Title

An Examination of the Relationship Between the Temporal and Spatial Organization of a Student's Handwritten Statics Solution and Its Correctness

Author

Van Arsdale, Timothy

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

An Examination of the Relationship Between the Temporal and Spatial Organization of a Student's Handwritten Statics Solution and Its Correctness

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Mechanical Engineering

by

Timothy Scott Van Arsdale

September 2012

Thesis Committee:

Dr. Thomas Stahovich, Chairperson
Dr. Javier Garay
Dr. Sundararajan Venkatadriagaram

The Thesis of Timothy Scott Van Arsdale is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I want to thank my advisor, Dr. Thomas Stahovich, for his guidance throughout my time at the University of California, Riverside. Without his input this project would not have been possible. I also want to thank Dr. Javier Garay and Dr. Sundararajan Venkatadriagaram for their time serving on my committee as well as the valuable assistance they provided along the way. Lastly, I want to thank my fellow lab members for providing useful suggestions throughout my research.

To my parents,

Jim and Carol,

for their support throughout my academic career.

ABSTRACT OF THE THESIS

An Examination of the Relationship Between the Temporal and Spatial Organization of a Student's Handwritten Statics Solution and Its Correctness

by

Timothy Scott Van Arsdale

Master of Science, Graduate Program in Mechanical Engineering
University of California, Riverside, September 2012
Dr. Thomas Stahovich, Chairperson

The purpose of this project is to understand how the organization of a student's solution to a problem relates to the correctness of that work. Understanding this relationship will enable software to provide early warnings and targeted feedback to students who are struggling in a course. In this study, students in an undergraduate statics course completed their work, including homework, quizzes, and exams, using Livescribe™ Smartpens. These devices record the handwritten solutions as time-stamped pen strokes, enabling the examination of not only the final ink on the page, but also the order in which it was written. This unique database of student work was used to examine how the history of the solution construction process correlates with the correctness of the work. Solution histories were characterized by a number of quantitative features describing the temporal and spatial organization of the work. For example, some features describe the order in which various problem-solving activities, such as the construction of free body diagrams and equilibrium equations, are performed and others describe the amount of time spent on each activity. The

spatial organization of the work is characterized by the extent to which a student revisits earlier parts of a solution to revise his/her work. Cross-validated regression models were constructed using the relaxed lasso method to determine the correlation between these features and student performance. On average, the models explained 43% of the variance in performance. This is a surprising result in that the features do not actually consider the semantic content of the writing. The relaxed lasso method also identified which features were most predictive of problem correctness, thus giving insights into which student behaviors are indicative of high or low performance. For example, revising work long after it was written indicates low performance. While our work has focused on engineering statics, we expect that these techniques will generalize to other domains for which problem solutions include both diagrams and equations.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Motivation

The high enrollment in many undergraduate engineering courses often makes manually grading every homework assignment prohibitively time consuming for instructors. However, when instructors do not grade homework assignments, students may not have sufficient incentive to complete their homework assignments diligently. As a compromise, instructors use tactics such as grading only a subset of problems on each homework assignment, providing grades for homework assignments solely based on completion, or selecting one question from each assignment for use as a quiz problem. These strategies reduce the work load for the instructor, but they severely limit the feedback that students receive.

As a remedy, we envision a *digital work cycle* in which student coursework is recorded electronically, uploaded to a server, and automatically analyzed to estimate student performance. This analysis would both provide the instructor with an assessment of the students' performance and would provide feedback to the students, all without the time-consuming task of manually inspecting the work. In this thesis, we explore the development of techniques for automatically estimating a student's performance on a problem from a digital record of the work.

1.2 Overview

In this project, we seek to understand how the organization of a student's solution to a problem relates to the correctness of the work. More precisely, we seek to understand how the history of the solution construction process correlates with the correctness of the work. Understanding this relationship will enable us to create software to provide early warnings to students who may be struggling in a course.

We have conducted a study in which students in an undergraduate statics course completed all of their work, including homework, quizzes, and exams, using Livescribe™ Smartpens. These devices record the solutions as time-stamped pen strokes, enabling us to see not only the final ink on the page, but also the order in which it was written. While previous studies have used video cameras to record problem-solving activities, the analysis of such data is a difficult and time-consuming task that requires human judgment[1]. Capturing the work as time-stamped pen strokes enables a much more precise and efficient analysis of the work.

We seek to understand the relationship between how students construct their solutions and their performance on those problems. We refer to the sequence of problem-solving steps as a *solution history*. We characterize solution histories with a number of quantitative features describing the temporal and spatial distribution of the work. For example, there are features that describe the order in which various problem-solving activities (such as the construction of free body diagrams and equilibrium equations) are performed, and the amount of time spent on each activity. Because Smartpens use ink, students cannot erase their errors and must cross them out. We characterize cross-outs by the delay between when ink was written and when it was crossed out. The spatial organization of the work is characterized by the extent to which a student revisits earlier parts of a solution to revise the work. We then construct cross-validated regression models to determine the extent to which these features correlate with the correctness of the solution. In the examples we considered, on average, about 43% of the variance in performance could be explained by these features. This is a surprising result in that the features do not consider the semantic content of the writing.

1.3 Related Work

Our work is a form of educational data mining: a research discipline that uses machine learning techniques, data mining techniques, and other similar techniques to examine education research issues. Romero and Ventura[2] provide a recent overview of work in this area. Much of this work relies on data collected in online environments such as web applications and intelligent tutoring systems. Our work is unique in that we use digital records of students' handwritten solutions, enabling us to study work habits in a more natural environment. The work of Oviatt *et al.*[3] suggests that natural work environments are critical to student performance. In their examinations of computer interfaces for completing geometry problems, they found that "as the interfaces departed more from familiar work practice..., students would experience greater cognitive load such that performance would deteriorate in speed, attentional focus, meta-cognitive control, correctness of problem solutions, and memory."

There have been several studies examining student work habits and performance in statics. For example, Steif and Dollár[4] examined usage patterns of a web-based statics tutoring system to determine the effects on learning. They found that learning gains increased with the number of tutorial elements completed. This study relied on an online learning environment, while we consider ordinary handwritten work. In another study, Steif *et al.*[5] examined whether students can be induced to talk and think about the bodies in a statics problem, and if doing so can increase a student's performance. They used tablet PCs to record the students' spoken explanations and capture their handwritten solutions as time-stamped pen strokes. The study focused on the spoken explanations, with the record of written work left mostly unanalyzed.

Researchers have used video recordings to examine student problem solving. For example, Blanc[6] examined video recordings of student work in mathematics and analyzed the path that students used to solve an example problem. Although Blanc recorded more than 75 problem solutions, only two were analyzed in his paper. This speaks to the labor-intensive nature of analyzing video records. Our pen stroke data is more amenable to automated analysis.

Other researchers have used journaling to examine student work habits. For example, Orr *et al.*[7] examined students' journal responses about their study habits, including factors such as when and how they completed their homework, and if they took advantage of assistance programs. While the results proved interesting, journals capture students' perceptions of their work habits rather than an objective characterization of them. Our work provides a nice complement to this work as we capture a detailed time-stamped record of a student's work over the duration of the course.

The ultimate goal of our work is to rapidly and inexpensively identify students who may be struggling in a course so that extra assistance can be provided. Other researchers have explored various mechanisms for providing rapid feedback. For example, Rasila *et al.*[8] explored the benefits of an online assessment tool for engineering mathematics. They found that automatic assessment was highly useful and improved the feedback provided to students. Chen *et al.*[9] used electronic conceptual quizzes during lectures within a statics course to help guide the lecture content. They found that the rapid feedback produced a significant increase in student performance.

Chapter 2. Data Set

We conducted a large-scale study in which over 120 students from an undergraduate mechanical engineering course in statics were given Livescribe^{TM} Smartpens. In addition to serving the same function as a traditional ink pen, these devices digitize the pen strokes and store them as sequences of time-stamped coordinates. Figure 1 provides an example of pen-stroke data from a Smartpen. Starting in the third week of the quarter, students were asked to complete all coursework using Smartpens. This included seven homework assignments with 44 problems in total, six quizzes each with a single problem, and three exams with a total of 13 problems. The resulting database contains over four million digitized pen strokes.

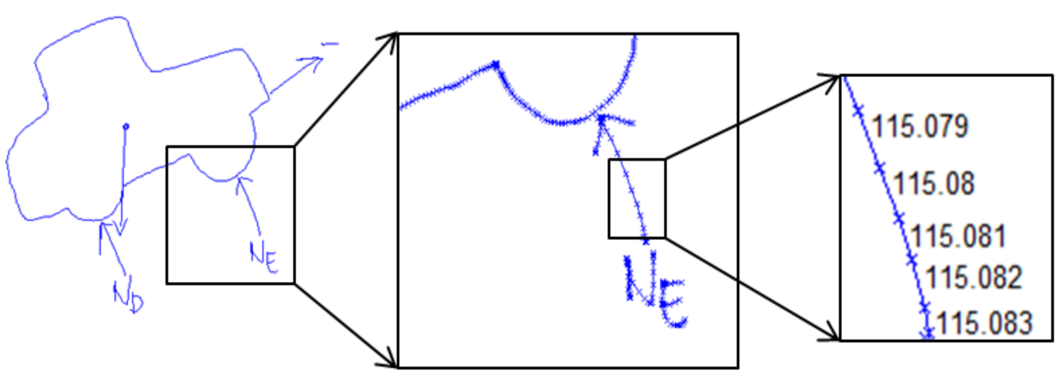

Figure 1: A typical example of digitized pen stroke data. Left – a free body diagram rendered from pen stroke data. Center – a selection of the diagram showing data points. Right – a smaller selection of data points including their time stamps (in seconds elapsed from the start of the problem).

We restrict our present analysis to problems from the two midterm exams and the final exam, because detailed grading information was available for these problems. Also, the data collected for exam problems comprises the students' entire solution effort. By contrast, grading information was available for only a subset of the homework; only one problem was graded for each assignment. Also, some students began their homework on scratch paper, and then constructed a neat final solution with the digital pen, so that only a portion of the effort was recorded. (Despite these issues, homework data is still useful for understanding students' performance in the course, but that work is beyond the scope of this project.) While we do have detailed grading information and a complete record of the solution effort for quizzes, we do not consider them here because the quiz problems required only brief solutions.

The midterm and final exams comprise a total of 13 problems (Table 1), each with between 97 and 122 student solutions. The variation in the number of students completing each problem is due to a variety of causes; some students forgot to bring their pens to the exam, some students did not complete all exam questions, and some students dropped the course. The problems were graded by teaching assistants based on rubrics developed by the course instructor. These rubrics assigned credit for the correctness of individual problem-solving steps as well as the overall correctness of the solution.

Our analysis of problem-solving activities requires knowledge of the kind of solution element each pen stroke comprises. Specifically, each pen stroke is categorized as an element of a free body diagram, an element of an equation, or a cross-out (i.e., a stroke used to cross-out work). We use an automatic stroke-labeling system developed by Lin[10] to determine the category of each pen stroke. In his experiments, Lin's system achieved 93% accuracy at this task.

Exam and Problem Number	Problem Type	Number of Unknowns	Number of FBDs	Number of Equilibrium Equations
Midterm 1 Problem 1	2-D Single Body	2	1	3
Midterm 1 Problem 2	2-D Single Body	3	1	3
Midterm 1 Problem 3	3-D Single Body	3	1	6
Midterm 2 Problem 1	2-D Machine	4	3	6
Midterm 2 Problem 2	2-D Machine	4	4	10
Midterm 2 Problem 3	2-D Truss	3	3	5
Final Exam Problem 1	2-D Single Body	3	1	3
Final Exam Problem 2	3-D Single Body	3	1	6
Final Exam Problem 3	2-D Machine	1	2	3
Final Exam Problem 4	2-D Truss	4	3	6
Final Exam Problem 5	2-D Machine	4	4	10
Final Exam Problem 6	2-D Machine with Belt Slip	1	2	2
Final Exam Problem 7	2-D Centroid	2	N/A	N/A

Table 1: Summary of exam problems analyzed. Column two contains the problem type; all except the last are equilibrium problems. Column three contains the number of unknown values (typically forces) that the student was asked to determine. Columns four and five are the numbers of free body diagrams (FBDs) and equilibrium equations in the instructor's solution to the problem.

Chapter 3. Manual Investigation of Solutions

We manually inspected student solutions to a single final exam problem to identify which aspects of solution histories were potentially correlated with performance. Initially, we printed the solutions on paper and grouped them based on their assigned score. Any differences between the solutions in different score groups were identified for possible use as features. For example, many solutions in the highest scoring group were concise. This analysis considered only the final ink on the page.

To examine the temporal properties of the work, we initially considered viewing the solutions by replaying them like a movie. However, this was very time consuming and thus impractical. Instead, we used color-coding to represent the dynamic properties of the solution. The pen strokes in each solution were color-coded according to the time at which they were written. A sequence of 64 colors was used to represent time, with the first stroke rendered with the first color in the sequence and the last stroke rendered with the last color. Figure 2 shows an example of a color-coded solution. We printed the color-coded solutions, allowing us to directly compare multiple solutions to each other.

We also represented the work abstractly as a sequence of color-coded activities rendered on a timeline similar to the representation in Figure 3. We constructed two versions of the timeline, one using absolute time and one using normalized time. For the latter, time was scaled so that the entire solution process took one unit of time. The former representation facilitates examination of the amount of effort spent on a problem. The latter facilitates comparison of the sequencing of the activities between different problems. For example, some students constructed all their free body diagrams before writing any equilibrium equations, while others switched between these activities throughout their solution.

The insights we gained from this manual inspection of solutions inspired the features described in Chapter 4. For example, the “Out-of-Order” features described in Section 4.2 were directly inspired by our examination of color-coded renderings like that in Figure 2.

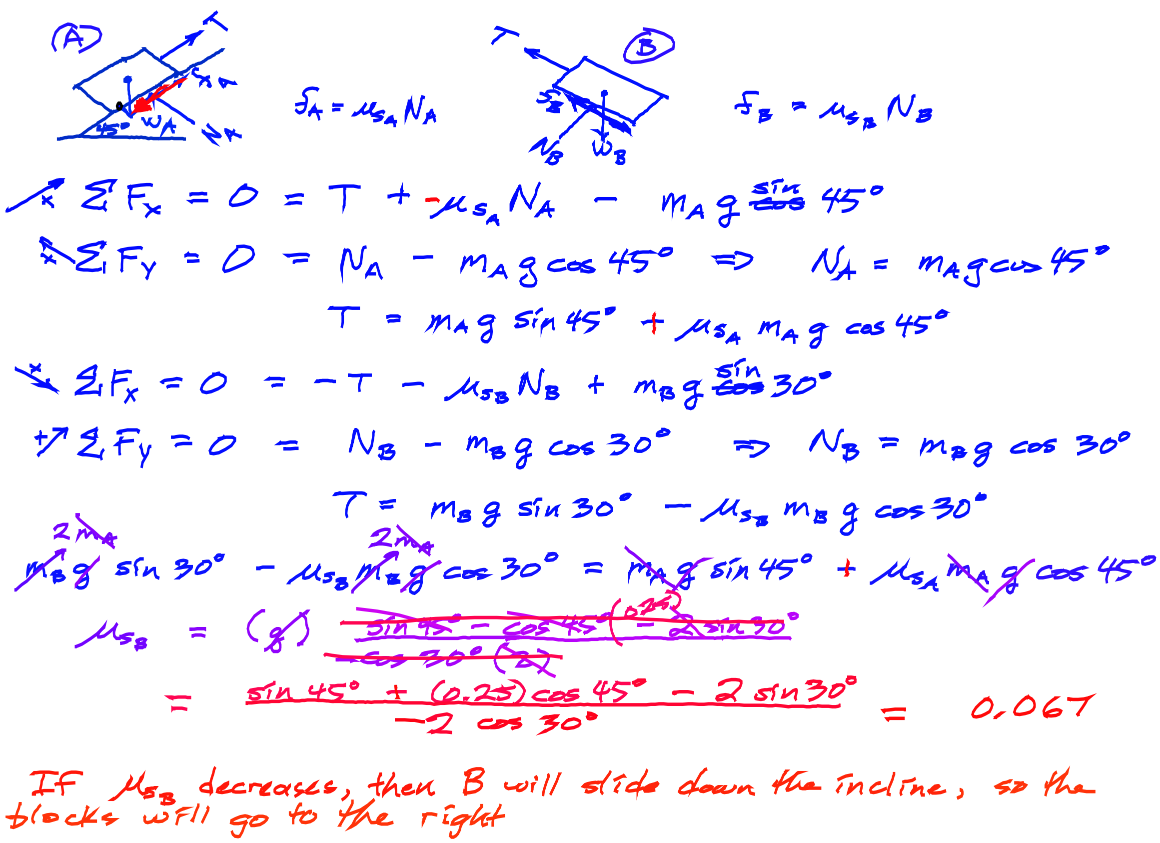

Figure 2: An example problem solution with the pen strokes color-coded to indicate the time at which they were written. Color progresses from blue (first stroke) to orange (last stroke).

Chapter 4. Characterization of Solution Histories

To examine the correlation between the properties of the solution histories and the correctness of the work, we first represent those properties quantitatively. We characterize a solution history in terms of the temporal and spatial distribution of the work. More specifically, we consider five types of features: properties of the temporal organization of the work, properties of the spatial organization of the work, properties of the spatial clustering of the work, properties of the cross-outs, and basic pen stroke properties. These features are described in detail in the following sections.

4.1 Temporal Organization Features

In characterizing the temporal distribution of the work in a solution history, we distinguish between four solution activities: drawing free body diagrams (FBDs), constructing and solving equilibrium equations, crossing out work, and working on other problems. The first three activities are inferred from the stroke labels described in Chapter 2.

To represent the sequence of solution activities, we divide the problem solution into n equal-time intervals. Each interval is labeled according to the solution activity that occurs most frequently during that interval, which is computed using the pen stroke labels. For example, if 70% of the drawing time in an interval was spent drawing free body diagram pen strokes, and the remaining time was spent drawing equation pen strokes, the interval as a whole would be characterized by the free body diagram activity. If no writing occurs during an interval, it is labeled as a break. In practice, we have found that using a value of 400 for n provides adequate detail to enable meaningful analysis of the solution. One advantage of this representation is that it abstracts away the total elapsed time, making it possible to directly compare the work of all students regardless of their total solution time.

If the student interrupts his or her work on a problem to work on other problems, we modify this representation slightly. If there are m such interruptions, we divide the work on the problem in question into $n - m$ equal intervals and compute their labels as before. Each of the m interruptions is then represented by an additional interval labeled as "other problem." Figure 3 shows a portion of a typical activity sequence.

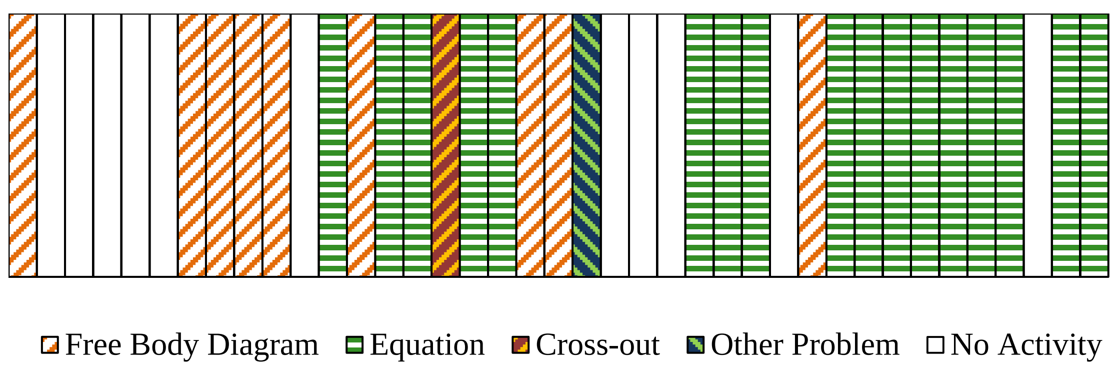

Figure 3: A portion of a typical discretized activity sequence.

The distribution of activities in the discretized solution history gives important insights into the student's thought process. We have designed a set of eight features to capture these insights. These features are summarized in Table 2. The first four features describe the amount of time spent on various activities. *FBD Effort* is the total number of activity intervals spent on free body diagrams, while *EQN Effort* is the number spent on equations. The *Break* feature is the number of intervals in which no work was done, while the *Other-Problem* feature is the number of times the student interrupted his/her work on the problem to work on other problems (this is the value "*m*" described above).

Taking breaks and working on other problems may indicate that the student was struggling on the current problem.

Temporal Organization Features	*FBD Effort*	Number of activity intervals spent on FBD activity.
	EQN Effort	Number of activity intervals spent on equation activity.
	Break	Number of activity intervals in which a student had no activity.
	Other-Problem	Number of times the student interrupted their work on a problem to work on other problems.
	Entropy	Entropy of the discretized activity sequence.
	Complexity	Complexity of the discretized activity sequence.
	FBD to Equation Activity Change	Number of activity changes from FBDs to equations.
	Equation to FBD Activity Change	Number of activity changes from equations to FBDs.
	Num Small Breaks	Number of breaks between 2 and 40 seconds in duration.
	Num Medium Breaks	Number of breaks between 40 and 160 seconds in duration.
	Num Large Breaks	Number of breaks at least 160 seconds in duration.

Table 2: Summary of the Temporal Organization features.

These first four features describe only the amount of effort spent on each type of activity. Four additional features describe the sequencing of the activities. An expert might solve a problem by first constructing all of the free body diagrams, and then constructing all of the equations. This would result in a very simple activity distribution. A novice student who is struggling on a problem might repeatedly move from one activity to another in a much more complex pattern. We use information theory notions of complexity and entropy to capture these distinctions.

The Kolmogorov complexity[11] of a sequence is a measure of the minimum length required to describe it. To estimate this value, we first represent the sequence as a character string, assigning a unique letter to each of the four types of activities. We then use a standard data compression algorithm (the ZLIB[12] implementation of DEFLATE[13]) to compress the string. We define the *Complexity* of the sequence as the length of the compressed string. A random sequence of activities will result in a large value for this feature, while a sequence comprised of a few large blocks of activities will result in a small value.

We use the *Entropy* of the sequence to measure the balance of effort between the activities. If the sequence contains, for example, only one type of activity, the entropy is relatively small. If, on the other hand, an equal amount of time is spent on each of the two types of activities, the entropy is maximal. We compute the *Entropy* using the usual approach:

$$Entropy = \sum_i -(n_i/n)\, ln(n_i/n)$$

where n_i is the number of occurrences of a particular type of activity, n is the total number of activities, and the sum is taken over the two main types of activities. (In this computation, we assume $ln(0) = 0$.)

Two additional features consider transitions between free body diagram activity and equation activity. The number of transitions from the former to the latter is represented by the *FBD to Equation Activity Change* feature, while the converse is represented by the *Equation to FBD Activity Change* feature. These features are calculated from the discretized activity sequence with the cross-out, break, and "other problem" intervals removed. Free body diagrams are a tool for constructing equilibrium equations and thus the former often precede the latter. These two activity change features are useful for detecting if students perform these tasks sequentially or if they iterate between them, for example.

The *Break* feature provides a measure of the total fraction of the activity sequence during which the student was not working on any solution activity. Three additional features characterize the size distribution of the individual periods of non-activity. More specifically, these features count the *Num Small Breaks* (breaks between 2 and 40 seconds in duration), the *Num Medium Breaks* (breaks between 40 and 160 seconds in duration), and the *Num Large Breaks* (breaks at least 160 seconds in duration). These features are computed directly from the original timeline of the solution history, not from the normalized discrete activity sequence.

4.2 Spatial Organization Features

The spatial organization of a solution on the page gives additional insights about the student's problem-solving process. For example, a student who starts at the top of a page and progresses downward may understand the problem better than a student who frequently revisits earlier work and revises it. We describe the spatial organization with two types of features (Table 3) that consider the progression of the work on the page and the local temporal history in the neighborhood of each stroke.

Spatial Organization Features	*Out-of-Order-10-20*	Fraction of strokes that differ from their reference time by 10% to 20% of the total problem time.
	Out-of-Order-20-30	Fraction that differ by 20%-30%.
	Out-of-Order-30-40	Fraction that differ by 30%-40%.
	Out-of-Order-40-50	Fraction that differ by 40%-50%.
	Out-of-Order-50-60	Fraction that differ by 50%-60%.
	Out-of-Order-60+	Fraction that differ by over 60%.
	Earlier-Neighbor-10-20	Fraction of strokes that have a delay from neighboring strokes of 10% to 20% of the total problem time.
	Earlier-Neighbor-20-30	Fraction that have a delay of 20%-30%.
	Earlier-Neighbor-30-40	Fraction that have a delay of 30%-40%.
	Earlier-Neighbor-40-50	Fraction that have a delay of 40%-50%.
	Earlier-Neighbor-50-60	Fraction that have a delay of 50%-60%.
	Earlier-Neighbor-60+	Fraction that have a delay over 60%.

Table 3: Summary of the Spatial Organization features.

We describe progression down the page in terms of deviation from a reference progression in which each stroke is drawn later than the ones above it. We use a two-inch-tall sliding window to construct this reference timeline as illustrated in Figure 4. The height of the window was chosen based on the inspection of the resulting timelines. The window is initially placed at the top of the work. The reference time assigned to the location of the top of the window is computed as the time of the earliest stroke in the window. The center point of a stroke's bounding box is used to determine if the stroke is in the window. In the example in Figure 4, the pen stroke for the letter "P" in "problem" determines the first location of the window and the reference time assigned is that of the earliest stroke in the window. The window is then slid down the page a small distance. The reference time assigned to the new location of the top of the window is again that of the earliest stroke in the window, unless that is earlier than the time assigned to the previous window. In that case, the reference time is taken to be that of the previous window. The process is repeated every inch until the bottom of the solution is reached, resulting in a sequence of monotonically increasing reference time values, equally spaced down the solution page. The distance between the window positions was chosen based on the inspection of the resulting timelines. If a solution spans multiple pages, the pages are ordered by the average stroke time on each page and are stacked vertically, with a 0.5 in. gap between each. This results in a single progression of work for each problem solution.

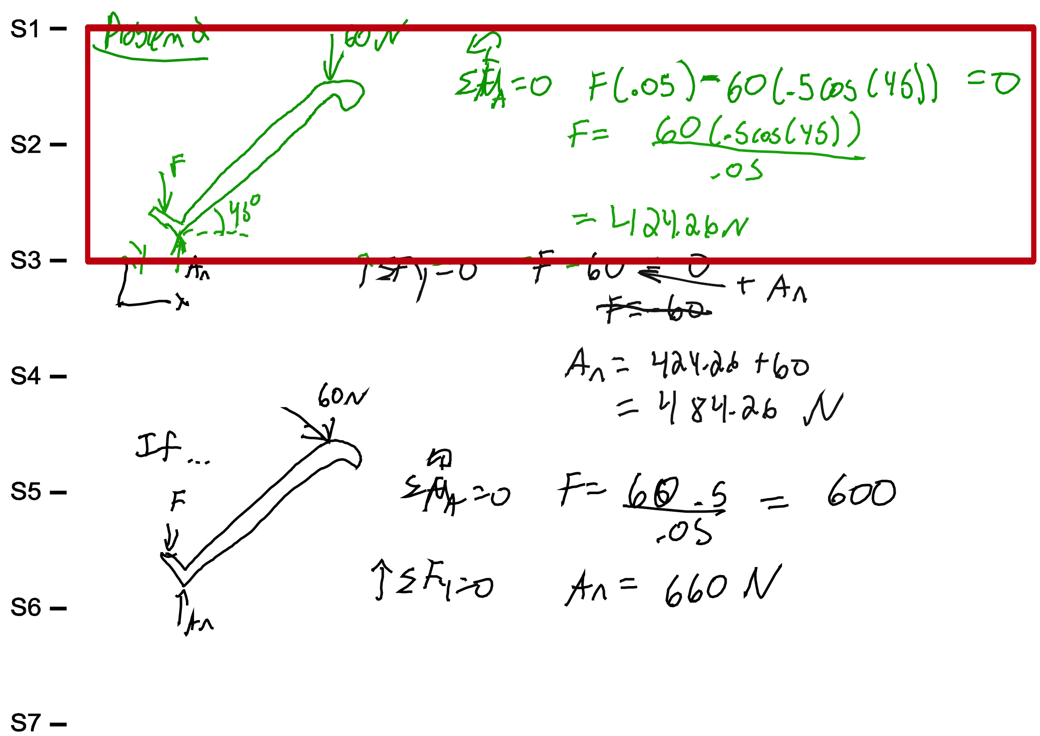

Figure 4: A sliding window (red box) is used to compute a reference timeline. The time stamp of the earliest stroke in the window is assigned to the location of the top of the window. Strokes inside the window are shown in green. Sample locations are indicated by S1 through S7 on the left-hand side of the figure.

Once the reference timeline has been constructed, it is used to identify strokes that are inconsistent with a top-down spatial progression, which are called “out-of-order” strokes. To do this, we compute the reference time for each stroke’s location (its midpoint) by linear interpolation of the reference timeline. If the time at which a pen stroke was drawn differs from this reference by at least 10% of the total solution time, the stroke is considered to be out-of-order. Figure 5 shows the out-of-order strokes from Figure 4. Six features are used to further characterize the out-of-order strokes by the extent to which they differ from their reference time as described in Table 3. For example, *Out-of-Order-10-20* is the fraction of strokes that differ from the reference time by between 10% and 20% of the total solution time, while *Out-of-Order-60+* is the fraction of strokes that differ by 60% or more.

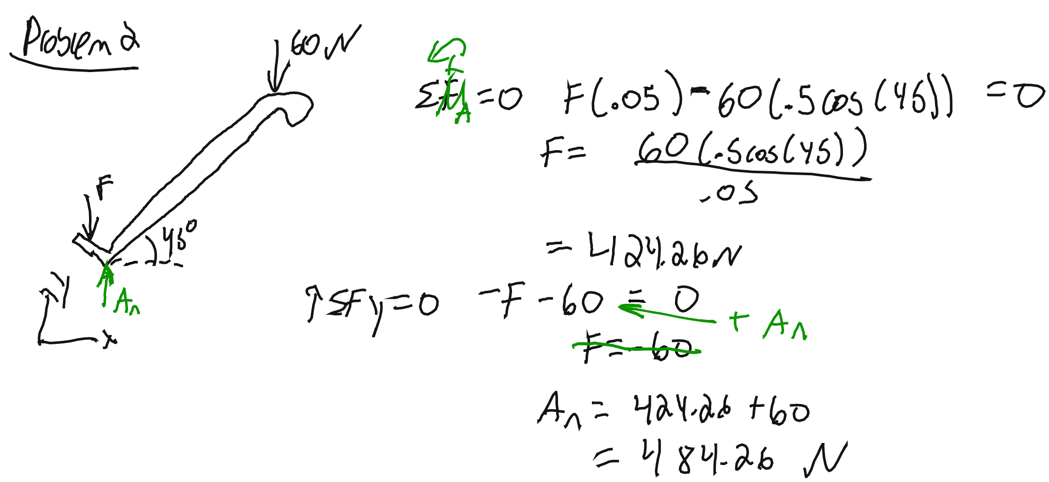

Figure 5: Hypothetical example of out-of-order work. Out-of-order strokes are shown in green. In this example, the student revised the free body diagram by adding an additional force after beginning the equilibrium equations.

The reference timeline provides a global view of the progression of work. A second type of feature provides a more local view of the progression by comparing the time stamp of a stroke to those of the nearby strokes that were drawn earlier. Two strokes are considered to be near each other if their expanded bounding boxes intersect (the strokes may actually intersect each other). For this calculation, the coordinate-aligned bounding boxes of the strokes are expanded in all directions by 0.8 in.; this value was obtained with the optimization procedure described in Section 4.6. Each stroke is then characterized by the time delay between it and its earliest nearby stroke. Analogous to the Out-of-Order features, six features are used to characterize this time delay as described in Table 3. For example, *Earlier-Neighbor-10-20* is the fraction of strokes with a delay between 10% and 20% of the total solution time. Strokes with a large delay may correspond to the student revising his or her work after an error is detected much later in the solution. This could occur, for example, if the student detects an inconsistency in the equilibrium equations and must revisit the free body diagram to fix the error. Students who frequently revisit earlier portions of their solution may be struggling with the concepts.

4.3 Spatial Cluster Features

Typical statics solutions are often organized into spatially distinct clusters of work. Each individual cluster typically represents a single substantial solution element, such as a free body diagram or a set of equilibrium equations. Figure 7 shows an example with seven clusters: four containing free body diagrams and three containing equations.

We compute several features that describe the spatial clustering of the work and the extent to which the work in the clusters is revised during the solution process. In computing these features, we define a cluster as a region on the page containing strokes that represent a single solution activity (either free body diagram or equation activity), are near each other, and are distant from other strokes of the same activity.

To compute the locations of clusters, we use a Gaussian function to represent the "spatial influence" of each stroke. More specifically, at each point on the page, we define an influence function that sums the signed influences of the strokes. Strokes from free body diagrams exert a positive influence, while those from equations exert a negative influence. This influence function, which is illustrated in Figure 6, is computed as:

$$H(x,y) = \sum_i A * S e^{-\frac{{d_i}^2}{2c^2}}$$

$$A = \begin{cases} 1 \text{ for FBD strokes} \\ -1 \text{ for equations strokes} \end{cases}$$

$$S = maximum\{1.5\ in., 0.75\ in. + \frac{l_i}{3}\}$$

Here, d_i is the minimum distance from stroke *i* to the point *(x,y)*, *c* is a constant equal to 1.4 in., and l_i is the length of the stroke (i.e., its arc length). The parameter *S* controls the maximum amplitude of the influence of a stroke, which increases with the stroke's length. However, the constant in the denominator (i.e., the "3") ensures that very long strokes do not dominate the calculation. Conversely, *S* has a minimum value (1.5 in) to ensure that even very short strokes have an appreciable maximum amplitude.

Because of the exponential nature of *H(x,y)*, strokes far from the point *(x,y)* exert little influence on that point. Thus, to achieve efficiency, the sum is taken over only those strokes that are near the point *(x,y)*. A stroke is considered to be near if its coordinate-aligned bounding box is within 1.1 in. of the point.

We then compute the cluster boundaries as level curves of the function $H(x,y)$. The boundaries of the free body diagram clusters are defined as level curves at $H(x,y) = 0.2$, while the boundaries of the equation clusters are defined as level curves at $H(x,y) = -0.2$. (When computing the level curves, $H(x,y)$ is sampled on a uniform grid with a spacing of 0.3 in.) Using values of +/- 0.2 for the level curves tends to place the cluster boundaries near the periphery of the ink they enclose. By contrast, if the level curves were taken at $H(x,y) = 0$, all regions of the page, even empty regions, would belong to some cluster. The parameters used for computing clusters were manually tuned so that the clusters closely matched the major solution elements for a set of sample sketches.

Figure 7 shows the set of clusters computed from the influence function in Figure 6. There are four clusters each representing a single, isolated free body diagram. There are three equation clusters. The top two each represent tight groupings of equations. The bottom cluster has a “C” shape. This cluster appears to contain two groupings of equations that are linked by a few pen strokes to the left of the lowest free body diagram cluster. Using a larger threshold for the level curves might split this cluster appropriately. However, this might also split portions off of the other clusters, such as the top of the uppermost equation cluster.

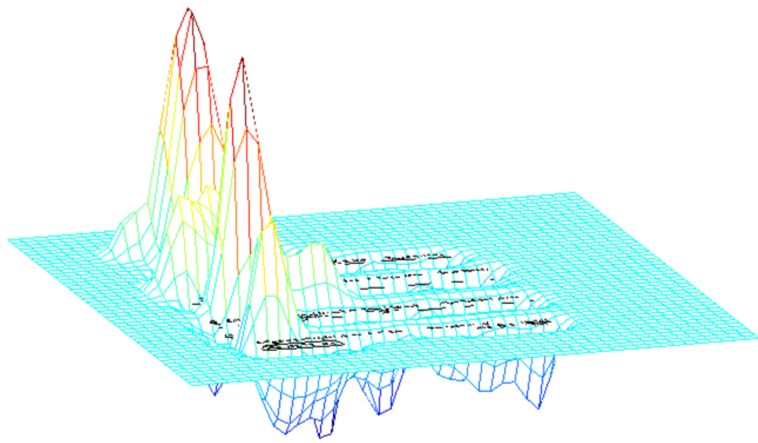

Figure 6: A typical influence function for computing clusters. Portions of the surface above the plane are most strongly influenced by free body diagram strokes, while regions below the plane are most strongly influenced by equation strokes. The black lines on the plane are individual pen strokes which are shown for reference.

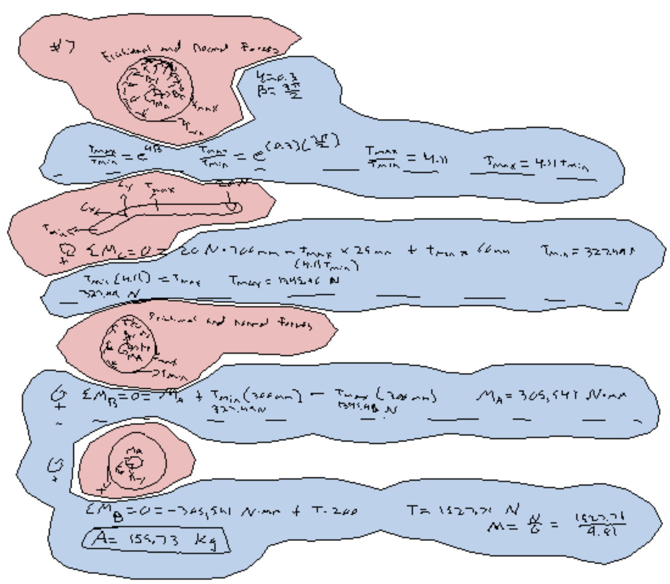

Figure 7: Clusters computed from the influence function in Figure 6. Red regions represent free body diagram clusters; blue regions represent equation clusters.

From these clusters we compute seven features, which are summarized in Table 4. Three characterize the number and size of the clusters. This includes the number of free body diagram clusters (*Num FBD Clusters*), the number of equation clusters (*Num

Equation Clusters), and the ratio of the net area of the equation clusters to the total area of all clusters (*Equation Area Fraction*).

The remaining four cluster features describe the student's temporal progression through the clusters. *FBD Revisits* is the number of times the student interrupted his or her work to add additional pen strokes to an existing free body diagram cluster. *FBD Revisit Strokes* is the fraction of all pen strokes that were added to free body diagram clusters in this way. *Equation Revisits and Equation Revisit Strokes* are defined analogously.

Spatial Cluster Features	*Num FBD Clusters*	Number of FBD pen stroke clusters.
	FBD Revisits	Number of times a student returned to a previous FBD cluster.
	FBD Revisit Strokes	Fraction of strokes in a solution that were added during FBD revisits.
	Num Equation Clusters	Number of equation pen stroke clusters.
	Equation Area Fraction	Ratio of the net area of the equation clusters to the total area of all clusters.
	Equation Revisits	Number of times a student returned to a previous equation cluster.
	Equation Revisit Strokes	Fraction of strokes in a solution that were added during equation revisits.

Table 4: Summary of the Spatial Cluster features.

4.4 Cross-out Features

Cross-outs are a direct indication of revised work. We characterize cross-outs in terms of the strokes that are deleted or "crossed out." The stroke labeler described in Chapter 2 identifies individual cross-out strokes, but not complete cross-out gestures. For example, an "X" drawn with two pen strokes is often used to cross-out erroneous work. We define a cross-out gesture as a set of consecutively drawn cross-out strokes that are all near each other. Cross-out strokes are near each other if the minimum distance between them is less than 1 in. or 15% of the stroke's arc length, whichever is smaller. These values were manually selected to achieve a balance between grouping the components of an intended gesture without erroneously grouping strokes from unrelated gestures.

To determine which strokes have been deleted by a cross-out gesture, we compute the convex hull of the strokes comprising that gesture. Any other pen strokes which have a convex hull that intersect the convex hull of a gesture are considered to have been deleted. Figure 8 shows the convex hulls of a zigzag-shaped cross-out gesture and a "X" cross-out gesture as well as the strokes that they delete.

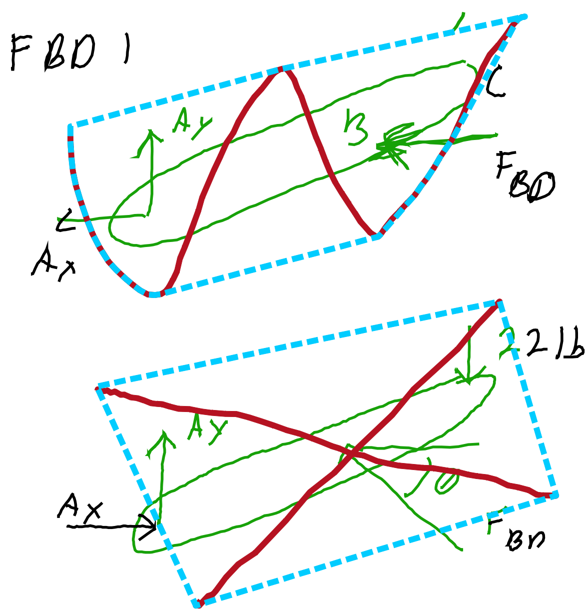

Figure 8: Strokes deleted by cross-out gestures. The cross-out gestures are shown in red, the convex hulls in dashed blue, and the deleted strokes in green.

We distinguish between two kinds of cross-out gestures, which we call "typo cross-outs" and "problem-solving cross-outs". The former are cases in which the student writes something and quickly crosses it out, as if correcting a typographical error. The latter are cases in which there is a substantial delay between the time the ink was written and when it was crossed out: these cases are more likely to be corrections of problem-solving errors. We use a threshold of 16 seconds as the boundary between the two types of cross-outs. This threshold was based on the optimization procedure described in Section 4.6.

We characterize cross-out gestures with five features which are summarized in Table 5. The *Typo-Cross-Outs* and *PS-Cross-Outs* features are the numbers of typo and problem-solving cross-out gestures, respectively. The *Big-Cross-Outs* feature is the number of cross-out gestures that delete (cover) 10 or more pen strokes and thus represents a revision of a substantial amount of work. (This threshold was also set based on the optimization procedure described in Section 4.6.) Additionally, we count the total number of free body diagram and equation strokes that were deleted by cross-out gestures producing the features *FBD Strokes Crossed-Out* and *Equation Strokes Crossed-Out*, respectively.

Cross-out Features	*FBD Strokes Crossed-Out*	Number of FBD strokes that were crossed-out.
	Equation Strokes Crossed-Out	Number of equation strokes that were crossed-out.
	Big-Cross-Outs	Number of cross-out gestures which removed 10 or more strokes.
	Typo-Cross-Outs	Number of cross-out gestures which occurred within 16 seconds of underlying ink.
	PS-Cross-Outs	Number of cross-out gestures which occurred after 16 seconds of underlying ink.

Table 5: Summary of the Cross-out features.

4.5 Basic Pen Stroke Features

We include six Basic Pen Stroke features in order to provide a measure of the amount of work in a solution and the student's writing style. These are summarized in Table 6. These features include the number of strokes written for each activity category (*Num FBD Strokes*, *Num of Equation Strokes, and Num Cross-Out Strokes*), as well as the median stroke length for each category (*Median FBD Stroke Length, Median Equation Stroke Length, and Median Cross-Out Stroke Length*).

Basic Pen Stroke Features	*Median FBD Stroke Length*	Median length of FBD strokes in the problem solution.
	Median Equation Stroke Length	Median length of equation strokes in the problem solution.
	Median Cross-Out Stroke Length	Median length of cross-out strokes in the problem solution.
	Num FBD Strokes	The total number of FBD strokes in the problem solution.
	Num Equation Strokes	The total number of equation strokes in the problem solution.
	Num Cross-Out Strokes	The total number of cross-out strokes in the problem solution.

Table 6: Summary of the Basic Pen Stroke features.

4.6 Selection of Feature Parameter Values

We used a simple optimization process to select parameter values for seven of the features. The parameters for related features were optimized simultaneously. Table 7 lists the sets of related features. The optimization process used search to select parameter values that maximized the predictive ability of ordinary least squares regression models.

To begin the search, we enumerated a small set of parameter values to explore. We used our experience with the features to select a reasonable default parameter value. We then enumerated smaller values by successively dividing the default value by two, and larger values by successively multiplying by two. This resulted in six values for each parameter. For example, the default value of the parameter "small break lower bound" was 8 seconds. From this, we generated values of 1, 2, 4, 8, 16, and 32 seconds. The optimization of a set of related features exhaustively explored all combinations of the enumerated parameter values. For example, the optimization of the break features would nominally explore $6^3 = 216$ combinations. If the optimum occurred at the boundary of the set of enumerated values, the set was expanded using the above method, and the search repeated.

Sets of Features	**Parameters (Selected Value)**	**Search Space Size**
Num Small Breaks *Num Medium Breaks* *Num Large Breaks*	Small break lower bound (2 seconds) Boundary between small and medium breaks (40 seconds) Boundary between medium and large breaks (160 seconds)	n^3
Earlier-Neighbor features	Bounding box expansion (0.8 in.)	n
Typo-Cross-Outs *PS-Cross-Outs*	Boundary between typo and PS cross-outs (16 seconds)	n
Big-Cross-Outs	Minimum number of strokes crossed-out (10)	n

Table 7: Sets of features, the parameters that were chosen to maximize the features' combined predictive ability, and the size of the resulting search spaces ($n \sim 6$).

Chapter 5. Modeling Performance using Solution Histories

We used linear regression to evaluate the ability of our set of 41 features to predict student performance on an exam problem. We also used regression to determine which of the five subsets of features (Temporal Organization, Spatial Organization, Spatial Cluster, Cross-out, and Basic Pen Stroke) and which individual features are most predictive. To prevent over-fitting, we use the relaxed lasso[14,15] regression technique as it performs feature selection. Relaxed lasso is an extension of the lasso[16] method, a regularized version of linear least-squares regression in which the weighted sum of the absolute value of the coefficients is added as a penalty. (We also considered the elastic net[17] method, but the models were not as predictive as those from relaxed lasso.) More specifically, the penalty has the form $\lambda \sum_{j=1}^{p} |\beta_j|$ where λ is a tuning parameter and β_j are the coefficients in the regression model. This penalty term helps to eliminate unimportant features from the model. The relaxed lasso method is a two-step version of lasso. An initial set of lasso models are constructed, any features not selected in the initial models are removed, and the models are retrained with the penalty terms relaxed. We implemented the relaxed lasso technique using the lasso function in the Matlab® Statistics Toolbox™. When constructing lasso models, we use 10-fold cross-validation and repeat the cross-validation 10 times using random seeds for the splits.

Additionally, we use stepwise linear regression to identify which features have statistically significant predictive ability. For this analysis, we use the stepwise linear regression function in IBM® SPSS® Statistics version 20. The threshold for including a feature in the model was a p-value less than 0.05 (based on the F value), while the threshold for removal was a p-value greater than 0.10.

To provide a basis for interpreting the regression results, descriptive statistics of the features, averaged over the problems for each exam, are reported in Table 8. Descriptive statistics for each individual exam problem are contained in Appendix A.

Features	Midterm 1		Midterm 2		Final Exam	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
FBD Effort	50.42	18.51	60.36	20.41	48.67	20.91
Equation Effort	95.87	37.50	104.46	39.33	99.62	39.77
Break	245.33	41.93	226.53	43.71	244.15	48.12
Other-Problem	0.96	1.29	0.85	0.92	0.84	1.00
Entropy	0.61	0.08	0.61	0.10	0.59	0.11
Complexity	99.71	11.69	104.72	13.19	98.31	14.99
FBD to Equation Activity Change	7.59	4.78	10.18	5.64	8.64	5.49
Equation to FBD Activity Change	7.10	4.88	9.83	5.62	8.21	5.54
Num Small Breaks	84.76	32.43	98.20	39.73	79.65	38.23
Num Medium Breaks	5.46	3.10	4.66	2.98	4.45	3.24
Num Large Breaks	0.52	0.87	0.35	0.62	0.59	0.94
Out-of-Order-10-20	0.19	0.09	0.21	0.11	0.19	0.10
Out-of-Order-20-30	0.10	0.08	0.11	0.09	0.09	0.08
Out-of-Order-30-40	0.07	0.07	0.06	0.07	0.05	0.06
Out-of-Order-40-50	0.04	0.06	0.04	0.05	0.03	0.06
Out-of-Order-50-60	0.03	0.06	0.02	0.05	0.02	0.04
Out-of-Order-60+	0.05	0.09	0.04	0.09	0.04	0.09
Earlier-Neighbor-10-20	0.24	0.11	0.26	0.11	0.25	0.12
Earlier-Neighbor-20-30	0.18	0.10	0.18	0.09	0.19	0.12
Earlier-Neighbor-30-40	0.12	0.09	0.11	0.08	0.11	0.09
Earlier-Neighbor-40-50	0.08	0.07	0.06	0.06	0.07	0.07
Earlier-Neighbor-50-60	0.06	0.06	0.04	0.05	0.04	0.05
Earlier-Neighbor-60+	0.09	0.10	0.05	0.07	0.06	0.09
Num FBD Clusters	2.87	1.91	4.05	2.29	3.69	2.28
FBD Revisits	8.44	5.90	9.26	6.33	7.54	5.82
FBD Revisit Strokes	0.16	0.12	0.15	0.13	0.12	0.10
Num Equation Clusters	2.75	1.55	3.33	1.79	3.08	1.79
Equation Area Fraction	0.66	0.14	0.62	0.16	0.66	0.15
Equation Revisits	3.89	4.36	5.81	5.50	5.32	5.45
Equation Revisit Strokes	0.32	0.27	0.33	0.24	0.34	0.25
FBD Strokes Crossed-Out	23.91	27.49	25.89	26.67	19.87	24.75
Equation Strokes Crossed-Out	56.34	68.33	57.04	58.41	53.57	70.10
Big-Cross-Outs	1.82	2.03	2.19	2.41	1.78	2.08
Typo-Cross-Outs	29.24	26.33	39.81	37.78	30.80	31.66
PS-Cross-Outs	62.41	84.84	54.95	64.75	54.80	94.14
Median FBD Stroke Length	127.13	48.00	131.09	45.36	138.47	49.29
Median Equation Stroke Length	115.67	28.41	109.82	29.71	118.30	33.73
Median Cross-Out Stroke Length	720.70	501.7	760.89	447.9	842.26	665.7
Num FBD Strokes	165.40	87.04	205.41	106.3	157.71	90.50
Num Equation Strokes	432.58	228.6	528.45	281.7	457.81	277.4
Num Cross-Out Strokes	13.98	10.71	14.79	11.56	12.57	12.51
Score	0.64	0.28	0.61	0.31	0.74	0.25

Table 8: The mean and standard deviation of features for each assignment.

5.1 Predictive Ability of Solution History Features

We identified the aspects of solution histories that are the most predictive of student performance by analyzing the five feature subsets (Temporal Organization, Spatial Organization, Spatial Cluster, Cross-out, and Basic Pen Stroke). We trained regression models on each subset of features, using the relaxed lasso method, to predict the score on individual problems. The R^2 values, estimated through cross-validation and averaged for all 13 exam problems, are reported in Table 9.

The Temporal Organization features have the most predictive ability, and explain an average of 36% of the variance in performance. The next-best subsets, Spatial Cluster features and Basic Pen Stroke features, are significantly less predictive. They explain only 25% and 27% of the variance, respectively. The Spatial Organization features have the lowest useful predictive ability, explaining only 19% of the variance. The Cross-out features have no useful predictive ability.

	Temporal Organization	**Spatial Organization**	**Spatial Cluster**	**Cross-out**	**Basic Pen Stroke**
Temporal Organization	0.36	0.39	0.37	0.37	0.37
Spatial Organization		0.19	0.32	0.20	0.33
Spatial Cluster			0.25	0.26	0.33
Cross-out				0.02	0.28
Basic Pen Stroke					0.27

Table 9: The average R^2 values of relaxed lasso models trained on feature subsets. The values in light grey cells (diagonal) are from models trained on individual subsets, while the values in white cells are from models trained on pairwise combinations of subsets.

To evaluate the unique information in each feature subset, we constructed regression models with pairs of the subsets. The resulting R^2 values are again reported in Table 9. The combination of the Temporal Organization features and the Spatial Organization features has the highest predictive ability, with an average R^2 of 0.39. The combination of the Spatial Organization features with the Spatial Cluster features provides the largest increase (0.07) in R^2 over the individual subsets.

Table 10 shows the results of relaxed lasso models trained on the complete set of 41 features. These models have an average R^2 of 0.43, an improvement of 0.04 over the best pairwise combination of subsets. The lowest-performing model on an individual problem explained only 18% of the variance in performance on that problem, while the highest-performing model explained 65%. To provide another perspective on the predictive ability of our features, we also performed stepwise linear regression which produced models with an average R^2 value of 0.48, a minimum of 0.25, and a maximum of 0.64.

	Midterm 1 Problem 1	Midterm 1 Problem 2	Midterm 1 Problem 3	Midterm 2 Problem 1	Midterm 2 Problem 2	Midterm 2 Problem 3	Final Exam Problem 1	Final Exam Problem 2	Final Exam Problem 3	Final Exam Problem 4	Final Exam Problem 5	Final Exam Problem 6	Final Exam Problem 7	Average
Relaxed Lasso	0.34	0.50	0.57	0.36	0.65	0.18	0.43	0.49	0.24	0.31	0.52	0.62	0.38	0.43
Stepwise Regression	0.40	0.54	0.64	0.41	0.64	0.25	0.52	0.55	0.31	0.46	0.48	0.64	0.43	0.48
Absolute Difference	0.06	0.04	0.07	0.05	0.01	0.07	0.09	0.06	0.07	0.15	0.04	0.02	0.05	0.05
Mean of Scores	0.71	0.68	0.53	0.66	0.43	0.72	0.75	0.71	0.81	0.84	0.72	0.62	0.70	0.68
Stdev (σ) of Scores	0.31	0.27	0.22	0.28	0.29	0.28	0.23	0.21	0.17	0.20	0.24	0.32	0.27	0.25

Table 10: The R^2 values of regression models trained on the complete set of 41 features as well as the mean and standard deviation of the scores for each problem

5.2 Important Features

One measure of the usefulness of a feature is the frequency with which it is selected by the relaxed lasso method for the set of 13 exam problems. By contrast, features that are infrequently selected either have no predictive ability or are correlated with a more predictive feature. Figure 9 through Figure 13 show the features selected for each of the five feature subsets. Green is used to indicate features that are positively correlated with performance, while red is used to indicate negative correlation. The actual magnitudes of the coefficients are not displayed in the tables because the use of a tuning parameter (λ) in the relaxed lasso method makes it difficult to compare the magnitudes of the coefficients across models.[15]

Figure 9 shows the features selected from the Temporal Organization features. The two most important features are *Equation Effort* and *Other-Problem*. Both were selected for over half of the problems. The former was always positively correlated with performance, while the latter was always negatively correlated. *Num Small Breaks* was also selected for over half of the problems, but the sign of the coefficient varied: the correlation was positive for six problems and negative for one.

Exam	Midterm 1			Midterm 2			Final Exam						
Problem	1	2	3	1	2	3	1	2	3	4	5	6	7
Code	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4	C5	C6	C7

Table 11: Codes used to represent problem names. These are used in Figure 9 through Figure 14 and Table 12 through Table 16.

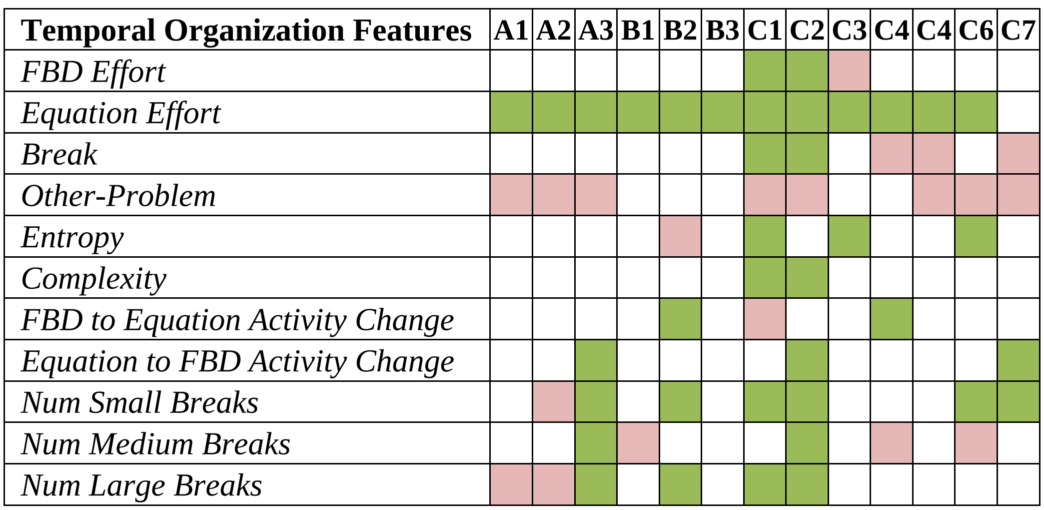

Figure 9: The sign of coefficients selected by the relaxed lasso method trained on the Temporal Organization features. Blank cells indicate unselected features, red cells indicate negative coefficients, and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

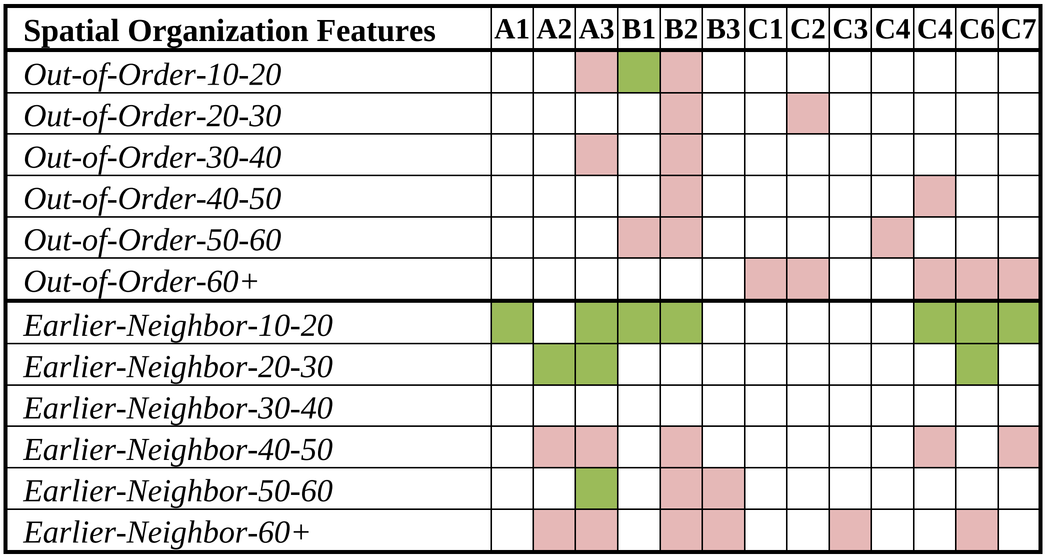

Figure 10: The sign of coefficients selected by the relaxed lasso method trained on Spatial Organization features. Blank cells indicate unselected features, red cells indicate negative coefficients, and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

The most important Spatial Organization feature (Figure 10) is the *Earlier-Neighbor-10-20* feature, which was selected for over half of the problems and is positively correlated with performance. Although selected fewer times, the *Earlier-Neighbor-20-30* feature also has a positive correlation with performance. By contrast, the features representing a greater degree of delay, i.e., *Earlier-Neighbor-40-50, Earlier-Neighbor-50-60*, and *Earlier-Neighbor-60+*, are negatively correlated with performance. Furthermore, while no individual Out-of-order feature was selected with high frequency, when such features were selected, they were almost always negatively correlated with performance.

Spatial Cluster Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
Num FBD Clusters													
FBD Revisits													
FBD Revisit Strokes													
Num Equation Clusters													
Equation Area Fraction													
Equation Revisits													
Equation Revisit Strokes													

Figure 11: The sign of coefficients selected by the relaxed lasso method trained on the Spatial Cluster features. Blank cells indicate unselected features, red cells indicate negative coefficients, and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

The most important Spatial Cluster features (Figure 11) are *Num FBD Clusters, FBD Revisits*, and *Equation Area Fraction*. Each was selected for at least half of the problems, and all are positively correlated with performance.

Cross-Out Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
FBD Strokes Crossed-Out													
Equation Strokes Crossed-Out													
Big-Cross-Outs													
Typo-Cross-Outs													
PS-Cross-Outs													

Figure 12: The sign of coefficients selected by the relaxed lasso method trained on the Cross-out features. Blank cells indicate unselected features, red cells indicate negative coefficients, and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

No single Cross-out feature was consistently selected (Figure 12). *Equation Strokes Crossed-Out* was selected for five the problems and was always positively correlated with performance. However, as the Cross-out features as a whole have no predictive ability (Table 9), this feature is unimportant.

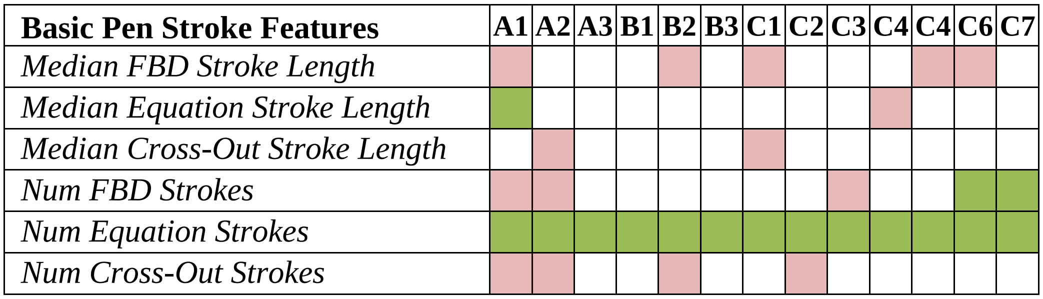

Figure 13: The sign of coefficients selected by the relaxed lasso method trained on the Basic Pen Stroke features. Blank cells indicate unselected features, red cells indicate negative coefficients, and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

The most important of the Basic Pen Stroke features (Figure 13) is the *Num Equation Strokes* feature which was selected for all problems and is always positively correlated with performance. None of the other Basic Pen Stroke features are selected frequently, although when they are selected, they are typically negatively correlated with performance.

Figure 9 through Figure 13 show the importance of the features within a subset. Figure 14 reveals the important features when they are examined as a complete set of 41. The rectangular cells in the figure are color-coded to indicate which features relaxed lasso selected from the complete set. The small circles are color-coded to indicate which features relaxed lasso selected from the subsets examined individually. Here again, green is used to indicate features that are positively correlated with performance, while red is used to indicate negative correlation.

For the complete set of features, only one is consistently selected: *Equation Area Fraction*. As would be expected, many of the features that were selected for individual subsets are not selected for the complete feature set. When a feature is selected for both, the sign of the coefficient is typically consistent across the two analyses. There are only three exceptions: this occurred twice for the *Equation Effort* feature and once for the *Equation Strokes Crossed-Out* feature. The former is likely due to an interaction between features, while the latter is consistent with lack of predictive ability of Cross-out features.

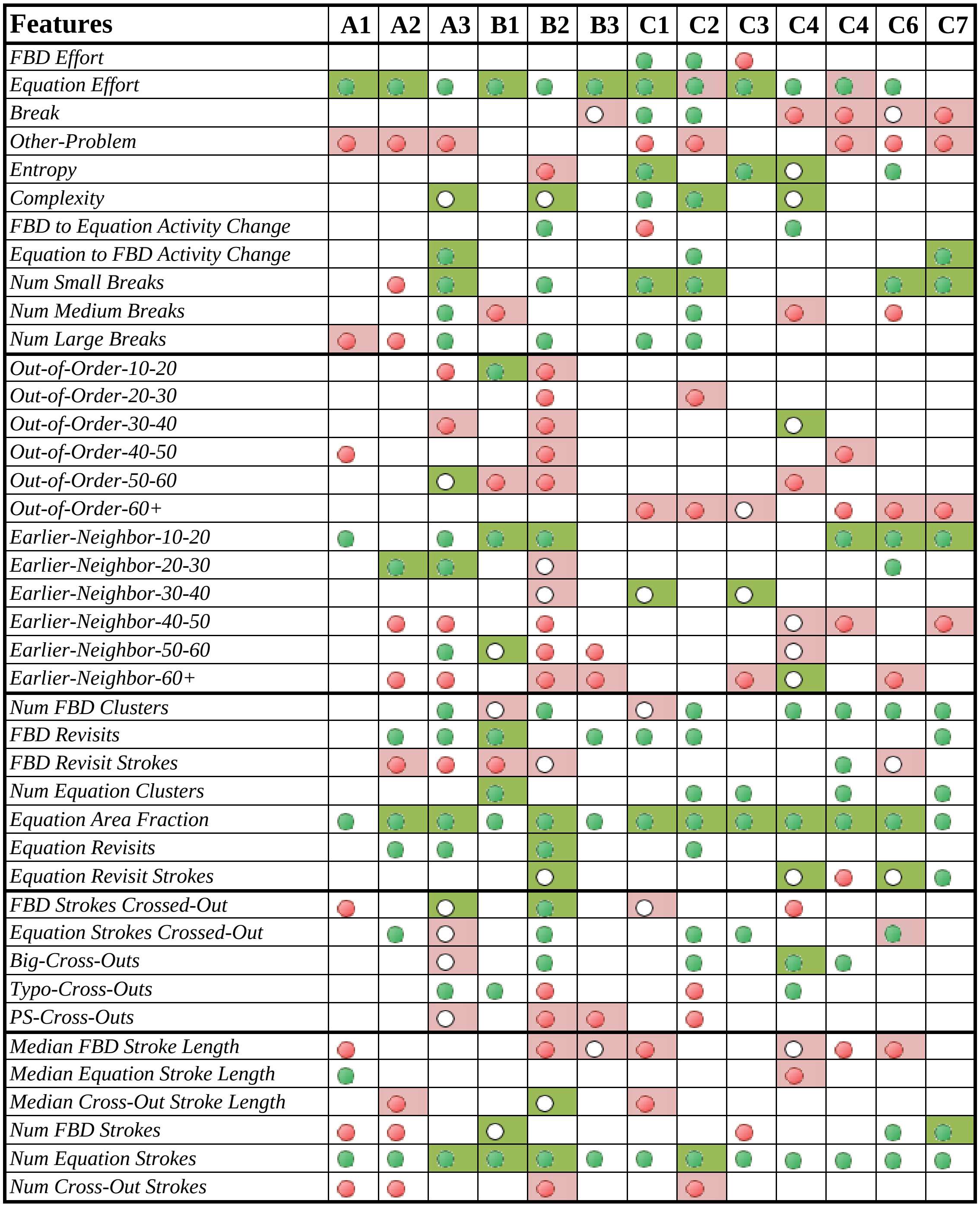

Figure 14: Comparison of relaxed lasso models trained on complete set of features to models trained on feature subsets. The rectangular cells indicate the features selected from the complete set. Circles indicate the features selected from the subsets when examined individually. Green is used to indicate features that are positively correlated with performance, while red is used to indicate negative correlation. Columns represent exam problems in order (as shown in Table 11).

5.3 Most Significant Features

We used stepwise linear regression to determine the statistical significance of the features. We constructed stepwise regression models for each of the five subsets of features. The standardized coefficients, β, and p-values for the models are shown in Table 12 through Table 16. Stepwise regression tends to select fewer features than the relaxed lasso method. Additionally, the signs of the coefficients tend to be consistent across the two methods.

Out of all five feature subsets, a total of three features were selected with high frequency and high confidence (i.e., small p-value). From the Temporal Organization features (Table 12), *Equation Effort* was selected for all but two problems with a p-value less than 0.005. From the Spatial Cluster features (Table 14), *Equation Area Fraction* was selected for all problems with a p-value less than 0.005. Finally, from the Basic Pen Stroke features (Table 16), *Num Equation Strokes* was selected for all problems with a p-value less than 0.005. In all three cases, the features are positively correlated with performance. Furthermore, all three features provide some measure of the amount of equation work in the solution. We provide additional analysis of these three features in the next section.

Temporal Organization Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
FBD Effort													
Equation Effort	0.46 (0.00)	0.60 (0.00)	0.51 (0.00)	0.54 (0.00)	0.74 (0.00)	0.42 (0.00)	0.64 (0.00)	0.40 (0.00)	0.39 (0.00)	0.48 (0.00)	0.67 (0.00)		
Break												-0.46 (0.00)	-0.50 (0.00)
Other-Problem		-0.20 (0.01)	-0.18 (0.01)										
Entropy							0.42 (0.00)			0.18 (0.03)			
Complexity													
FBD to Equation Activity Change							-0.18 (0.04)						
Equation to FBD Activity Change													
Num Small Breaks	-0.18 (0.04)		0.26 (0.01)					0.29 (0.01)				0.28 (0.02)	
Num Medium Breaks				-0.18 (0.02)									
Num Large Breaks	-0.29 (0.00)		0.15 (0.03)										

Table 12: The standardized coefficients, β, and p-values (in parentheses) of selected features from stepwise linear regression models trained on the Temporal Organization features. Red cells indicate negative coefficients and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

Spatial Organization Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
Out-of-Order-10-20					-0.21 (0.01)								
Out-of-Order-20-30								-0.31 (0.00)					
Out-of-Order-30-40			-0.23 (0.01)										
Out-of-Order-40-50											-0.28 (0.00)		
Out-of-Order-50-60				-0.24 (0.01)	-0.31 (0.00)					-0.35 (0.00)			
Out-of-Order-60+							-0.43 (0.00)	-0.34 (0.00)		-0.22 (0.02)	-0.37 (0.00)	-0.42 (0.00)	-0.42 (0.00)
Earlier-Neighbor-10-20	0.25 (0.01)				0.29 (0.00)							0.36 (0.00)	0.20 (0.03)
Earlier-Neighbor-20-30		0.22 (0.02)										0.19 (0.02)	
Earlier-Neighbor-30-40													
Earlier-Neighbor-40-50		-0.17 (0.05)	-0.22 (0.01)		-0.30 (0.00)						-0.31 (0.00)		-0.42 (0.00)
Earlier-Neighbor-50-60													
Earlier-Neighbor-60+		-0.30 (0.00)	-0.32 (0.00)			-0.31 (0.00)			-0.39 (0.00)				

Table 13: The standardized coefficients, β, and p-values (in parentheses) of selected features from stepwise linear regression models trained on the Spatial Organization features. Red cells indicate negative coefficients and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

Spatial Cluster Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
Num FBD Clusters			0.17 (0.03)					0.30 (0.00)		0.19 (0.04)	0.23 (0.01)	0.36 (0.00)	0.33 (0.00)
FBD Revisits		0.21 (0.01)	0.21 (0.01)		0.16 (0.03)	0.21 (0.03)		0.21 (0.02)					
FBD Revisit Strokes		-0.38 (0.00)									0.44 (0.01)		0.41 (0.00)
Num Equation Clusters				0.19 (0.03)									
Equation Area Fraction	0.31 (0.00)	0.36 (0.00)	0.61 (0.00)	0.52 (0.00)	0.51 (0.00)	0.32 (0.00)	0.48 (0.00)	0.54 (0.00)	0.44 (0.00)	0.37 (0.00)	0.89 (0.00)	0.44 (0.00)	0.51 (0.00)
Equation Revisits					0.24 (0.00)			0.23 (0.01)					
Equation Revisit Strokes													

Table 14: The standardized coefficients, β, and *p*-values (in parentheses) of selected features from stepwise linear regression models trained on the Spatial Cluster features. Red cells indicate negative coefficients and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

Cross-Out Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
FBD Strokes Crossed-Out													
Equation Strokes Crossed-Out												0.20 (0.05)	
Big-Cross-Outs													
Typo-Cross-Outs			0.35 (0.00)		0.32 (0.00)								
PS-Cross-Outs													

Table 15: The standardized coefficients, β, and p-values (in parentheses) of selected features from stepwise linear regression models trained on the Cross-out features. Red cells indicate negative coefficients and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

Basic Pen Stroke Features	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**	**C4**	**C4**	**C6**	**C7**
Median FBD Stroke Length							-0.31 (0.00)					-0.24 (0.00)	
Median Equation Stroke Length													
Median Cross-Out Stroke Length							-0.23 (0.01)						
Num FBD Strokes									-0.22 (0.03)				0.27 (0.00)
Num Equation Strokes	0.26 (0.00)	0.65 (0.00)	0.70 (0.00)	0.52 (0.00)	0.90 (0.00)	0.36 (0.00)	0.26 (0.00)	0.71 (0.00)	0.43 (0.00)	0.34 (0.00)	0.50 (0.00)	0.55 (0.00)	0.36 (0.00)
Num Cross-Out Strokes		-0.25 (0.01)			-0.27 (0.00)			-0.23 (0.00)					

Table 16: The standardized coefficients, β, and *p*-values (in parentheses) of selected features from stepwise linear regression models trained on the Basic Pen Stroke features. Red cells indicate negative coefficients and green cells indicate positive coefficients. Columns represent exam problems in order (as shown in Table 11).

5.4 Predictive Ability of the Most Significant Features

Section 5.3 suggests that features measuring the amount of work on equations are consistently significant and positively correlated with performance. Here we compare the importance of such features to that of the other features. Specifically, we used relaxed lasso to constructed regression models for two set of features. The first set contains all features that measure the amount of work on equations: *Equation Effort*, *Equation Area Fraction*, *Num Equation Strokes*, and *Num Equation Clusters*. (Only the first three of these were consistently selected by stepwise regression.) The second set contains the other 37 features. The R^2 values for these models are listed in Table 17. The four equation features resulted in an R^2 value of 0.34, while the other features resulted in a value of 0.40. Thus, although the four equation features are useful for prediction, the set comprising the other features is more predictive.

	Midterm 1 Problem 1	Midterm 1 Problem 2	Midterm 1 Problem 3	Midterm 2 Problem 1	Midterm 2 Problem 2	Midterm 2 Problem 3	Final Exam Problem 1	Final Exam Problem 2	Final Exam Problem 3	Final Exam Problem 4	Final Exam Problem 5	Final Exam Problem 6	Final Exam Problem 7	Average
Amount of Equation Work Features	0.27	0.46	0.51	0.30	0.56	0.16	0.25	0.39	0.18	0.19	0.43	0.43	0.22	0.34
All Other Features	0.31	0.48	0.53	0.33	0.59	0.17	0.41	0.45	0.18	0.25	0.48	0.63	0.39	0.40
Absolute Difference	0.04	0.02	0.02	0.03	0.03	0.01	0.16	0.07	0.00	0.06	0.05	0.20	0.17	0.06

Table 17: The R^2 values of regression models trained using the relaxed lasso method on the 4 features that measure the amount of equation work and on the 37 other features.

Chapter 6. Discussion

The analysis summarized in in Table 9 indicates that the Temporal Organization features are the most predictive of student performance. When pairs of feature subsets were analyzed, combining the Spatial Organization features with the Temporal Organization features provided the greatest predictive ability, reaffirming our intuition that both the spatial and temporal organization of the work are indicative of performance. For example, it is not surprising that out-of-order work is negatively correlated with performance.

Interestingly, the Cross-out features did not have any useful ability to predict performance. For example, for 10 of the 13 problems, stepwise linear regression found that no Cross-out feature had a statistically significant predictive ability (Table 15). We believe that cross-outs may still provide important information about a student's performance, but our features are ineffective. For example, our current features measure the absolute number of cross-out gestures and crossed-out strokes. It may be more useful to normalize these features by the total number of strokes in the solution to provide a measure of the fraction of the work that was corrected. Similarly, it is likely that other properties of cross-outs may make useful features such as the area of a page that is crossed out or the time in the sequence of activities at which the cross-outs occur.

The regression models from the relaxed lasso method explained between 18% and 65% of the variance in student performance on individual exam problems (Table 10). This large variation in the coefficient of determination suggests that differences in problem type and difficulty may affect the predictive ability of our approach.

Interestingly, the regression models had the best predictive ability on the problems with low average scores (Figure 15). One might expect that this was the result of a high degree of variation in student performance on such problems. However, as illustrated in Figure 16, there is little correlation between the variance in the scores and the accuracy of our models. More specifically, there is little correlation between the standard deviation of the scores and the coefficients of determination of our regression models. These results suggest that it may be possible to design problems for which our techniques are better able to predict performance and thus are better able to provide assessment results to the both the students and the instructor. However, developing principles for designing such problems will take additional research.

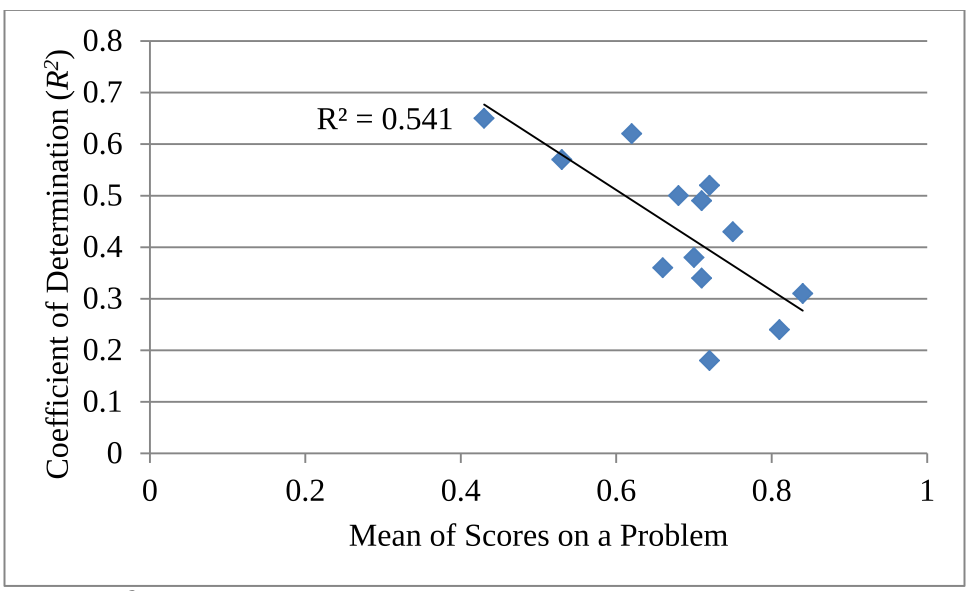

Figure 15: The R^2 values of regression models trained using the relaxed lasso method plotted verses the mean of scores on a problem (values are from Table 10). A linear least-squares regression line is shown with its corresponding R^2 value.

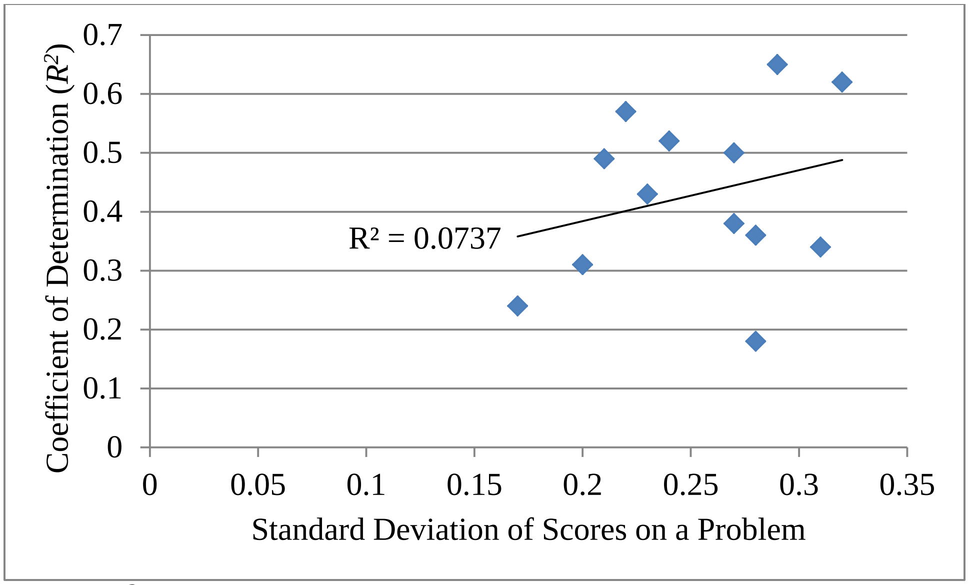

Figure 16: The R^2 values of regression models trained using the relaxed lasso method plotted verses the standard deviation of scores on a problem (values are from Table 10). A linear least-squares regression line is shown with its corresponding R^2 value.

Our current techniques examine student performance on individual exam problems considered in isolation. We may be able to provide better assessment results by tracking individual students throughout a course. For example, building models that simultaneously consider multiple exams and assignments may allow us to determine if a student's understanding is improving with time. In fact, it may be possible to explicitly model a student's development during a course.

The analysis in Section 5.3 revealed that the most important features for predicting performance all capture aspects of the amount of equation work in a solution. Furthermore, the analysis in Section 5.4 indicated that, taken together, the four features that directly measure the amount of work on equations can explain 34% of the variance in performance. To gain insights into the power of these features, we informally examined a few examples of student work. We found that it was not uncommon for a student to achieve a low score on a problem because he or she oversimplified the problem setup, enabling an incorrect solution to be obtained with little equation work. Similarly, other students scored low because they failed to construct many equations, likely due to a lack of time, understanding, or both. It appears that a student who produces the necessary amount of equation work is more likely to produce the correct answer. This simple fact enables a very inexpensive means for providing a rough assessment of student performance.

Our analysis of the most important and most significant features in Chapter 5 presents a profile of a high-performing statics student. Such students tend to spend large portions of their effort working on equations rather than free body diagrams (Figure 9 and Figure 13). Similarly, they tend to cover more of the page with equation work than with free body diagram work (Figure 11). They also tend to complete each problem without switching to other problems mid solution (Figure 9). Although no single Spatial Organization feature was consistently selected by the regression techniques, it appears that high performing students tend to work from the top of the page to the bottom without returning to add additional work to earlier sections of the page (Figure 10). When they do revise their work, they tend to do so soon after the original work was written (Figure 10). Students who do not match this profile may need extra guidance in the course.

The variation in the features selected for various problems is due in part to the relaxed lasso's attempt to select the best features out of a set of correlated features.[15] Newer regression methods such as the random lasso[18] method or the forward-lasso adaptive shrinkage (FLASH)[19] method may provide more stable selection of features without sacrificing predictive ability. We may also be able to improve the stability of the feature selection by simplifying the feature set. For example, the Out-of-Order features were all negatively correlated with performance (Figure 10), and thus could be combined into a single feature. Likewise, the Earlier-Neighbor features could be combined into two distinct features based on the positive and negative coefficients demonstrated.

Chapter 7. Conclusion

We have examined how the organization of a student's solution to a problem relates to the correctness of the work. In this study, students in an undergraduate statics course completed all of their work (homework, quizzes, and exams) using digital pens that recorded the work as time-stamped pen strokes. We characterized the solution history of each problem with a number of quantitative features describing the organization of the work. Regression models revealed that, on average, about 43% of the variance in student performance could be explained by these features. Additionally, the Temporal Organization features were the most predictive of performance.

Our examination of the predictive ability of our features presents a profile of a high-performing statics student. For example, such students tend to spend more effort on equations than on free body diagrams, they tend to complete each problem without switching to other problems mid solution, and they tend to work from the top of the page to the bottom without returning to revise earlier work. Students who do not match this profile may need extra guidance in the course.

This work is a first step at building techniques that can provide automated assessment of performance from handwritten student work. Our results demonstrate that the temporal and spatial organization of a student's work is indeed indicative of performance. The features that we have developed have produced promising results. However, we believe that it may be possible to provide even better assessment by developing new features that capture other properties of the solutions histories. While our work has focused on engineering statics, we expect that these techniques will generalize to other domains for which problem solutions include both diagrams and equations.

References

[1] Rogers Hall. Video Recording as Theory. In D. Lesh and A. Kelly (Eds.) *Handbook of Research Design in Mathematics and Science Education*, 647-664. Mahweh, NJ: Lawrence Erlbaum, 2000.

[2] Cristóbal Romero and Sebastián Ventura. Educational data mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 40(6):601-618, 2010.

[3] Sharon Oviatt, Alex Arthur, and Julia Cohen. Quiet interfaces that help students think. In *Proceedings of the 19th annual ACM symposium on user interface software and technology (UIST '06)*, 191–200, New York, NY, USA, 2006.

[4] Paul Steif and Anna Dollár. Study of Usage Patterns and Learning Gains in a Web-based Interactive Static Course. *Journal of Engineering Education*, 94(4):321-333, 2009.

[5] Paul Steif, Jamie Lobue, Anne Fay, and Levent Kara. Improving Problem Solving Performance by Inducing Talk about Salient Problem Features. *Journal of Engineering Education*, 99(2):135-142, 2010.

[6] Paul Blanc. Solving a Non-routine Problem: What Helps, What Hinders? In *Proceedings of the British Society for Research into Learning Mathematics*, 19(2):1-6, 1999.

[7] Marisa Orr, Lisa Benson, Matthew Ohland, and Sherrill Biggers. Student Study Habits and Their Effectiveness in an Integrated Statics and Dynamics Class. In *Proceedings of the 2008 American Society for Engineering Education Annual Conference and Exposition*, 2008.

[8] Antti Rasila, Linda Havola, Helle Majander, and Pekka Alestalo. Automatic assessment in engineering mathematics: evaluation of the impact. In Myller, E. (ed.), *ReflekTori 2010 Symposium of Engineering Education*, 37-45. Aalto University School of Science and Technology, 2010.

[9] John Chen, Dexter Whittinghill, and Jennifer Kadlowec. Classes that click: Fast, rich feedback to enhance student learning and satisfaction. *Journal of Engineering Education*, 99(2):159-168, 2010.

[10] Han-lung Lin, Thomas Stahovich, and James Herold. Automatic Handwritten Statics Solution Classification and Its Applications in Predicting Student Performance. In *Proceedings of the 2012 American Society for Engineering Education Annual Conference*, 2012.

[11] C. S. Wallace and D. L. Dowe. Minimum Message Length and Kolmogorov Complexity. *Computer Journal*, 42(4):270-283, 1999.

[12] P. Deutsch and J. L. Gailly. ZLIB Compressed Data Format Specification version 3.3. *RFC Editor*, 1996.

[13] P. Deutsch. DEFLATE Compressed Data Format Specification version 1.3. RFC 1951, Aladdin Enterprises, May 1996.

[14] Nicolai Meinshausen. Relaxed Lasso. *Computational Statistics & Data Analysis,* Elsevier, 52(1):374-393 Sept. 2007.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer, 2009.

[16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-288, 1996.

[17] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301–320, April 2005.

[18] S. Wang, B. Nan, S. Rosset, J. Zhu. Random lasso. *The Annals of Applied Statistics*, 5(1):468–485, 2011.

[19] P. Radchenko and G. M. James. Improved variable selection with forward-lasso adaptive shrinkage. *The Annals of Applied Statistics* 5(1):427-448, 2011.

Appendix A. Descriptive Statistics of the Features

This appendix contains descriptive statistics of each feature, as well as the score, for each exam problem. More specifically, the mean and standard deviation for each are contained in the following tables.

Features	**Midterm 1**		**Midterm 2**		**Final Exam**	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
FBD Effort	50.42	18.51	60.36	20.41	48.67	20.91
Equation Effort	95.87	37.50	104.46	39.33	99.62	39.77
Break	245.33	41.93	226.53	43.71	244.15	48.12
Other-Problem	0.96	1.29	0.85	0.92	0.84	1.00
Entropy	0.61	0.08	0.61	0.10	0.59	0.11
Complexity	99.71	11.69	104.72	13.19	98.31	14.99
FBD to Equation Activity Change	7.59	4.78	10.18	5.64	8.64	5.49
Equation to FBD Activity Change	7.10	4.88	9.83	5.62	8.21	5.54
Num Small Breaks	84.76	32.43	98.20	39.73	79.65	38.23
Num Medium Breaks	5.46	3.10	4.66	2.98	4.45	3.24
Num Large Breaks	0.52	0.87	0.35	0.62	0.59	0.94
Out-of-Order-10-20	0.19	0.09	0.21	0.11	0.19	0.10
Out-of-Order-20-30	0.10	0.08	0.11	0.09	0.09	0.08
Out-of-Order-30-40	0.07	0.07	0.06	0.07	0.05	0.06
Out-of-Order-40-50	0.04	0.06	0.04	0.05	0.03	0.06
Out-of-Order-50-60	0.03	0.06	0.02	0.05	0.02	0.04
Out-of-Order-60+	0.05	0.09	0.04	0.09	0.04	0.09
Earlier-Neighbor-10-20	0.24	0.11	0.26	0.11	0.25	0.12
Earlier-Neighbor-20-30	0.18	0.10	0.18	0.09	0.19	0.12
Earlier-Neighbor-30-40	0.12	0.09	0.11	0.08	0.11	0.09
Earlier-Neighbor-40-50	0.08	0.07	0.06	0.06	0.07	0.07
Earlier-Neighbor-50-60	0.06	0.06	0.04	0.05	0.04	0.05
Earlier-Neighbor-60+	0.09	0.10	0.05	0.07	0.06	0.09
Num FBD Clusters	2.87	1.91	4.05	2.29	3.69	2.28
FBD Revisits	8.44	5.90	9.26	6.33	7.54	5.82
FBD Revisit Strokes	0.16	0.12	0.15	0.13	0.12	0.10
Num Equation Clusters	2.75	1.55	3.33	1.79	3.08	1.79
Equation Area Fraction	0.66	0.14	0.62	0.16	0.66	0.15
Equation Revisits	3.89	4.36	5.81	5.50	5.32	5.45
Equation Revisit Strokes	0.32	0.27	0.33	0.24	0.34	0.25
FBD Strokes Crossed-Out	23.91	27.49	25.89	26.67	19.87	24.75
Equation Strokes Crossed-Out	56.34	68.33	57.04	58.41	53.57	70.10
Big-Cross-Outs	1.82	2.03	2.19	2.41	1.78	2.08
Typo-Cross-Outs	29.24	26.33	39.81	37.78	30.80	31.66
PS-Cross-Outs	62.41	84.84	54.95	64.75	54.80	94.14
Median FBD Stroke Length	127.13	48.00	131.09	45.36	138.47	49.29
Median Equation Stroke Length	115.67	28.41	109.82	29.71	118.30	33.73
Median Cross-Out Stroke Length	720.70	501.7	760.89	447.9	842.26	665.7
Num FBD Strokes	165.40	87.04	205.41	106.3	157.71	90.50
Num Equation Strokes	432.58	228.6	528.45	281.7	457.81	277.4
Num Cross-Out Strokes	13.98	10.71	14.79	11.56	12.57	12.51
Score	0.64	0.28	0.61	0.31	0.74	0.25

Features	**Midterm 1 Problem 1**		**Midterm 1 Problem 2**		**Midterm 1 Problem 3**	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
FBD Effort	49.46	19.46	49.89	18.30	51.91	17.79
Equation Effort	102.43	33.06	79.46	32.18	105.58	41.27
Break	239.81	37.65	262.26	39.79	234.05	43.09
Other-Problem	0.89	1.19	1.20	1.28	0.80	1.38
Entropy	0.60	0.08	0.63	0.09	0.60	0.08
Complexity	99.46	10.52	97.13	12.71	102.52	11.20
FBD to Equation Activity Change	7.00	4.22	7.28	4.20	8.49	5.67
Equation to FBD Activity Change	6.46	4.35	6.87	4.30	7.98	5.77
Num Small Breaks	86.59	25.61	69.67	29.82	97.91	34.95
Num Medium Breaks	5.20	3.28	5.54	2.90	5.66	3.11
Num Large Breaks	0.60	1.03	0.47	0.68	0.49	0.86
Out-of-Order-10-20	0.19	0.08	0.19	0.09	0.19	0.09
Out-of-Order-20-30	0.11	0.07	0.11	0.08	0.09	0.08
Out-of-Order-30-40	0.07	0.06	0.08	0.07	0.05	0.07
Out-of-Order-40-50	0.04	0.06	0.04	0.06	0.04	0.06
Out-of-Order-50-60	0.04	0.07	0.03	0.05	0.03	0.05
Out-of-Order-60+	0.06	0.10	0.06	0.10	0.04	0.07
Earlier-Neighbor-10-20	0.25	0.12	0.20	0.10	0.26	0.11
Earlier-Neighbor-20-30	0.18	0.09	0.17	0.11	0.19	0.10
Earlier-Neighbor-30-40	0.13	0.09	0.13	0.09	0.10	0.08
Earlier-Neighbor-40-50	0.08	0.06	0.09	0.08	0.07	0.06
Earlier-Neighbor-50-60	0.05	0.06	0.07	0.06	0.04	0.05
Earlier-Neighbor-60+	0.07	0.08	0.11	0.12	0.08	0.09
Num FBD Clusters	2.62	1.62	2.63	1.84	3.36	2.16
FBD Revisits	7.57	5.55	8.59	5.89	9.16	6.17
FBD Revisit Strokes	0.14	0.10	0.19	0.14	0.16	0.12
Num Equation Clusters	2.25	1.28	2.91	1.68	3.07	1.55
Equation Area Fraction	0.70	0.12	0.61	0.14	0.68	0.14
Equation Revisits	3.58	4.22	3.51	3.92	4.57	4.83
Equation Revisit Strokes	0.34	0.29	0.28	0.26	0.33	0.25
FBD Strokes Crossed-Out	23.97	30.74	26.35	29.75	21.44	20.92
Equation Strokes Crossed-Out	63.41	79.89	44.51	58.34	60.99	63.83
Big-Cross-Outs	1.97	2.19	1.52	1.68	1.98	2.16
Typo-Cross-Outs	26.93	24.53	26.31	28.24	34.46	25.51
PS-Cross-Outs	72.51	103.68	55.58	75.03	59.08	71.83
Median FBD Stroke Length	119.76	48.16	132.88	53.45	128.80	41.09
Median Equation Stroke Length	111.20	26.14	124.14	30.43	111.82	26.84
Median Cross-Out Stroke Length	751.34	457.96	762.72	619.62	648.64	398.34
Num FBD Strokes	154.01	82.56	157.87	85.70	184.27	90.23
Num Equation Strokes	432.32	176.97	339.07	179.52	525.57	276.61
Num Cross-Out Strokes	13.73	11.73	12.74	10.31	15.47	9.89
Score	0.71	0.31	0.68	0.27	0.53	0.22

Features	Midterm 2 Problem 1		Midterm 2 Problem 2		Midterm 2 Problem 3	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
FBD Effort	63.97	21.12	58.38	20.56	58.70	19.20
Equation Effort	112.98	38.26	93.50	46.40	106.80	29.29
Break	213.63	42.71	240.11	47.87	225.96	36.08
Other-Problem	0.92	1.01	1.13	0.88	0.52	0.73
Entropy	0.62	0.09	0.59	0.13	0.62	0.08
Complexity	108.31	14.01	100.76	13.36	105.05	11.02
FBD to Equation Activity Change	12.68	5.55	8.73	5.71	9.10	4.79
Equation to FBD Activity Change	12.35	5.52	8.37	5.63	8.75	4.83
Num Small Breaks	127.33	36.89	81.00	34.97	86.12	29.65
Num Medium Breaks	6.37	3.08	4.30	2.49	3.29	2.47
Num Large Breaks	0.61	0.80	0.28	0.49	0.17	0.40
Out-of-Order-10-20	0.22	0.10	0.19	0.11	0.23	0.11
Out-of-Order-20-30	0.10	0.08	0.10	0.08	0.14	0.09
Out-of-Order-30-40	0.05	0.06	0.07	0.08	0.06	0.06
Out-of-Order-40-50	0.03	0.05	0.04	0.06	0.04	0.05
Out-of-Order-50-60	0.02	0.05	0.03	0.05	0.02	0.03
Out-of-Order-60+	0.04	0.09	0.04	0.08	0.03	0.06
Earlier-Neighbor-10-20	0.26	0.09	0.25	0.12	0.28	0.12
Earlier-Neighbor-20-30	0.18	0.09	0.17	0.10	0.20	0.09
Earlier-Neighbor-30-40	0.10	0.07	0.11	0.07	0.13	0.09
Earlier-Neighbor-40-50	0.06	0.05	0.07	0.06	0.06	0.06
Earlier-Neighbor-50-60	0.04	0.04	0.04	0.05	0.04	0.04
Earlier-Neighbor-60+	0.05	0.06	0.06	0.09	0.04	0.06
Num FBD Clusters	4.80	2.33	3.67	2.29	3.67	2.05
FBD Revisits	11.60	6.83	7.86	6.11	8.31	5.35
FBD Revisit Strokes	0.14	0.10	0.19	0.17	0.13	0.10
Num Equation Clusters	4.15	1.76	3.09	1.84	2.75	1.43
Equation Area Fraction	0.62	0.13	0.59	0.21	0.66	0.12
Equation Revisits	7.23	6.14	5.21	5.24	5.00	4.82
Equation Revisit Strokes	0.36	0.22	0.34	0.26	0.30	0.24
FBD Strokes Crossed-Out	31.35	28.74	25.11	26.43	21.21	23.85
Equation Strokes Crossed-Out	71.55	66.03	51.46	52.54	48.07	53.35
Big-Cross-Outs	2.47	2.47	2.09	2.52	2.02	2.21
Typo-Cross-Outs	46.25	42.87	38.11	37.55	35.06	31.49
PS-Cross-Outs	72.22	74.81	46.43	51.93	46.12	62.40
Median FBD Stroke Length	124.56	38.43	141.44	50.80	127.37	44.65
Median Equation Stroke Length	102.98	25.47	117.72	35.46	108.96	25.66
Median Cross-Out Stroke Length	704.21	338.13	760.74	459.29	818.25	524.11
Num FBD Strokes	257.62	103.78	184.06	110.93	174.35	82.87
Num Equation Strokes	671.60	291.35	460.32	307.65	452.81	170.83
Num Cross-Out Strokes	18.59	12.37	13.17	10.27	12.59	11.08
Score	0.66	0.28	0.43	0.29	0.72	0.28

Features	Final Problem 1		Final Problem 2		Final Problem 3	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
FBD Effort	44.57	17.22	43.41	16.64	46.96	17.38
Equation Effort	87.79	33.61	123.39	36.04	71.86	22.96
Break	260.53	37.14	222.80	38.90	274.91	30.62
Other-Problem	1.08	1.16	1.11	1.12	0.91	1.03
Entropy	0.60	0.10	0.55	0.09	0.64	0.08
Complexity	95.09	13.17	105.49	11.21	90.67	12.04
FBD to Equation Activity Change	6.57	3.80	9.88	6.44	5.83	3.41
Equation to FBD Activity Change	6.14	3.85	9.38	6.49	5.35	3.54
Num Small Breaks	73.55	27.61	117.92	36.93	55.28	22.25
Num Medium Breaks	5.39	3.34	7.50	4.26	4.15	2.70
Num Large Breaks	0.86	1.07	0.84	1.28	0.71	0.91
Out-of-Order-10-20	0.18	0.10	0.18	0.09	0.18	0.10
Out-of-Order-20-30	0.10	0.09	0.07	0.05	0.12	0.09
Out-of-Order-30-40	0.05	0.06	0.03	0.05	0.07	0.07
Out-of-Order-40-50	0.04	0.06	0.02	0.04	0.06	0.07
Out-of-Order-50-60	0.02	0.04	0.01	0.04	0.03	0.06
Out-of-Order-60+	0.03	0.08	0.03	0.07	0.04	0.08
Earlier-Neighbor-10-20	0.23	0.12	0.26	0.11	0.22	0.11
Earlier-Neighbor-20-30	0.18	0.12	0.17	0.10	0.18	0.12
Earlier-Neighbor-30-40	0.13	0.10	0.10	0.08	0.14	0.10
Earlier-Neighbor-40-50	0.08	0.07	0.05	0.05	0.09	0.08
Earlier-Neighbor-50-60	0.04	0.06	0.02	0.03	0.05	0.05
Earlier-Neighbor-60+	0.06	0.09	0.03	0.06	0.07	0.10
Num FBD Clusters	2.91	1.79	4.27	2.71	2.35	1.24
FBD Revisits	7.82	5.77	10.50	6.74	5.82	4.91
FBD Revisit Strokes	0.13	0.10	0.10	0.07	0.15	0.10
Num Equation Clusters	2.73	1.54	3.42	1.81	2.55	1.41
Equation Area Fraction	0.64	0.16	0.75	0.11	0.58	0.13
Equation Revisits	2.64	2.94	5.60	5.41	2.79	3.72
Equation Revisit Strokes	0.28	0.27	0.40	0.25	0.20	0.22
FBD Strokes Crossed-Out	18.30	23.12	28.34	29.65	17.45	23.86
Equation Strokes Crossed-Out	47.33	53.59	97.89	104.03	34.17	56.01
Big-Cross-Outs	1.75	1.59	3.05	2.76	1.13	1.66
Typo-Cross-Outs	26.20	23.09	45.30	33.95	19.09	21.37
PS-Cross-Outs	52.18	78.18	104.52	155.06	42.89	86.40
Median FBD Stroke Length	136.83	55.25	125.32	40.32	151.90	46.46
Median Equation Stroke Length	121.88	38.18	112.37	25.99	120.33	32.12
Median Cross-Out Stroke Length	845.78	688.23	747.39	378.39	858.77	836.14
Num FBD Strokes	143.11	76.56	194.89	99.42	121.36	57.73
Num Equation Strokes	401.34	193.85	716.87	268.64	274.62	120.33
Num Cross-Out Strokes	11.15	8.37	21.22	15.44	8.96	11.62
Score	0.75	0.23	0.71	0.21	0.81	0.17

Features	Final Problem 4		Final Problem 5		Final Problem 6	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
FBD Effort	63.36	21.74	56.24	16.79	34.88	17.37
Equation Effort	107.10	31.26	128.18	36.14	77.18	36.42
Break	221.58	38.51	207.10	40.64	282.43	43.47
Other-Problem	0.65	0.97	0.74	0.94	0.81	0.87
Entropy	0.63	0.08	0.59	0.07	0.56	0.15
Complexity	104.35	10.00	107.50	11.17	85.34	16.55
FBD to Equation Activity Change	9.60	4.55	11.50	5.61	5.45	3.27
Equation to FBD Activity Change	9.11	4.59	11.16	5.67	5.07	3.34
Num Small Breaks	90.41	31.18	108.84	32.71	47.98	23.05
Num Medium Breaks	3.48	2.62	4.33	2.37	3.55	2.11
Num Large Breaks	0.39	0.84	0.38	0.70	0.52	0.79
Out-of-Order-10-20	0.24	0.11	0.19	0.09	0.19	0.11
Out-of-Order-20-30	0.08	0.08	0.07	0.06	0.10	0.09
Out-of-Order-30-40	0.04	0.06	0.03	0.05	0.06	0.06
Out-of-Order-40-50	0.02	0.04	0.02	0.04	0.03	0.06
Out-of-Order-50-60	0.01	0.04	0.01	0.02	0.02	0.04
Out-of-Order-60+	0.02	0.06	0.02	0.04	0.08	0.16
Earlier-Neighbor-10-20	0.32	0.13	0.29	0.10	0.22	0.14
Earlier-Neighbor-20-30	0.23	0.13	0.19	0.11	0.20	0.12
Earlier-Neighbor-30-40	0.11	0.08	0.10	0.06	0.12	0.09
Earlier-Neighbor-40-50	0.05	0.06	0.05	0.04	0.08	0.07
Earlier-Neighbor-50-60	0.03	0.04	0.03	0.04	0.04	0.05
Earlier-Neighbor-60+	0.03	0.06	0.05	0.07	0.08	0.12
Num FBD Clusters	4.04	1.86	5.02	2.52	2.73	1.60
FBD Revisits	7.72	4.96	8.17	6.17	4.09	3.63
FBD Revisit Strokes	0.12	0.09	0.11	0.12	0.11	0.10
Num Equation Clusters	2.99	1.64	3.59	1.91	2.58	1.54
Equation Area Fraction	0.64	0.12	0.70	0.13	0.66	0.17
Equation Revisits	6.37	4.99	7.94	5.60	3.82	4.13
Equation Revisit Strokes	0.36	0.23	0.42	0.22	0.32	0.28
FBD Strokes Crossed-Out	20.90	23.94	25.09	26.32	12.05	17.22
Equation Strokes Crossed-Out	54.97	53.54	67.83	77.00	30.41	43.36
Big-Cross-Outs	2.18	2.11	2.49	2.33	0.84	1.17
Typo-Cross-Outs	37.31	32.09	51.06	43.20	16.57	22.20
PS-Cross-Outs	49.79	63.75	54.65	86.78	31.94	54.33
Median FBD Stroke Length	147.23	47.36	124.97	42.25	161.58	59.16
Median Equation Stroke Length	123.11	33.62	108.68	30.31	132.67	36.66
Median Cross-Out Stroke Length	971.11	807.43	831.31	540.69	924.20	731.09
Num FBD Strokes	183.61	85.52	220.70	91.01	84.53	53.17
Num Equation Strokes	461.05	199.38	736.16	304.97	266.85	157.61
Num Cross-Out Strokes	12.68	11.25	15.66	14.74	7.46	8.54
Score	0.84	0.20	0.72	0.24	0.62	0.32

Features	Final Problem 7	
	Mean	Stdev
FBD Effort	50.37	25.34
Equation Effort	100.86	41.79
Break	241.64	49.92
Other-Problem	0.55	0.77
Entropy	0.58	0.12
Complexity	99.06	15.59
FBD to Equation Activity Change	11.50	6.36
Equation to FBD Activity Change	11.14	6.36
Num Small Breaks	61.75	26.52
Num Medium Breaks	2.67	2.24
Num Large Breaks	0.42	0.70
Out-of-Order-10-20	0.20	0.10
Out-of-Order-20-30	0.10	0.09
Out-of-Order-30-40	0.06	0.07
Out-of-Order-40-50	0.04	0.06
Out-of-Order-50-60	0.03	0.05
Out-of-Order-60+	0.05	0.09
Earlier-Neighbor-10-20	0.24	0.12
Earlier-Neighbor-20-30	0.17	0.11
Earlier-Neighbor-30-40	0.11	0.08
Earlier-Neighbor-40-50	0.07	0.08
Earlier-Neighbor-50-60	0.05	0.06
Earlier-Neighbor-60+	0.07	0.09
Num FBD Clusters	4.51	2.44
FBD Revisits	8.45	5.91
FBD Revisit Strokes	0.13	0.13
Num Equation Clusters	3.70	2.21
Equation Area Fraction	0.66	0.17
Equation Revisits	8.06	7.26
Equation Revisit Strokes	0.38	0.23
FBD Strokes Crossed-Out	16.52	24.11
Equation Strokes Crossed-Out	41.17	61.38
Big-Cross-Outs	1.00	1.38
Typo-Cross-Outs	19.22	18.94
PS-Cross-Outs	46.40	82.90
Median FBD Stroke Length	122.89	38.65
Median Equation Stroke Length	109.88	32.37
Median Cross-Out Stroke Length	719.19	541.26
Num FBD Strokes	151.72	85.32
Num Equation Strokes	337.79	163.77
Num Cross-Out Strokes	10.61	10.50
Score	0.70	0.27