignments and surveys. As of this writing, OWL can be found at http://owl.oit.umass.edu/

PCA: Principal Component Analysis is a multilinear method that attempts to reduce a set of data (composed of multiple measurements on a number of objects) to a small number of underlying components or factors. Each component can be expressed as linear combination of the original measurements. Combinations of these components yields a data matrix substantially similar to the original. Alternative terms for PCA include Eigenanalysis and Principal Factor Analysis. In this thesis, the "objects" are students, and the measurements can be either gauges or behaviors, with the underlying factors being higher-order measurements.

Physics 151: An introductory mechanics course designed for engineering and many science majors. Biology students and other life-science majors typically take Physics 131 instead, so 151 is typically composed of freshman and sophomore engineers and chemists. The course covers basic motion, forces, energy, rotational mechanics, and universal gravitation. Physics 181 (below) is considered a more difficult course, as it uses calculus more extensively and goes into greater depth on the same number of topics.

Physics 181: An introductory mechanics course, designed for first-year physics majors. A moderate number of astronomy and engineering majors can also typically be found in this course, as well as students from other majors desiring a greater challenge.

PRS: "Personal Response System," a classroom feedback and response system designed by the GTCO CalComp company. Used for formative assessment in the sections of Physics 151 and Physics 181 studied in this thesis. Problems and surveys are posted in-class, for the whole class to see. Each student uses his or her own "clicker" to respond

within a certain time limit (typically less than 3 minutes). The teacher has the option to show histograms of the responses to the class. PRS was also used to collect attendance-related data.

Traditional Questions: See page 18 for descriptions and examples of traditional homework questions.

UMass Grading Schemes: Until recently, UMass Amherst used a grading scheme slightly different from the usual plus/minus system. The highest grade was an A, followed by AB, B, BC, C, CD, D, and F. A grade of "Incomplete" was also possible. There was no "DF" grade. Physics 151 was graded using this scheme, while Physics 181 uses the more conventional A, A–, B+, ..., F grading scheme that was introduced in the intervening year. For the purposes of this study letter grades were generally eschewed in favor of raw scores, but the coarser-grained letter grades were occasionally useful for histograms.

Appendix 2: Igor Code

The following routines were used in Igor for the varimax rotation method we employed in principal component analysis. These functions were included with the Igor software distribution. Double slashes (//) precede comments. Note that some lines wrap unavoidably due to the size of this page.

```
    ////////////////////////////////////////////////////////////
    // 14NOV02
    // The following function performs a Varimax rotation of inWave subject to the specified epsilon.
    // The algorithm follows the paper by Henry F Kaiser 1959 and involves normalization followed by
    // rotation of two vectors at a time.
    // The value of epsilon determines convergence.  The algorithm computes the tangent of 4*rotation
    // angle and the value is compared to epsilon.  If it is less than epsilon it is assumed to be essentially
    // zero and hence no rotation.  A smaller value of epsilon leads to a larger number of rotations.
    // The function returns the number of rotations peformed (each rotation is on two vectors).  The function
    // creates the wave M_Varimax that contains the rotated matrix.
    ////////////////////////////////////////////////////////////

    Function WM_VarimaxRotation(inWave,epsilon)
          Wave inWave
          Variable epsilon

          Variable rows=DimSize(inWave,0)
          Variable  cols= DimSize(inWave,1)

          // start by computing the "communalities"
           Make/O/N=(cols) communalities
           Variable i,j,theSum
           for(i=0;i<cols;i+=1)
               theSum=0
               for(j=0;j<rows;j+=1)
                    theSum+=inWave[j][i]*inWave[j][i]
               endfor
               communalities[i]=sqrt(theSum)
           endfor

           Make/O/N=(2,2) rotationMatrix
           Make/O/N=(rows,2) twoColMatrix
```

```
          Duplicate/O inWave, M_Varimax         // the calculation is
done in place so M_Varimax will be the wave holding the rotated vectors.
          // normalize the wave
          for(i=0;i<cols;i+=1)
              for(j=0;j<rows;j+=1)
                   M_Varimax[j][i]/=communalities[i]
              endfor
          endfor

          // now start rotating vectors:
          Variable convergenceLevel=cols*(cols-1)/2
          Variable rotation,col1,col2
          Variable rotationCount=0
          do
              for(col1=0;col1<cols-1;col1+=1)
                   for(col2=col1+1;col2<cols;col2+=1)
                        rotation=doOneVarimaxRotation(M_Varimax,rotation
Matrix,twoColMatrix,col1,col2,rows,epsilon)
                        rotationCount+=1
                        if(rotation)
                              convergenceLevel=cols*(cols-1)/2
                        else
                              convergenceLevel-=1
                              if(convergenceLevel<=0)
                                     for(i=0;i<cols;i+=1)
                                          for(j=0;j<rows;j+=1)
                                                M_Varimax[j][i]*=communa
lities[i]
                                          endfor
                                     endfor
                                    KillWaves/Z rotationMatrix,twoColMat
rix,communalities,M_Product
                                    return rotationCount
                              endif
                        endif
                   endfor
              endfor
         while(convergenceLevel>0)

         KillWaves/Z rotationMatrix,twoColMatrix,communalities,M_Product
         return rotationCount
    End




    ////////////////////////////////////////////////////////////////
    // this function is being called by WM_VarimaxRotation(); it has no
use on its own.  The function
    // rotates a couple of vectors at a time.  We keep rotationMatrix,tw
oColMatrix in the calling routine
    // so that they are allocated only once and not each time this func-
tion is called.
```

```
    // To optimize things further consider the xx and yy assignments.
    ////////////////////////////////////////////////////////////////////
    Function  doOneVarimaxRotation(norWave,rotationMatrix,twoColMatrix,c
ol1,col2,rows,epsilon)
          wave norWave,rotationMatrix,twoColMatrix
          Variable col1,col2,rows,epsilon

          Variable A,B,C,D
          Variable i,xx,yy
          Variable sqrt2=sqrt(2)/2

          A=0
          B=0
          C=0
          D=0

          for(i=0;i<rows;i+=1)
               xx=norWave[i][col1]
               yy=norWave[i][col2]
               twoColMatrix[i][0]=xx
               twoColMatrix[i][1]=yy
               A+=(xx-yy)*(xx+yy)
               B+=2*xx*yy
               C+=xx^4-6.*xx^2*yy^2+yy^4
               D+=4*xx^3*yy-4*yy^3*xx
          endfor

          Variable numerator,denominator,absNumerator,absDenominator
          numerator=D-2*A*B/rows
          denominator=C-(A*A-B*B)/rows
          absNumerator=abs(numerator)
          absDenominator=abs(denominator)

          Variable cs4t,sn4t,cs2t,sn2t,tan4t,ctn4t

          // handle here all the cases :
          if(absNumerator<absDenominator)
               tan4t=absNumerator/absDenominator
               if(tan4t<epsilon)
                    return 0
// no rotation
               endif
               cs4t=1/sqrt(1+tan4t*tan4t)
               sn4t=tan4t*cs4t

          elseif(absNumerator>absDenominator)
               ctn4t=absDenominator/absNumerator
               if(ctn4t<epsilon)
// paper sec 9
                    sn4t=1
                    cs4t=0
               else
                    sn4t=1/sqrt(1+ctn4t*ctn4t)
                    cs4t=ctn4t*sn4t
```

```
            endif
        elseif(absNumerator==absDenominator)
            if(absNumerator==0)
                return 0;
// undefined so we do not rotate.
            else
                sn4t=sqrt2
                cs4t=sqrt2
            endif
        endif

        // at this point we should have sn4t and cs4t
        cs2t=sqrt((1+cs4t)/2)
        sn2t=sn4t/(2*cs2t)

        Variable cst=sqrt((1+cs2t)/2)
        Variable snt=sn2t/(2*cst)

        // now converting from t to the rotation angle phi based on the
signs of the numerator and denominator
        Variable csphi,snphi

        if(denominator<0)
            csphi=sqrt2*(cst+snt)
            snphi=sqrt2*(cst-snt)
        else
            csphi=cst
            snphi=snt
        endif

        if(numerator<0)
            snphi=-snt
        endif

        // perform the rotation using matrix multiplication
        rotationMatrix={{csphi,snphi},{-snphi,csphi}}
        MatrixMultiply twoColMatrix,rotationMatrix
        // now write the rotation back into the wave
        Wave M_Product
        for(i=0;i<rows;i+=1)
            norWave[i][col1]=M_Product[i][0]
            norWave[i][col2]=M_Product[i][1]
        endfor
        return 1
    End
```

Appendix 3: Minaei-Bidgoli's Thesis:

Behrouz Minaei-Bidgoli's thesis, submitted to Michigan State University in 2004, bears some similarities to our work. His thesis used more sophisticated modeling tools than we did (cluster analysis, genetic algorithms, pattern recognition, etc.), with much more technical detail overall. This is to be expected in a computer science thesis focusing on the methodology of data mining.

His study also had a much larger sample size, from many different courses and departments. The homework system examined was LON-CAPA, widely used at Michigan State University.

Overall, we examined a larger number of gauges than he did (his work calls them features). His gauges are:

1. Total number of correct answers - corresponds to our Full Credit gauge.

2. Getting the problem right on the first try, vs. those with high number of submissions - no direct correspondence, but has some relation to Number of Attempts and Average Score per Attempt.

3. Total number of attempts before final answer is derived - corresponds to our Number of Attempts gauge.

4. Total time that passed from the first attempt, until the correct solution was demonstrated, regardless of the time spent logged in to the system. Also, the time at

which the student got the problem correct relative to the due date. Usually better students get the homework completed earlier. - These correspond to my Elapsed Time and Start Time, but specify that the problem was completed correctly.

5. Total time spent on the problem regardless of whether they got the correct answer or not. Total time that passed from the first attempt through subsequent attempts until the last submission was demonstrated. - This is probably Elapsed Time, though it might be a cumulative Seconds To Respond.

6. Participating in the communication mechanisms, vs. those working alone. LON-CAPA provides online interaction both with other students and with the instructor. - We have no gauge that is comparable to this feature.

His "performance percentage" for classifying students's final grades accurately hovers around 80%, which is roughly the same number as the best correlation factors we can obtain with final grade (see page 104 for the tenacity and efficiency combination). It is encouraging that our very different methods yielded similar outcomes. His optimized results, after genetic algorithms were applied, came out to about 90%, which is better than what we have been able to achieve.

Minaei-Bidgoli's concerns involve minimum times for classification and optimization, because he hopes to use his methods in a real-time system. If our methods are generalizable, they should be able to be implemented in such a system without as much computational overhead, but they don't have the inherent adaptability of his methods to other classes and disciplines.

Appendix 4: Subspaces and Ideal Vectors

The correlation tables we had created between gauges and performance could be seen as an exceptionally complex vector function: the worksheet took an 18-dimensional vector, with each component being a weighting for one gauge, and returned a series of scalars (the correlation factors). Each linear combination of gauges could be seen as a vector in a very non-orthogonal 18-dimensional space.

After examining the PCA data from Physics 181, we performed a test of Physics 181's "ideal" linear combination — or ideal vector — to see whether it was a good predictor for Physics 151. High correlation factors were indeed obtained, meaning that this particular combination was good for both classes. Testing Physics 151's "ideal for exams" vector in Physics 181 showed similar results. Neither one was quite as good in the other class as in their own, and neither one was quite as good as the "native" ideal vector, but both ideals were very good predictors in both courses.

What made PCA particularly interesting in this case is that it returns not merely a series of numbers, but a reduced set of orthogonal vectors that spanned a subspace of the original 18-dimensional space. It seemed possible that the four basis vectors returned by PCA (two from Physics 151 and two from Physics 181) were actually located in the same plane, and simply rotated from each other by some amount.

Unfortunately, this conjecture turned out to be incorrect. The plane created by Physics 151's two factors is "close" to, but not exactly coplanar with, the plane created by Physics 181's factors. If we call one of the factors X and the other Y in each of Phys-

ics 151 and Physics 181, then X151 and X181 were ~28 degrees apart, while Y151 and Y181 were ~25 degrees apart. This indicates some additional rotation through at least one more dimension.

While the classes' ideal vectors were in their respective planes, neither of them was in both planes (nor could the intersection line between the two planes be considered the least bit "ideal"). The two ideal vectors turned out to be ~22 degrees apart from each other, which is not really that much in an 18-dimensional space.

While the hypothesis for this portion of the study turned out to be incorrect, we found that merely considering and testing it provided a useful and different viewpoint on reduced factors, and led to a deeper understanding of what PCA actually returned.

Appendix 5: "Raw" Data

For those who are interested, we present some less processed forms of the data gathered in this study. The truly raw data returned by OWL is noted on page 13; there are over 150,000 entries of this type.

Gauges were calculated through the use of Excel's Pivot Tables. Here is a small sampling of the Elapsed Time data table from Physics 181:

Table A.1: Elapsed Time Data

UMass ID	Analysis	Conceptual	MC/Def.	Traditional
(student #1)	0:18:54	0:00:59	0:12:08	0:29:39
(student #2)	0:14:59	6:57:32	0:06:23	0:21:22
(student #3)	0:44:28	18:59:11	9:22:41	0:16:47
(student #4)	1:16:08	9:12:33	0:22:52	5:39:20
(student #5)	40:54:53	0:55:51	0:03:55	17:30:41
(student #6)	0:12:28	8:19:04	0:22:17	1:42:39
...				

Here is a sampling of the Average Score per Attempt gauge for the same students:

Table A.2: Average Score Data

UMass ID	Analysis	Conceptual	MC/Def	Traditional
(student #1)	0.75	0.5	0.6072727	0.539375
(student #2)	0.9054545	0.7	0.7783333	0.6459375
(student #3)	0.3738888	0.4642857	0.4738461	0.3306521
(student #4)	0.6618181	0.7727272	0.6855555	0.5664285
(student #5)	0.62	0.6	0.4166666	0.58
(student #6)	0.7492857	0.4583333	0.8	0.4695744

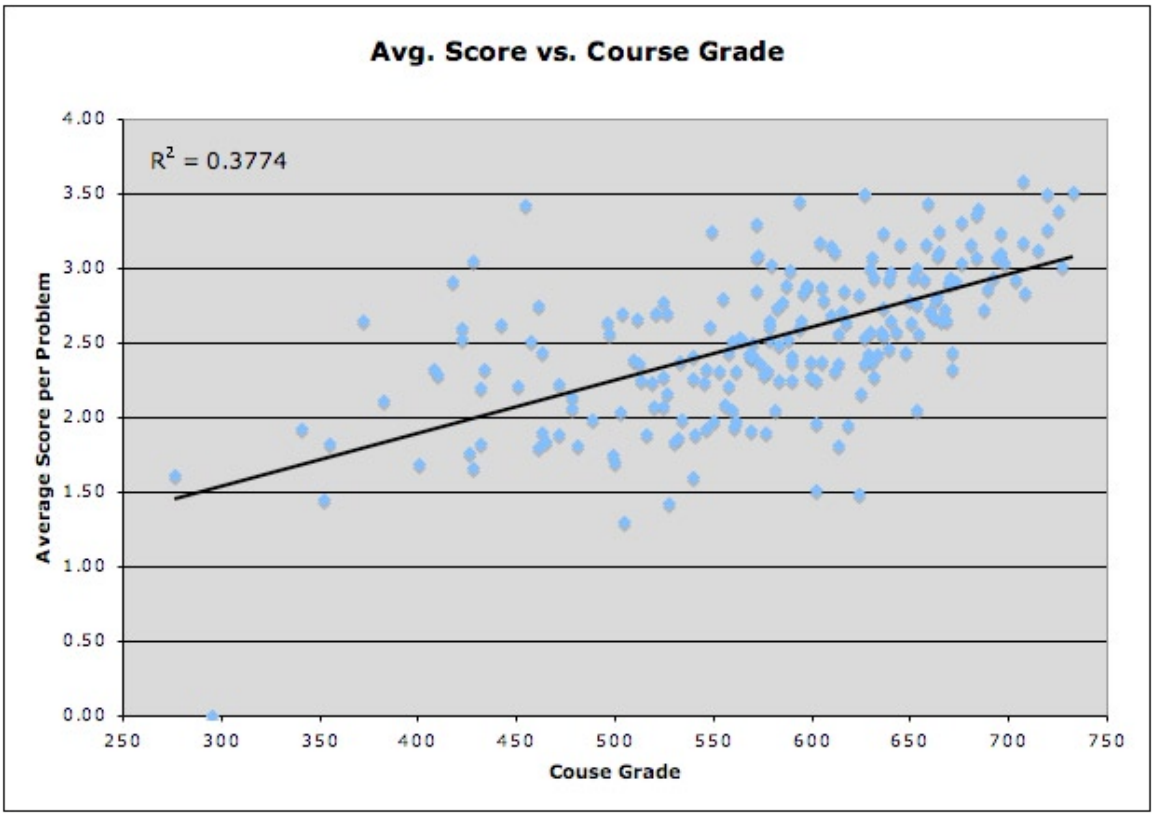
Here are some raw scores for Behaviors for a different group of students. Note that because Behaviors were created from differing numbers of Gauges, they have different ranges:

Table A.3: Behavior Data

Inactive	Uncertain	Tenacious	Efficient	Frustrated	G-C	S&S
-2.48	3.79	0.46	1.21	2.54	6.37	5.72
-2.05	6.59	7.97	2.52	2.82	7.97	8.53
-0.50	5.37	2.00	0.22	4.94	5.52	6.84
-2.18	8.02	6.87	1.44	5.63	9.76	8.86
-2.86	4.15	-0.54	1.24	3.16	4.96	7.03
-0.66	4.68	3.10	1.73	3.57	6.35	6.24

On the next two pages are scatter plots of various gauges compared with the final course grade. The gauges have all been scaled to a 0-4 value. The gauges were chosen for variety; some of them show relatively good correlation, while others are weak or difficult to see. Only the Elapsed Time graph shows an insignificant correlation. All of these plots are taken from Physics 151.

Fig. A.1: Scatter Plots A



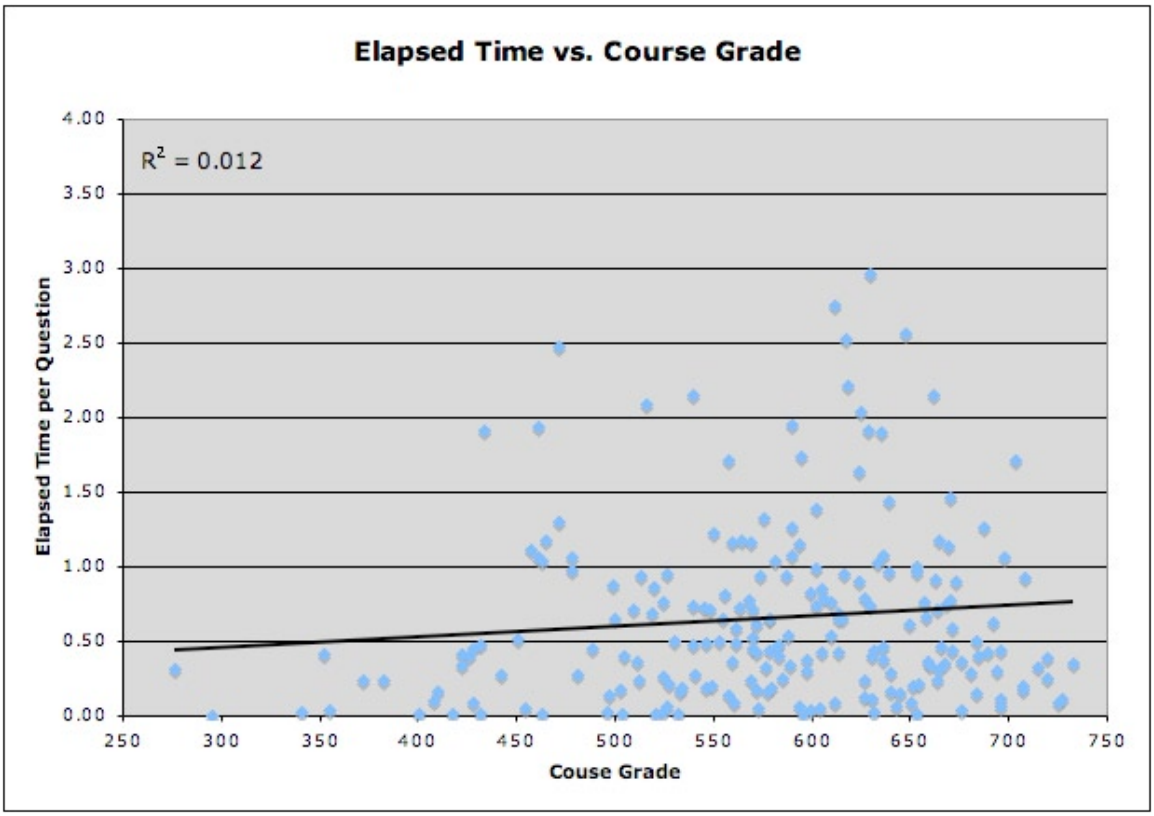
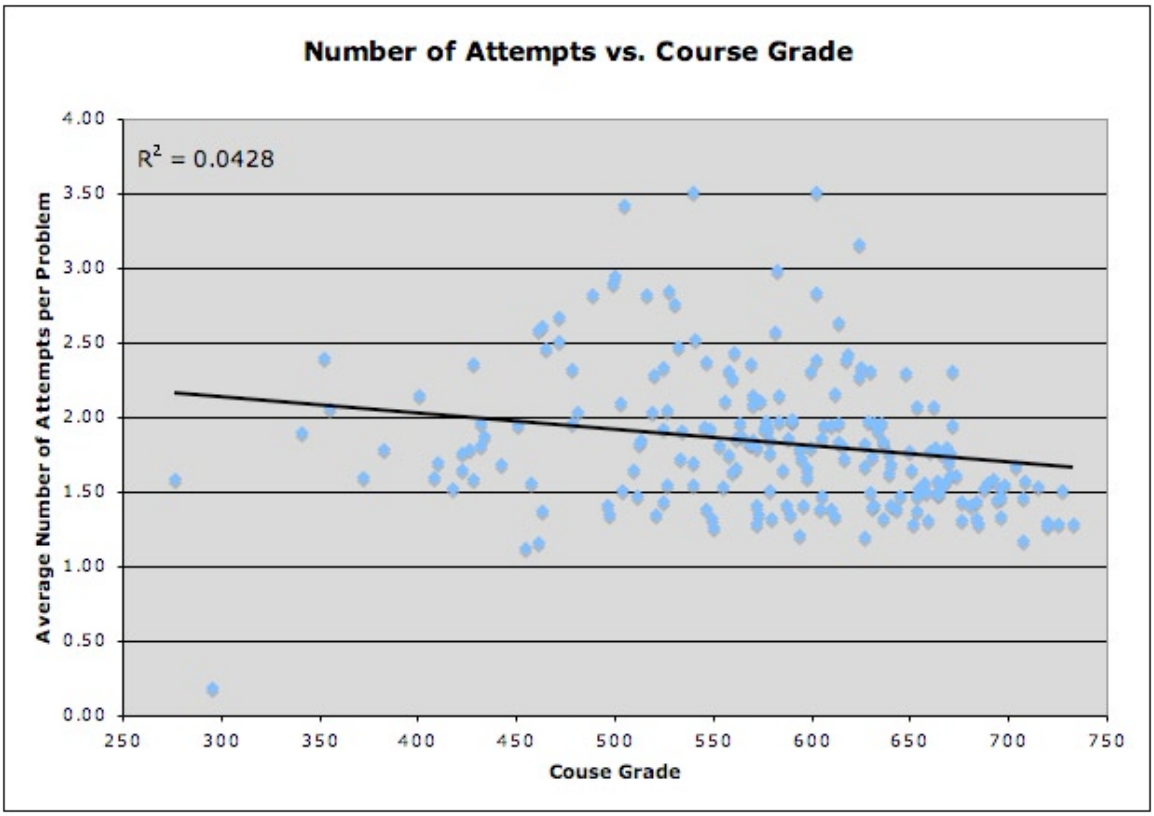
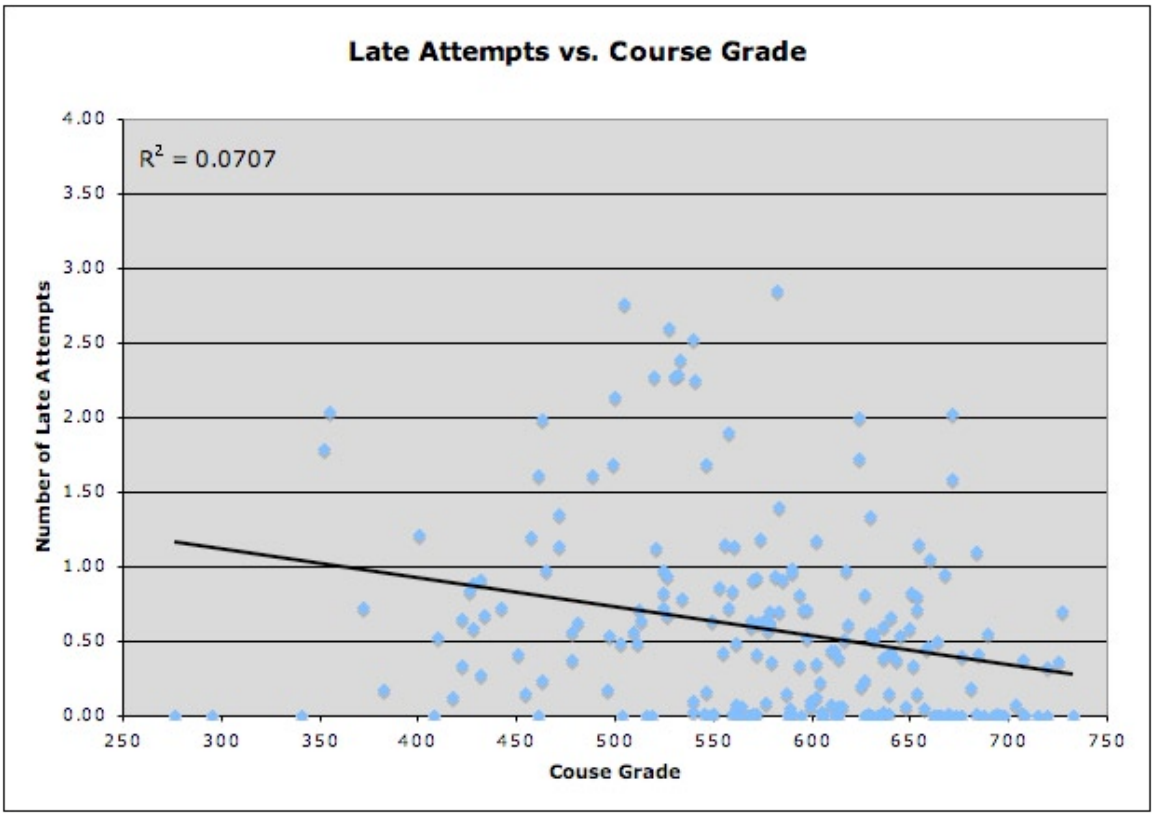


Fig. A.1: Scatter Plots B





Appendix 6: Student Homework Types

The creation of student homework types or categories was one of the original goals of this study (see page 1), giving the ability to characterize students in a quick and meaningful manner. Many different methods were attempted, with varying degrees of failure. In the end it was decided that all of these approaches to homework types were either untenable or did not contribute significantly to our understanding of student activity and its impact. This section describes several attempts at creating such categories, their advantages and disadvantages, and the reasons they were eventually discarded.

Homework Types via Simple Statistics

There were three attempts made to use relatively simple statistics in the creation of student homework types. One used the students' rankings in various behaviors, another looked at which behaviors were strongest for each student, and a third examined consistency across different homework types.

The first attempt to create student homework types determined whether a particular student was rated as "high," "low," or "medium" in a particular behavior, broken down by problem type. We tried several different methods for creating cutoffs, including ones based on percentiles, standard deviation, and standard error, eventually settling on cutoffs of 80% and 20%. The larger the "medium" section, the fewer homework types would be populated. This was important, as three options in each of seven different behaviors gives $3^7 = 2187$ different possible homework types. However, a large central section also

tended to put most students into the “average in all categories” type. The 80%/20% cutoff seemed to give the best split.

Unfortunately, no matter where the cutoffs were placed, many students in both courses ended up in a category on their own, leading to small-number statistics and unreliability. Even with all problem types aggregated together, there were rarely more than ten students in a particular category. In those categories for which average grades and standard deviations could be found, the categories seemed well-separated, but there was a general feeling that too much was “slipping through the cracks.” For all these reasons, this method was abandoned.

Another attempt looked at whether or not students were consistent in their behaviors across problem type — for example, were they tenacious on both Analysis and Multiple Choice problems? After a suitable measurement of overall consistency was created, this method was discarded, because said measurement had no significant correlation with performance. Placing students into a particular category seems less worthwhile when there are no noticeable effects of being in said category.

Still another use of simple statistics started with a gaussian normalization of all the student behavior data, so that each behavior had an average of zero and a standard deviation of one. Each behavior was thus equally weighted numerically, so that one could tell whether a student was, say, more tenacious than he or she was efficient. Using this idea, we found which behavior was strongest for each student, in either the positive or negative direction, and used this as their homework type. Each student thus fell into one of fourteen categories, with each category contained between zero and about twenty students.

The advantage of this approach is that it underlined the results discovered through the use of correlation factors. Behaviors with a strong correlation showed a great amount of separation between their two types — students in the "tenacious is strongest" and "not-tenacious is strongest" categories had final grade averages that were separated by multiple standard errors. Behaviors with lower correlation factors showed averages that were within a standard error of each other, and thus statistically indistinguishable. This method's drawback is that it gave us no information that was not already present in the behavior correlations. The reinforcement of our existing conclusions was comforting, but not useful for our intended purpose.

Homework Types via PCA

Realizing that simple statistics would not give us any more insight than correlation factors already had, we turned to Principal Component Analysis. We hoped that factors obtained through PCA of student behaviors would yield useful categories into which our students could be placed.

Depending on which course was being analyzed and how one interpreted the "scree test," one could find two or three different factors through PCA. This was a significant improvement in the number of categories over our first method listed above: rather than $3^7 = 2187$ possible homework types, we had $3^3 = 27$ of them. Averages and standard deviations indicated that there were several types that were well-separated when it came to final performance.

One major drawback prevented us from going forward with this method: the categories created in this way were nearly uninterpretable. Even those that corresponded to just two behaviors were sometimes nonsensical, combining (for instance) students who were both highly lazy and highly tenacious. Some categories would turn out to have significant percentages of all seven behaviors. While these PCA-based categories were effective at separating students, they seemed to have little real information content and little that we could understand.

Intentions and Attitudes

Because all attempts at discovering emergent categories had yielded either uninterpretable or less-than-useful results, we considered the possibility of creating homework types from behaviors in the same way that behaviors were created from gauges (see pages 62-65). In this way the meaning could be built into the categories to begin with, rather than expected to arise from them spontaneously. Two conceptually different but procedurally identical approaches were taken: one to create a set of Intentions, another to create Attitudes.

Many students come into a course with a particular intention: a goal for their performance or personal improvement. Some students intend to get a good grade, some intend to understand the material well, others are there to fulfill a departmental requirement or simply to pass. Some want to skate through with the minimum amount of effort. Still others are likely to come in with no discernible goals, and are simply there to "take the class." Such intentions are not necessarily orthogonal — someone could want to under-

stand the material *and* get a good grade, for example — but there seems to be a greater conceptual orthogonality inherent in student intention than is found in behaviors.

While intentions can be seen as cerebral approaches to the course, attitudes can be seen as emotional approaches to the homework. Students can be persistent (working hard for a long time), they can feel lost (acting frustrated, inefficient and unsure), they can be industrious (using their time efficiently to quickly get through the homework), or they can be uncaring (showing little interest in doing work or getting a good score). Other attitudes are likely possible as well.

The goal of creating a set of intentions was abandoned before any statistics were applied. The entire approach felt like attempting to build a house of cards — there was little to no indication that our calculated "intentions" would correspond to what students actually wanted to accomplish. While students were asked this sort of question at the beginning of the year, the unreliability of self-reported survey data and the spotty response to surveys meant that there was essentially no way to check our results. The "attitudes" approach felt slightly more reliable, and showed some promise when it came to correlation and other statistical measures, but the question as to whether the underlying behaviors were accurately named remained.

In the end, it was decided that these two approaches were untenable. When and if student behaviors (as measured in this study) are verified, it may be possible to return to these methods. For now, we have abandoned them.

Non-Exclusive Types

Our final attempt at creating student homework types involved non-exclusive categories. Each category was a bottom or top quartile in a particular behavior (the middle two quartiles were ignored). There were therefore fourteen categories in total, with each student falling into between zero and seven categories. Inclusion in one category did not preclude students from falling into another. Only eleven students (four in Physics 181, seven in Physics 151) fell into no category at all. The average number of categories per student was about 3.5.

The advantage of this approach was that each category had the same number of students in it: 25% of the course. This made statistical comparison between the homework types very easy, and the increased number of students per category narrowed down standard errors significantly in some cases. As before, standard errors and averages of final grade performance agreed with the results found through correlations between behavior and performance (see chart on next page). Unfortunately, again, there was little to no additional information that could be obtained through this approach.

We feel that non-exclusive homework types are still more promising than the exclusive types used in our early steps, and can provide more useful information than one can obtain through exclusive types. They still fall short of our final goal, which was to create descriptive homework types that significantly improved our understanding of student activity. After seven failed attempts at creating such categories, we have decided that such lines of inquiry may be more fruitful at a later time, when certain aspects of this study (e.g. behaviors) have been verified and there is more appropriate data available.

BIBLIOGRAPHY

Journal Articles:

S. Bonham, D. Deardorff, R. Beichner, "Comparison of student performance using web and paper-based homework in college-level physics," Journal of Research in Science Teaching, 40(10), pp 1050 - 1071 (2003)

K. Cheng, C. Crouch, "Using an online homework system enhances students' learning of physics concepts in an introductory physics course," Am J. Phys. 72 (11), pp. 1447-1453 (2004)

M. Chi, P. Feltovich, & R. Glaser, "Categorization and representation of physics problems by experts and novices." Cognitive Science, 5, pp. 121-152 (1981)

I. Clarke, T. Flaherty, S. Mottner, "Student perceptions of educational technology tools," Journal of Marketing Education 23(3) pp. 169-177 (2001)

R. Cole, J. Todd, "Effects of Web-Based Multimedia Homework with Immediate Rich Feedback on Student Learning in General Chemistry," Journal of Chemical Education, 80(11), pp. 1338-1343 (2003)

H. Cooper, J Valentine, "Using Research to Answer Practical Questions about Homework," Education Phychology, 36(3), pp. 143-153 (2001)

R. Dufresne, J. Mestre, D. M. Hart, and K. A. Rath, "The effect of web-based homework on test performance in large enrollment introductory physics courses," J. Comput. Math. Sci. Teach. 213, pp. 229 – 251 (2002)

A. Elby, "Another reason that physics students learn by rote," Phys. Educ. Res., Am. J. Phys. Suppl. 67(7), pp S52 - S57 (1999)

J. Fredricks, P. Blumenfeld, A. Paris, "School Engagement: Potential of the Concept, State of the Evidence," Review of Educational Research, 74(1), pp. 59-109 (2004)

I. Halloun, D. Hestenes, "The Initial Knowledge State of College Physics Students," Am. J. Phys. 53(11), p. 1043-1055 (1985)

S. Hauk, A. Segalla, "Student perceptions of the web-based homework program WeBWorK in moderate enrollment college algebra courses," Journal of Computers in Mathematics and Science Teaching 24(3), pp. 229-253 (2005)

D. Hestenes, M. Wells, and G. Swackhammer, "Force Concept Inventory," The Physics Teacher 30, pp 141-158 (1992)

L. Jones and D. Kane, "Student evaluation of computer-based instruction in a large university mechanics course," American Journal of Physics 62(9), pp. 832-836 (1994)

P. Kotas and J Finck, "Collaborative Learning and Other Successful Strategies for On-line Homework", Proceedings of the International Conference on Computers in Education, pp 1068 (2002)

G. Kuh, "What We're Learning about Student Engagement from NSSE: Benchmarks for Effective Educational Practices," Change, 35(2) pp. 24-32 (2003)

W. Leonard, W. Gerace, R. Dufresne, "Analysis-Based Problem Solving: Making analysis and reasoning the focus of physics instruction," UMPERG technical report 2001#12-AUG#3-v.2-23pp (2001). Published in Spanish under the title, "Resolutión de Problemas Basada en el Análisis: Hacer del análisis y del razonamiento el foco de la enseñanza de la física," Enseñanza de las Ciencias 20(3, November): 387–400 (2002).

E. Lust, P. Vuchetich, "Comparison of Students' Performance and Perceptions of a Web-based Distance Pharmacy Calculations Course to a Campus-based Course," International Journal of Pharmacy Education, 2 (1) online journal (2004)

D. Pritchard, R. Warnakulasooriya, "Data from a web-based homework tutor can predict student's final exam score," Conference Proceedings: EdMedia -World Conference on Educational Multimedia, Hypermedia & Telecommunications Vol. 2005, pp. 2523-2529. (2005)

D. Spangler, "Assessing students' beliefs about mathematics," The Mathematics Educator 3(1), pp. 19-23 (1992)

G. Stewart, "Precise Modeling of Available Student Homework Behavior", paper presented at APS/AAPT Joint Meeting, May 2-5 (1996).

E. Thomas, R. Hume, "Relationship of Homework Complexity of Quiz Scores," Frontiers in Education Conference, 1997. 27th Annual Conference. 'Teaching and Learning in an Era of Change'. Proceedings. pp. 558-561 (1997)

R. Thornton and D. Sokoloff, "Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula," Am. J. Phys 66(4), pp 338-352(1998)

R. Warnakulasooriya, D. Pritchard, "Time to completion reveals problem-solving transfer," Proceedings of the 2004 Physics Education Research Conference, pp. 205-208 (2004)

Books and Theses:

R. Khattree and D. Naik, *Multivariate Data Reduction and Discrimination with SAS Software* (J. Wiley, New York, 2000)

P. Kotas, "Homework Behavior in an Introductory Physics Course," Thesis submitted to Central Michigan University, Dept. of Physics (2000)

E. Malinowski, *Factor Analysis in Chemistry*, 3rd ed. (J. Wiley, New York, 2002).

B. Minaei-Bidgoli, "Data Mining for a Web-Based Educational System," Thesis submitted to Michigan State University, Dept. of Computer Science and Engineering (2004)

W. Press, W. Vetterling, S. Teukolsky, and B. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, New York, 1992).

B. Loh, B. Reiser, J. Radinsky, D. Edelson, L. Gomez, S. Marshall, "Developing Reflective Inquiry Practices: A Case Study of Software, the Teacher, and Students." In K. Crowley, C. Schunn, & T. Okada, (Eds.), *Designing for Science: Implications from Everyday, Classroom, and Professional Settings*. (Erlbaum, Mahwah, NJ, 2001).