Perspectives, Performance Measuring, and a Powerful Multivariate Analysis of Grades In Teaching Statistics

By Bill Seaver, Ph.D., Issac Edmiston, and Missy Morris

One of the authors has been teaching statistics over thirty years but also working as a consultant, both full-time and part-time many years as well. This research deals with some suggestions on measuring performance on homework, quizzes, exams, and projects as they relate to the final course grade. Typically, instructors compute grades as to the contract from their syllabus. On close calls, every instructor has his/her ways to make close calls, such as attendance, final exam grade, homework, trends, etc. In addition, it is not unusual for today's students to nit-pick about grades, a point here and there. As such, large classes create a lot of work addressing this issue. Also, how does one deal with the variation in grading that could come from the teaching assistant? Or what about the instructor second guessing the weighting scheme for the class? This instructor has pulled together a former student and teaching assistant for such additional insights. Finally, as statisticians should we not use statistics to check our individual student grades, grading schemes and close calls. In this case, we rely on some sampling, multivariate statistics, and conceptual scores. Several different classes are used to display the options for deciding grades.

Literature Review

There is existing literature which addresses how student performance is being measured in college statistics courses in the form of assessment and as to cheating. This information has greatly affected how one instructor has changed his teaching and his dealing with grades into context as to what other academic professionals are saying. The following articles offer suggestions for improving both the forms of assessment and curriculum currently being used in order to more effectively measure the level of student learning and better set students up for success in their courses.

The first article that discusses the current assessment environment in college statistics comes from Garfield et al. (2011), who call for changes to be made in the way that instructors evaluate student performance. The authors first note that students will put effort into and care about the assignments that are graded and that because of this, instructors should attempt to use assessment to aid in the learning process, especially providing proactive feedback that will help students improve. This feedback should not simply be supplying students with the correct answers, but it should be very specific in nature and be connected to the learning goals for the course. In order to promote the most accommodating learning environment, the authors suggest that instructors ought to keep assessment aligned with the curriculum being taught in the class at the time and to allow students to provide their input on what is expected of them (Garfield et al., 2011). Other suggestions are made; but the final section of their article deals with fairness, where the authors acknowledge that introductory statistics courses are becoming more common in the curriculum of a variety of disciplines and that low grades are frequently due to a lack of accommodation for students of different backgrounds and base levels of knowledge. In light of this, Garfield et al. (2011) make some profound proposals to fundamentally change the current landscape of assessment, which include grading students in context of their circumstances and learning preferences, rewarding effort with higher grades, and creating an environment where making mistakes is acceptable.

Moore and Kaplan (2015) further the discussion on assessment in undergraduate statistics courses by stating that instructors are being encouraged more and more to define learning outcomes for their students and put plans in place to evaluate these outcomes so that improvement can take place. In order to gain comprehensive measures of student performance, they suggest that instructors use multiple different types of assessment methods in their courses. The authors advocate the use of both direct assessment options, which relate the knowledge of the students to the instructors in the form of traditional assignments that are graded for accuracy, and indirect assessment options that ask students to comment on their levels of knowledge and understanding in the form of surveys (Moore & Kaplan, 2015). Doing this will provide instructors with a more thorough knowledge of what their students know, so that they can alter their teaching methods and materials appropriately to try to maximize students understanding. The authors also briefly discuss the use of rubrics in assessment, to which they say that rubrics can be beneficial in categorizing student performance and relating the assessment tasks back to the learning outcomes. They conclude their report arguing that in order for assessment to measure student performance as effectively as possible, instructors need to have the freedom to customize their assessment techniques to best meet the objectives of their course.

While the previous articles largely deal with assessment, an article by Green and Blankenship (2015) discusses changing the curriculum of statistics courses to try to promote more conceptual understanding. They begin by arguing that statistics students need sufficient conceptual knowledge to be able to think critically about real-world scenarios and that the current landscape of the statistics curriculum does not promote this kind of thinking. In order to keep students better engaged, the authors advocate using writing exercises that ask students to summarize the main points discussed during the lecture and encourage them to ask questions about anything that they did not fully understand. They argue that such exercises promote class-wide discussion and help bring students confidence to speak up during class (Green & Blankenship, 2015). The authors also go into detail about specific in-class activities that they have utilized in the past to promote a better understanding of the concepts, which the instructors usually grade simply for completion. These activities involve problems that are more challenging in nature that what the students would be expected to answer on homework or exams, with students working in groups to promote discussion of different approaches to and perspectives on the problems. In the final section of their report, Green and Blankenship (2015) do briefly discuss assessment, stating that the combination of having both assignments graded just for completion and assignments graded for accuracy is an effective way to keep students relaxed and encourage better learning. They do consent that exams should still count considerably toward assessment, but they propose that more conceptually-driven questions should be asked in addition to the traditional computational questions.

Moving on from discussion on assessment and curriculum to academic dishonesty, D’Souza and Siegfeldt (2017) offer a framework for instructors to identify and handle cases of cheating, which they cite as becoming a bigger issue in recent years with out-of-class exams being more common in order to avoid wasting class time. They successfully applied their method to a situation at a Virginia university where cheating was suspected because of students with low class attendance scoring unusually high on their online exams as compared to the year prior when the same exam was given in the classroom (D’Souza & Siegfeldt, 2017). By having three increasingly more sophisticated phases in their framework, D’Souza and Siegfeldt’s approach provides flexibility and the opportunity to evaluate instances in more than one way. The first level of their framework involves using descriptive statistics and graphs to try to find noticeable differences between the scores, followed by using a hypothesis test to determine whether the in-class and online exams had the same average score or were significantly different, and finally a regression analysis to look for significant variables and discrepancies between the R^2 values of regression for the in-class versus the online exam scores. In addition to their three-phase approach, they also include a test for unequal variance and a comparison test as two additional levels that instructors can implement at their own discretion (D'Souza & Siegfeldt, 2017). After providing the complete framework for determining whether or not cheating had taken place, they suggest ways for the instructors to penalize the behavior, affirming that the punishment needs to be high to discourage future violations, and suggest forcing the offending students to drop the course for the current semester.

Based on the presented in these articles, a general consensus appears to exist that the current landscape of measuring student performance in higher education statistics has plenty of room for improvement. Many proposals are being made to change the nature of assessment techniques that are used today. Furthermore, the fact that none of articles discussed in this report offer suggestions for dealing with students who fall on the border of a certain letter grade at the end of a course supports the notion that this instructor's approach is truly novel in nature and that his methods could significantly improve the quality of performance measurement in college-level statistics courses. In terms of curriculum, our opinion that there needs to be a prioritization of promoting conceptual understanding in statistics courses is supported in the literature.

Student Perspectives

One of the co-authors was a student who has taken both the statistics capstone and the time series class under the instructor in the past. A student's perspective on the instructor's approach to teaching and grading is given in the following comprehensive list:

- Being told on the first day of class that the syllabus grade would not necessarily be your final course grade and that there would be certain indicators throughout the course that might boast the grade up. This could be a little intimidating at first, but it helped students to not so much to focus on the scores for a high grade but learning and mastering the material
- Having each lecture recorded online made it easier for the student to focus more on the conceptual content during lectures, knowing that you could always go back to the recorded lectures to watch the software demonstrations to see which commands to perform.
- The grace period of one extra class after the "official" due date to turn in homework assignments was interesting. Some students would wait to turn in assignments on the grace period date, but others would plan to turn in the work on the official due date and treat the grace period as it was intended to be – an extra couple of days to work in case the assignment took longer than expected or in case of a busy week in other classes. Many good students found that they stayed on top of the material much better and had an easier time completing the assignment, since the appropriate lecture was fresh on the mind.
- Being told that the lowest one or two homework grades would be dropped made doing homework less stressful. Most students tried not to skip any assignments so that any homework assignments where a student did not perform well could be

dropped. It was nice to have that policy in place in case one was really pressed for time that week.

- Homework questions emphasizing interpretation rather than just asking for computer output helped students form a more thorough grasp of the material instead of just memorizing without understanding why.
- Short attendance quizzes in the capstone were helpful to learning in the long run and in paying attention in each class. They always came at the end of the class and were always unannounced
- The projects in both classes helped the student understand the material better and to prepare for the final exams since the projects required students to apply essentially all of the statistical and forecasting methods learned in class to real datasets. Interpretations and implications for management were required.
- Review sessions with TAs were helpful since one could ask questions about certain topics or homework problems. The TAs were always former students.
- The TAs were available during the week with office hours, prior to class, for help sessions prior to exams, and usually by email.
- The take-home exams were posted online after class on Thursdays to allow the students to have more time during the weekend instead of trying to find time to finish them during busy weekdays.
- Having the take-home exam due on Monday before the Tuesday in-class exam helped students perform better on the in-class exam. Completing the take-home exam essentially functioned as time studying for the in-class exam.
- Conceptual in-class exams were intimidating and challenging, but very beneficial in learning the material for the courses. There was no need of memorizing procedures or methodologies for problems. One had to focus on concepts not numbers to answer questions presented in situations that had not seen before.
- Increasing the percentage of the in-class exam that was made up of conceptual questions for each subsequent exam in the capstone was helpful in that it allowed one to get gradually accustomed to the kind of thinking required to answer the conceptual questions without being too overwhelmed right at the beginning of the course.
- Making both in-class exams in time series entirely conceptual is the biggest reason why most students feel that they learned and retained more from that course than any other statistics course they have taken. It really forced students to thoroughly know the concepts and content of the course so that one would be able to think critically and apply the knowledge to the conceptually framed questions that presented new situations and problems.
- Having in-class exams occasionally give some computer output and ask students to interpret it helped these students to understand the purpose of running certain tests or analyses in the statistical software used in the courses.
- In-class exams frequently included a section or two on a certain type of questions that was missed on a previous exam. This was beneficial to students in learning from past mistakes and helping students retain past course material better.
- Going over the graded exams in class was helpful to see what mistakes were made and what the correct answers were supposed to be.
- There was a policy of being willing to go back and change grades if grading mistakes were made, which assures each student that they would get the grade they deserved.

Cheating

It is most likely that cheating in upper-level courses is most common in the form of prohibited collaboration with other students. This cheating is more likely to happen in courses where students are aware of their current standing and know the exact amount of points they need to earn on remaining assessments in order to pass the course or obtain their desired letter grade. For this reason, students in this instructor's courses might be more deterred from cheating, where final scores are not necessarily based on the grading rubric in the syllabus but on the statistical analysis where certain indicator variables are used to help determine grades.

While D'Souza and Siegfeldt (2017) admit that their method has shortcomings when dealing with situations in which their tests do not clearly point in one direction or the other. Another possible limitation with this framework could be its reliance on a having a baseline in-class exam score to use to compare to the corresponding online exam score. The authors do not address how to proceed when there is not a baseline score for comparison as there was in the study or whether their approach could be used in different situations, such as looking for cheating on homework assignments. While D'Souza and Siegfeldt's method can work as a viable way to find and address possible cheating situations, a simpler and potentially more effective approach can be used.

One of the other topics we examine in this research is trying to find the best way to identify cases of cheating, which is especially important in classes that involve take-home exams. This instructor takes precautionary steps to limit students' cheating capabilities on take-home exams by giving each student a different dataset, which is generated by his or her student ID number, so that no two students will have the same exact answer. D'Souza and Siegfeldt (2017) offer a metric for identifying cheating on specific cases with exams but do not discuss a general framework to use on other types of assessment. A simple approach that this instructor uses for trying to check for potential cases of cheating on homework is to take a median of scores including each exam score, project grade, homework average, and the overall conceptual exam score. An average homework grade that is much greater than this median value could indicate possible cheating. While this approach does not have a highly sophisticated nature and does not involve any "true" statistical analysis such as hypothesis testing or regression, it is a simple technique that instructors could potentially use as a starting point to see if further investigation may be needed in cases of suspected cheating.

Assessment Measures

Typical measures of assessments include homework, quizzes, exams, projects, attendance quizzes, presentations, and maybe more over a semester. The projects can be individual or group or in-class or out-of-class. The homework can be on-line or off-line in paper form. In many of our undergraduate classes, some of us give short attendance quizzes that track with class material and that are easy to grade. These are frequently grouped with homework so that there is freedom to drop a low homework or quiz or missed assignment when computing final grades. Some use point values for these different assessments but one of us likes everything converted to percentiles so that many different grading schemes can be assessed.

Homework Assessments

Homework can be on-line and off-line with hard paper copies. However, research into cheating has shown a higher incidence of such with on-line and take-home assessments. There can be a lot of homework assignments (usually almost 25) for a senior capstone course in statistics while less than half as many for a senior course in applied time series.

Usually, a large introductory statistics class will have at least 25 assignments with most being on-line. In grading the homework, there is an emphasis on the techniques or methods but also assumptions and interpretations. Having taught for many years, it gets tedious in dealing with all of the excuses for not getting the homework in on time. Most of the time, there is a grace period for the homework. The homework is due the next class period but can still be turned in two class periods later. It is not accepted after that. Sometimes, students will begin turning homework in only at the grace period deadline so the grace period is suspended until all have caught up. This suspension always comes before an exam so that all homework appropriate to the coming exam has been turned in and hopefully graded and returned to the students for their exam. The grace period on homework does prevent working homework problems in class or delays it at least a week. But help sessions and office hours by the TA and instructor are an attempt to avoid this. It is really hard to grade late homework and be fair. Sometimes, there are extenuating circumstances, such as sickness, death in the family, an accident, or whatever, which take special treatment.

Almost all homework assignments are graded on a 10-point basis, but only select problems in the homework assignment are graded. For instance, an assignment with eight problems would only require the teaching assistant to grade 4 out of the eight so as to save the TAs time. It is easier to be diligent on fewer homework problems. There is a specific focus on problems that require interpretations and require dealing with assumptions.

Class Options

Over the last five years, we have been able to record our statistics classes so that the students can go back and watch the recording many times. Recently, one of us had to teach in a non-recordable room, and the anguish of the students was very great. Lots of TA and instructor office hours in this situation. The students really needed the recording when dealing with software issues and more. High tech classrooms open up options for videos too. The format of our classes is traditional lecture with an occasional flipped class. Of course, there is an emphasis on computing by hand and by software but also on assumptions and interpretations. Due to time in industry, there are always stories to tell about getting things done in an understandable way for a client or your boss. Soft skills, like writing, presentations, and team dynamics, are illustrated as well. A number of conceptual situations are posed, and students as small teams are asked to make suggestions. For instance, why might outliers in a data set be the most important observations in the whole analysis? In addition, if attendance is waning or the students are not paying close attention, an attendance quiz is given, which is added to the homework grades and increases the options of dropping low scores. Attendance quizzes are always conceptual or recognition of the type of probability distribution or analysis to be used.

Exams

With a junior level or higher course, we utilize exams with two parts: a take-home and an in-class. For the take-home, the student can use a calculator or software or both, but interpretations and assessment of assumptions are critical. However, sometimes on take-homes there are things we want to see by hand too. For instance, almost all statistical software assumes the sample is from an infinite population but frequently in the corporate work it may be a finite population where the finite population correction factor is needed. Those types of computations are minor but need to be shown. To avoid cheating (recent research has shown that cheating is much more common with on-line and take-home assessments (D'Souza and Siegfeldt (2017))), the students each have a different data set. The students ID is broken up into two digit numbers, where simulation creates up to three different data values according to an least two distributions. For instance, if the student ID was 123456, the 12, the 34, the 56 or other combinations would be the cumulative probability for a normal, exponential, chi-square or other distribution. The basic structure of everyone's data is very similar, but it is easy to compare means to make sure that everyone's data and interpretation is slightly different.

However, the in-class exam is conceptual or maybe the interpretation of parts of a computer output as well where most of the analysis is shown, not necessarily right or totally wrong. The take-home exam is handed out on a Thursday typically and due on a Monday at some unusual time, like 2:17 or 3:53, so that they will remember the time even though it is posted. While consulting in industry, these unusual times were easier to remember than 10:00am or 2:00pm. A day later, the in-class exam is given. It is hoped with the unique design of the take-home that the students will be better prepared for the conceptual exam. The conceptual exam portion is designed to easily be finished within 45 minutes. If we really want students to know the conceptual material, then there is increasing value on the conceptual material with each exam. For instance, the conceptual weight for a first exam might be 40%, second exam 60% and on the final maybe 80%.

The only negative to the conceptual exam is the grading. The instructor doesn't hand this grading off to a TA unless the TA has graded for the instructor at least a year. Thus, the instructor generally grades the conceptual. The conceptual questions are usually very short answers but no true and false or multiple choice type questions. One wants to see how the student expresses the concept issues on their own with brevity and clarity.

In the recording of the scores for an exam, there are three data points: a score on the conceptual, a score on the analysis and interpretation, and the total score on the whole exam. In addition, all the conceptual scores across all individual exams are totaled and converted to a percentage. So, two exams and a final would create ten bits of information.

Projects

There are always projects in these junior/senior level statistics courses. There are no team projects because the jobs these students take need someone semi-self-sufficient in time series, and everyone has different data for their project. For instance, one project used frequently is a team collection of stopping times at a four-way stop sign, which provides data for three projects: left versus a right turn, left turn versus straight and right turn versus straight. This data is rarely normal and has some unique challenges. In time series, each student not only has their own data, but they are required to submit a written report and to give a five-minute presentation With 30 to 40 students in a time series class, this requires two to three days of class time. Sad to say, this is usually the only presentation in our statistics program. Their presentation is to be geared for management with one statistician in the audience. The projects are graded by a TA with a tight grading sheet as to what is expected.

Grading

As noted, the conceptual portions of the exams are always graded by the instructor. The homework, the problem-oriented take-homes, the attendance quizzes, and the projects are graded by the TA. As to homework, a few are graded and a simple key is created; but only a few of the homework problems are graded to save the TA time. The attendance quizzes take about an hour to grade and record, and usually about five a semester. For the take-home exams, a few are graded by the instructor as to good students and C-type students. When the take-home exams are graded and returned, the instructor takes a random sample of the graded take-home exams to see if there is consistency in the grading. If not, the inconsistency is usually confined to an area of one or two exam problems. The TA then re-grades these few where scores can be changed. For projects, the same strategy is used even though everyone has different data, but the exact key is very exact as to what is wanted and to order of presentation.

A query on anything that the TA has done always comes back to the instructor. However, there are students that try to hustle every point taken off an exam. If a legitimate mistake in scoring as to adding up points, there is no problem. I tell my students upfront, if there is concern about the grading, then the instructor has the right to re-grade the entire exam. The usual result is no grade change or a lower grade. When the students begin to understand that their exams have been graded graciously and that they have gotten benefits of doubt on some questions or that they have not gotten slammed on grading, they relax and try to study more. For instance, if a student leaves a question blank, they tend to lose all of the points; but some answer gains points. The instructor reminds the students that scores are as to the syllabus, but the instructor takes another look at the data with multivariate methods and a point here or there is not going to make any difference in their grade.

The culmination of the grading is a percentile for homework (inclusive of attendance quizzes), for project, for each exam inclusive of the final, and for the conceptual only across all exams. Each exam is broken down into points for the conceptual and the problems. Thus for two exams and a final, the following scores would be recorded: (1) homework minus lowest two scores, (2) project, (3) conceptual and problem separated for each exam grade (two scores out of each exam), (4) exam grade (assuming two exams and a final), and (5) total percentile for conceptual across all exams. This would entail nine bits of information. At least five to six bits of this information will be used to figure the syllabus grade and the multivariate course grade.

The Multivariate Analysis Options

The multivariate options are many with a regular principal component analysis (PCA) or a robust principal component analysis (ROPCA). In difficult cases, one could follow this with a fuzzy clustering analysis to see which grade category a student really should be placed in. However, in this case, things are kept simple or practical with a robust PCA.

Principal component analysis (PCA) constructs new variables that are linear combinations of the original variables that have maximum-ordered variances (i.e, the first principal component has the largest variance, the second component has the second largest variance, and so forth), and are independent of one another as result of the eigenvalue and vector analysis. Once these transformed variables are created, they can be used as a distance measure for future analysis. The PCA solution is unique, but there are three issues that could affect that solution: a) whether to conduct a PCA on the variance-covariance or correlation matrix; b) whether to rotate the principal components or not and; c) how to account for the presence of outliers or influential observations.

First, principal component analysis is usually done using a correlation matrix if there are major differences in variability among variables (as one might expect when considering grade variables) (Jackson, 2003). The correlation matrix is also preferable when there may be an occasional binary variable in the subset subset indicating for example, missing a key assignment or not getting an assignment in on time.

Second, a choice has to be made whether to rotate the principal components or not. When interpretation of the individual principal components or loadings is important, rotations help; but one gives up the maximum-ordered variance characteristics. On the other hand, if the interpretation is not important, and if the PCA is a preliminary data analysis step to another approach, then non-rotation is preferred which was the authors' choice in this research.

Third, when examining grade variables, it is expected that one would encounter outliers that will impact estimates of the mean, variances, covariances, and correlations. In light of this inevitable dilemma, robust PCA is proposed. There is a considerable amount of research being conducted in robust PCA (see Hubert, Rousseeuw, and Vanden Branden, 2005) using different robust strategies (i.e., projection pursuit, etc.) but in this research the robust estimates of the correlation matrix and mean vector are weighted inversely proportional to the outlying-ness of the observations as indicated by the following equation:

$$dist_i = \{(z_i - a)' R^{*-1} (z_i - a)\}^{.5} \qquad \text{i = 1 to n observations} \qquad (1)$$

where z_i is the standardized observation vector, **a** is the robust estimate of the mean vector, and $\mathbf{R}^*$ is the robust estimate of the correlation matrix based on the EM algorithm (Dempster, Laird, Rubin, 1977), which provides a maximum likelihood estimate (Little and Rubin, 1987). The weights are 1/dist_i and have the same form as the inverse of Hotelling's T^2 for individual observation vectors. It takes an iterative process to estimate the weights (i.e., the higher the distance measure dist_i the lower the weight given to the observation); and **a** and $\mathbf{R}^*$ may change with each iteration as a result of outliers and their masking impacts. Various functions (see Jackson, 2003 for a quick overview of weights) of the above distance will down-weigh outliers rather than trim them. The mathematics of the PCA is to find a particular orthogonal matrix **P** such that $\mathbf{P}'\mathbf{R}^*\mathbf{P} = \mathbf{L}$, where $\mathbf{R}^*$ is the robust estimate of the correlation matrix and the diagonal elements of **L** are the eigenvalues or maximum-ordered variances of the new transformed variables, $\mathbf{y}=\mathbf{P}'(\mathbf{z}-\mathbf{a})$ that represent the PCA scores. Of course, y_1 is uncorrelated with y_2, …, y_p, y_2 is uncorrelated with y_3, …, y_p, and so forth. y_1 has the maximum variance, y_2 has the second largest variance, and so forth. These PCA scores are used in graphs or as inputs for a hard or fuzzy clustering algorithm. One can make this as difficult as one wants but simplicity seems to be the best option for real time. One of the authors has done a lot of multivariate strategies with corporate data.

As to the number of principal components to retain, Kaiser (1960) proposed dropping factors whose eigenvalues are less than one, since these may provide less information. On the other hand, Jolliffe (1972) felt that Kaiser's criterion was too large. His suggestion was using a cutoff on the eigenvalues of 0.7 when correlation matrices are analyzed. Other authors note that if the largest eigenvalue is close to one, then holding to a cutoff of one may cause useful factors to be dropped. However, if the largest factors are several times larger than one, then those near one may be reasonably dropped. In analyzing grades data, most of the time the third component had an eigenvalue less than 0.7. This greatly facilitates the graphic presentation to a two-dimensional graph.

Case 1: An Applied Time Series Class

In this undergraduate applied time series class, there were 31 students with two incompletes where one eventually finished the course. The scores used in the robust PCA were homework average, exam one, project (an average of report score and presentation), and the final. The factor loadings are show in Table 1 below with everything highly correlated with the first PCA and basically the percentage score on the final and the project percentage with the second component.

Table 1. Factor Loadings for Time Series

	Factors	
Variables	**Factor1**	**Factor2**
exam1	-0.846661	-0.145953
perfin	-0.748750	-0.592055
hwave	-0.811112	0.193387
projx	-0.765252	0.535792
eigenvalue	2.70	0.70
% of variation	63.0	17.40

Table 2 shows the four key scores for the time series class, the sorted syllabus weighed average, the conceptual percentage across both exams, and the initial letter grades. When there were not pluses and minuses with grades, grading by syllabus was easier. This multivariate approach was necessary to deal with the close calls. Figure 1 shows the initial robust PCA on the four key scores. There are several unusual observations flagged. First, there is a C+ student in the mass of the B students who is student ID 30 on row three in Table 2. Notice that this student has a syllabus average of 75 and a very low homework average at 32. However, the median score of the four key scores is 84.75 implying a B grade. When one looks at the conceptual average, it is a high B. This student understands the conceptual issues just didn't take care of the homework because of taking too many classes or outside activities. In actuality it was both. She was moved to a B since that is where the student fell by grouping. The second student of interest was ID 5 or the student in the first row of Table 2. Again, the syllabus average showed at best a C- and a median score of 71.5. This student too had a poor homework average. His circumstances too were unusual in that it was his second time thru the course, and the student worked about 30 hours a week so that he could pay for his college education.

Table 2. **Time Series Class Scores**

Row	**ID**	**exam1**	**perfin**	**projx**	**hwave**	**wtave**	**concept**	**grade**
1	5	67.0	76.0	77.5	47.7	68.1	75.4	C- to C
2	2272.0	84.7	88.0	45.5	73.9	82.3	C+	
3	30	80.0	90.3	89.5	32.3	75.0	89.2	C+ to B
4	3	74.0	87.0	85.5	49.1	75.4	81.5	B-
5	29	64.0	79.3	90.0	69.1	76.0	76.2	C+
6	20	74.0	86.0	87.0	52.7	76.2	78.5	B-
7	15	66.0	81.3	90.0	68.2	76.9	76.9	C+
8	16	73.0	89.3	86.0	58.2	78.0	84.6	B
9	18	68.0	93.7	86.0	59.1	78.6	87.7	B
10	12	88.5	86.7	86.0	50.0	78.8	83.8	B

11	2	88.0	86.0	86.5	55.5	79.8	82.3	B
12	24	73.0	86.3	80.5	79.1	80.5	80.0	B
13	27	71.0	84.7	85.5	79.1	80.6	79.2	B
14	11	81.0	93.3	82.0	60.0	80.7	90.0	B
15	10	69.0	91.3	82.0	78.2	81.4	82.3	B
16	21	81.0	86.7	87.5	68.2	81.5	78.5	B
17	19	79.0	85.3	89.0	77.3	83.0	84.6	B
18	13	82.0	88.7	90.5	71.8	83.9	88.5	B
19	14	78.0	87.7	94.0	79.1	85.0	82.3	B
20	6	79.0	90.3	88.0	80.9	85.2	88.5	B+
22	28	76.0	95.0	87.5	86.4	87.2	90.0	B to B+
23	8	82.0	88.7	90.0	94.5	88.8	83.8	A-
24	7	92.0	94.7	84.0	83.6	89.3	94.6	A-
25	23	85.0	91.7	94.0	86.4	89.5	86.2	A-
26	25	87.0	92.0	92.0	92.7	91.1	90.0	A
27	4	90.0	89.7	92.5	95.5	91.7	86.9	A
28	26	88.0	92.7	95.0	97.3	93.2	91.5	A
29	9	94.5	94.7	96.0	95.5	95.1	93.1	A
30	31	98.5	93.0	94.5	99.1	95.9	91.9	A

The grade of C enabled him to finally graduate. The final borderline student was ID 28 or row 22 in Table 2. The syllabus average was 87.2, a B by letter grade. This student did have one of the three highest grades on the final but botched the first exam and did not do as well on her project. Her median score was 86.9 so the grade of a B seems fair. However, her conceptual score was a 90, and in Figure 1 she fell close to a B+ or A-. The B+ seems to be a fairer grade

The point to be made is that there were three special cases! The multivariate analysis found them. Our experience has indicated these special cases can be anywhere from 10 to 20 percent in a class. In this case, there was a story behind each special case but that is not necessary. Those who have struggled with assigning grades in such cases should know that having a robust multivariate analysis brings clarity for difficult or close calls.

Figure 1. Robust PCA Scores for Time Series Class (Initial)

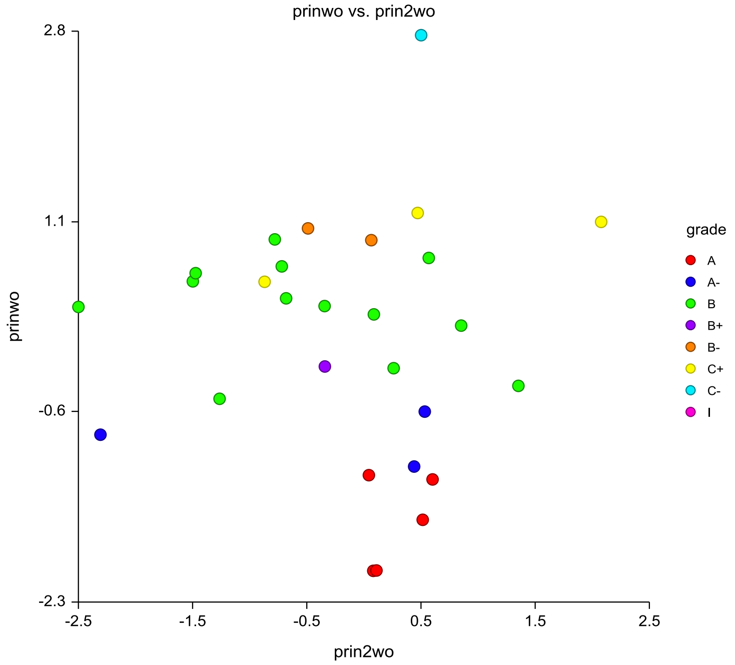
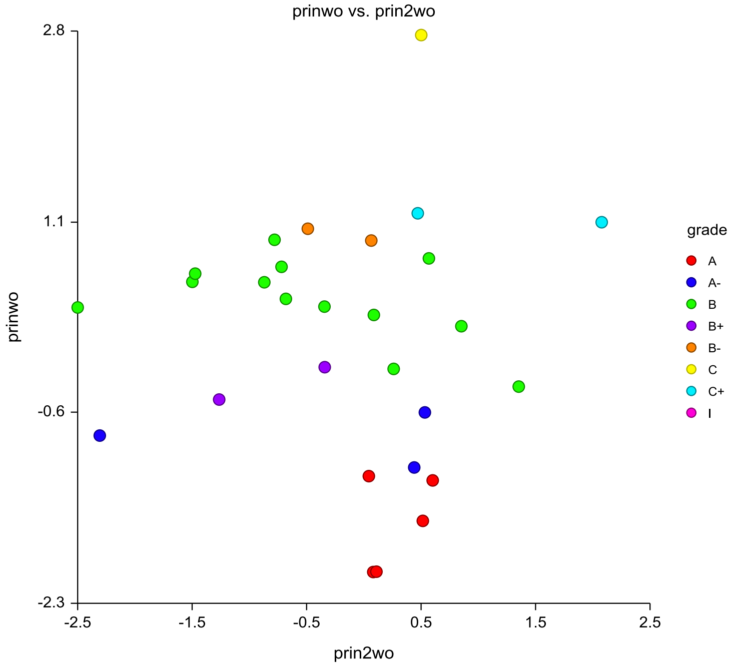


Figure 2. Robust PCA Scores for Time Series Class (Final version)



Case 2: The Capstone Class

In this capstone course, there are three exams, one project, and close to 25 homework assignments throughout the semester that receive a combined average score when final grades are determined. The following table summarizes these scores for all 39 students in this particular section, along with their overall conceptual score for in-class exams and their weighted average for the course.

Assignment	Mean	Median	Standard Deviation	Min	Max
Exam 1	82.51	84	7.70	65	94
Exam 2	78.74	81	8.39	52	94
Exam 3	75.93	75.67	7.18	64	92.67
Homework Avg	83.91	86.95	12.69	51.67	98.61
Project	90.36	90	7.04	66	100
Conceptual	72.71	72.63	8.01	56.84	89.47
Weighted Avg	80.85	81.07	6.64	62.77	93.19

As can be seen in the above results, the grades in this class have a decent amount of variability, with the highest weighted average being over 30 points more than the lowest. An interesting trend that is evident from this summary table is the drop in the average exam score that takes place from exam 1 to exam 3. Although this may appear to be opposite of what would be expected to occur with students usually scoring better on exams after they become more accustomed to the material and types of questions being asked, this result is likely due to the fact that the instructor increases the portion of each in-class exam that is made up of conceptual questions in each subsequent exam. The capstone class usually acts as the first time that statistics and business analytics students get exposed to this type of thought-provoking questioning, and the results show.

The robust principal components analysis for this class includes the variables of exam 1, exam 2, exam 3, homework average, and project grade. The eigenvalues and percentage of variation for each principal component is given below.

Number	Eigenvalue	Individual Percent	Cumulative Percent
1	2.940	58.81	58.81
2	0.872	17.43	76.24
3	0.626	12.53	88.77
4	0.345	6.90	95.66
5	0.217	4.34	100

Based on the values of the above eigenvalues, two principal components appear to be appropriate in this case, as they explain over three-quarters of the total variance. Adding a third principal component could be justified, as its eigenvalue falls less than 0.07 away from the 0.7 value for the Kaiser rule and it would add an additional 12.5 percent of the total variation, but using three components would make the plot of factor scores grouped by grade three-dimensional, which is more difficult to visualize and represent. Therefore, two principal components are used for this capstone class.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Exam 1	-0.761	0.275	-0.503	0.254	0.164
Exam 2	-0.913	-0.037	0.058	0.144	-0.375
Project	-0.713	-0.526	0.377	0.177	0.204
Homework Avg	-0.584	0.678	0.421	-0.121	0.081
Exam 3	-0.824	-0.239	-0.224	-0.462	0.030

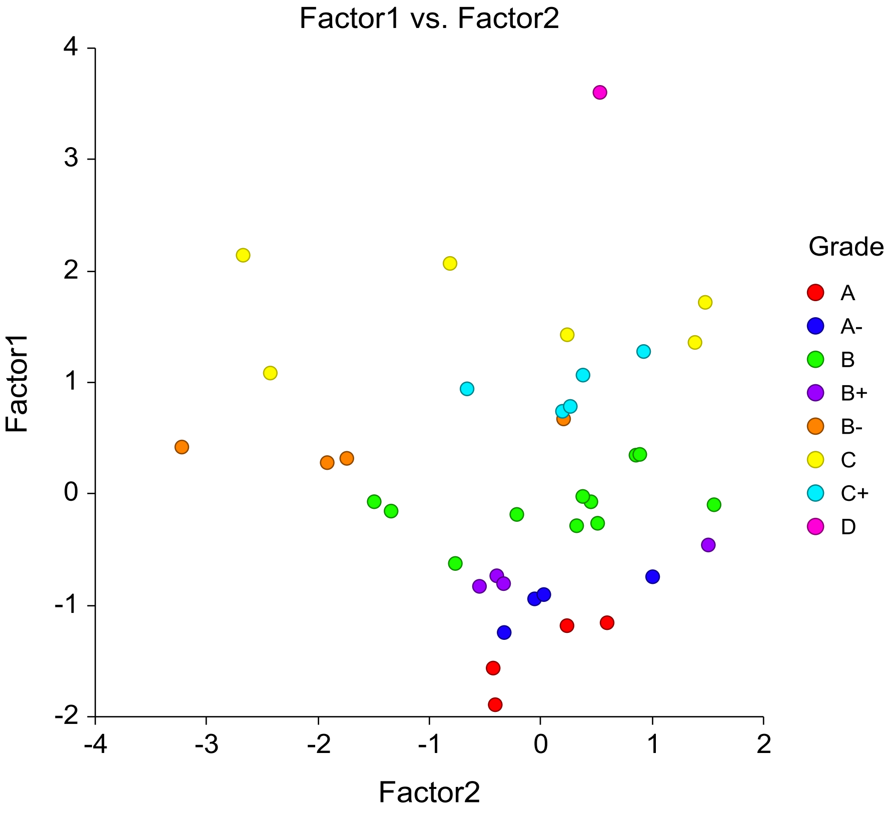


Figure 3: Robust PCA Scores for the Capstone Class

The green table on the previous page gives the values for the factor loadings, with the focus going to the first two loadings, since the first two principal components account for the majority of the variation. From the values of this table, the three exams stand out as being especially important in the first principal component, with homework average and project grade showing up in the second component. This suggests that the exams, especially exam 2, are the most significant variables in differentiating the students of this class. From the scatterplot of the first two factor scores grouped by grade (Figure 3), a few observations stand out as not appearing to fall in the same area of the graph as the rest of the observations in their grade grouping. Some of the observations that the instructor would have to give a second look to include the student with an A- who gets placed in the same cluster as students with A's and the B student who appears to be more closely related to the group of B+ students than the B students. The principal components analysis works to provide the instructor with an idea of which students appear to fall on the fringe of a particular letter grade, and he ultimately determines whether or not to boost their grade by looking back at how they improved throughout the course and considering special circumstances if any are applicable.

Case 3: A Large Introductory Statistics Class

This data was from three sections taught by one of the co-authors. There were 338 students from a variety of different disciplines. But the publisher software did not permit us to break this down by class. Robust PCA was done again, but the plots were not clear enough for details. We did bootstrapping to develop four proxy classes and analyze them. There were 12 scores coming from on-line quizzes, projects, pivot charts, in-class quizzes, clicker quality, and three exams. One-third of the 338 students were identified as outliers by on-line quizzes, clicker points and pivot charts (measures that could be cheated on). The robust PCA was done again, eliminating these measures. However, there were still a lot of outliers, around 20%. It will take some other approaches to get this solved, but the first is getting the publisher to extract the data by class.

Conclusions

Many creative options were suggested in this research, take-home exams with different data for each student, in-class conceptual exams with increasing weight throughout the semester or quarter, conceptual exams with no numbers or computations, robust PCA on scores to identify wrongly assigned grades, conceptual grade for the class, and more. In today’s classroom, instructors must think about how we score knowledge learned. Failure to ignore the possibilities of cheating and to teach concepts hurts the student, the employer, and the academic institution’s reputation. Finally, our best students are always the students that have the best conceptual understandings.

References

D’Souza, K. A., & Siegfeldt, D. V. (2017). A conceptual framework for detecting cheating in online and take-home exams. *Decision Sciences Journal of Innovative Education*, *15(4), 370-391.*

Filzmoser, P., & Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta*, *705(1-2)*, 2-14.

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G. W., Chance, B. L., & Holcomb, J. P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, *65(1)*, 1-10.

Green, J. L., & Blankenship, E. E. (2015). Fostering conceptual understanding in mathematical statistics. *The American Statistician*, *69(4)*, 315-325.

Herrera-Restrepo, O., Triantis, K., Seaver, W. L., Paradi, J. C., & Zhu, H. (2016). Banking branch operational performance: A robust multivariate and clustering approach. *Expert Systems with Applications*, *50(1)*, 107-119.

Moore, A. A., & Kaplan, J. J. (2015). Program assessment for an undergraduate statistics major. *The American Statistician*, *69(4)*, 417-424.

Saccenti, E., & Camacho, J. (2015). Determining the number of components in principal components analysis: A comparison of statistical, crossvalidation and approximated methods. *Chemometrics and Intelligent Laboratory Systems*, *149(1)*, 99-116.

Triantis, K., Sarayia, D., & Seaver, B. (2010). Using multivariate methods to incorporate environmental variables for local and global efficiency performance analysis. *Information Systems and Operational Research*, *48(1)*, 39-52.