Does Instructor Matter? Grade Variation Among Math Courses at CSUN, 2005-2014

Bruce E. Shapiro and Carol A. Shubin
Department of Mathematics, California State University, Northridge
18111 Nordhoff Street, Northridge, CA 91330-8313

Data Science and Mathematical Modelling Research Report 15-01. © 2015 All Rights Reserved

Abstract

Grading patterns across the Department of Mathematics at the California State University, Northridge over a ten year period (2005-2014) were examined to determine if (a) there were differences in grading pattern due to instructor rank and (b) inconsistencies in grading within individual classes. Significant difference is grading were found between instructors of different ranks in some classes (Kolmogorov-Smirnov P-value 0.46) but these differences are subtle, in that the shape of the distribution, not its mean or spread, are primarily affected. The variation between instructor is more noticeable when differences between rank are not taken into account, i.e. rank (tenure or adjunct) does not seem to be the primary determining factor in instructor grading pattern. Grading appears to be a highly individualized process. Students know this. Perhaps faculty and administrators should review their own practice in light of this observation.

Background

The California State University, Northridge (CSUN) is a comprehensive university offering 68 bachelors degree, 47 credential, 58 masters programs, and two doctoral programs. Located in the San Fernando Valley of northwestern Los Angeles with over 40,000 students, CSUN is one of the largest single campus universities in the United States. A total of 6814 baccalaureate and 1913 graduate degrees were awarded during the 2012-2013 academic year.

All undergraduates are required to complete at least one quantitative class to graduate, and at CSUN this requirement can only be met by completion of a course in the Mathematics department. Consequently some 10,000 students enroll in college level math courses every year at CSUN. The most popular courses are College Algebra (Math 102) and Introduction to Statistics (140) (for ten-year enrollments, see Table 4, discussed in more detail later in this report). In addition, another 3700 first time freshmen are required to complete one or two terms of developmental mathematics annually, prior to enrolling in any other math class. Developmental math courses must be completed during the first year at CSUN.

Grades in math lag significantly below grades in other courses on campus; in 2013 the average grade among all students enrolled in math class was 2.085, while among all courses on campus it was 2.795. For courses not offered by the math department, the average was even higher, 2.831, a differential of nearly three quarters of a grade point.

This differential alone indicates that either (a) students under-perform in math; (b) instructors under-perform in math; or (c) some combination of the two. Since many students still manage to thrive in the quantitative sciences, we undertook this study to determine if we identify any any of the root causes of grade variation in math classes.

Motivation

College professors typically have received extensive training in their disciplines and usually hold doctorates, but generally have little or no additional pedagogical training. This association between of the doctoral degree as a license to teach stretches back to at least twelfth century Europe, and probably even earlier. In fact, the term for doctor (latin doceo) became associate during the period with the term for license to teach (latin licentia docendi). The first doctoral degrees were probably awarded at the University of Paris in 1150. Thus a doctor was teacher. The more recent association of the term doctor as a medical or surgical professional did occur until nineteenth century in the United States.

The general belief that mastery of content knowledge in a discipline is satisfactory preparation for transferring such mastery to others persists today. In medieval Europe, discipline content was highly guarded by the trade
guilds and transference only occurred between master and apprentice, or between doctor and scholar in the early university. Pedagogy was not an issue; only the fittest needed to survive. It is unlikely that many medieval blacksmiths had issues meeting their statistics, US History, or upper division writing requirements.

So why do we retain this medieval attitude today, “I have a Ph.D, so therefore I know how to teach?” In the present study, we investigated the following questions across a single department at CSUN over a ten year period.

- Does instructor matter? If only content knowledge is important, verifying that instructors have sufficient content knowledge to teach as class should be sufficient qualification.
- Does professional level matter? On the one hand, anecdotal evidence suggests that tenured faculty should be better teachers because of the lack of pressure to publish, while on the other hand, adjuncts should be superior because teaching is the only criterion for their retention.

In terms of the differences between instructors, the questions we asked were:

- Were some instructors particularly hard graders? easy graders? Are the easy or hard graders correlated with professional level?
- How do grading patterns vary within single classes? A grading pattern is much more than a mean and standard deviation. For example, two instructors may have identical mean and standard deviations, but one may choose to give only A, B, C, D and F grades, while the other splits the B grades equally into B+, B, and B-, giving very different histograms.
- How do grading patterns vary across the department? Even though this information is considered public record (for cohorts greater than 10 students) in California and available in various places, most instructors are not aware when their grading patterns significantly deviate from others in their own department or university unless university officials go out of their way to make the statistical summaries available to them. Students, on the other hand, are extremely aware of this information. Sites like http://myedu.com and http://csunportal.com/professor-ratings/ post the actual aggregated grade histograms for each class taught by every instructor on campus. Students check these sites regularly before registering for class.

As discussed in the following sections, we found significant variations in grading patterns across the department.

- In most cases, there was little or no significant difference in terms of grade mean or spread between faculty in different rank (tenure vs. adjunct)
- In several cases, there was a significant difference in grading pattern between between faculty of different rank within specific classes.
- In many classes, there is a significant difference in grading pattern between faculty. It is not unusual, in a single course, for one faculty member, to give primarily D’s and F’s, and for another to give primarily A’s and B’s. There is no way to distinguish between poor pedagogy and poor academic performance in these situations, however, without further information that was not available in the data set.

Methods

Data Source. We examined a data set (Table 1) containing ten years of redacted instructor course grade history. Course grades were given as numerical histograms for each course section taught over a ten year period during the regular sessions (Fall and Spring sessions only) from the Fall 2005 through the Spring 2014 semesters. Each instructor was identified by a unique identifier code.

| Number of Class Sections |  |
|--------------------------|--
| Graded                   | 2922 |
| Credit/No Credit         | 2144 |

| Number of Students       |  |
|--------------------------|--
| Graded                   | 100,377 |
| Credit/No Credit         | 60,438 |

| Unique Instructors       |  |
|--------------------------|--
| Graded Classes           | 213 |
| Credit/No Credit         | 211 |

| Class Size, mean (s.d.)  |  |
|--------------------------|--
| Graded Classes           | 34.5 (24.2) |
| Credit/No Credit         | 28.2 (14.5) |

Table 1: Properties of data set used in the analysis.
categories: A, A-, B+, B-, C+, C, C-, D+, D-, F, CR, NC, I, W, RP, SP, WU; as well as the total number of students in each section.  

Each course section was identified by course number (e.g., Math 140), a unique instructor identifier (so that all the courses taught by instructor John Doe would have the same identifier), and the instructor was identified as either tenured, tenure track, adjunct, or teaching assistant.

We were not given access to the translation between the instructor identifier and the actual names of the instructors. This kept the analysis at least minimally blind, although with some diligence it would have been possible to track down and identify the instructors’ names because the data was identified by course term and name. We made no attempt to track this information down. However, it should be noted that the university provides this information to various data analysis corporations because they are considered California Public Records and generally not considered (by the courts) to violate FERPA when aggregated in cohorts of at least ten students. Such histograms of instructor grade can be found listed by professor name on public sites such as myedu.com and csunportal.com.

Kolmogorov-Smirnov Test. We examined the data for variations between instructor type, e.g., to see if there were different grading patterns between tenure-track and non-tenure-track faculty, using the Kolmogorov-Smirnov Test. (Smirnov, 1948; Kolmogorov, 1933) The Kolmogorov-Smirnov tests compares two samples to determine the likelihood that they are produced by the same probability distribution. A statistic called the Kolmogorov-Smirnov statistic is calculated to measure the distance between two empirically observed cumulative probability distribution and a p-value is calculate to determine if the two CDF’s come from the same distribution. A p-value close to 1 means that there is no significant difference between the observed distributions, and a p-value close to 0 means that there is a significant difference. We used stats.kstest in the Python scipy package (Oliphant et al., 2001) in iPython notebooks (Perez and Granger, 2007) to calculate the Kolmogorov-Smirnov distance and p-value for all math classes with more than 100 aggregate enrollments that were taught by by tenure-track and adjunct faculty over the analysis period.

Figure 2: Comparison of the range of average grades given by professors in math classes, and range of all grades given in all math classes.

Results

Variation in Grade Distribution We first examined the variation in grading pattern of all instructors who taught classes that were awarded A through F grades. Box plots representing the distribution of grades given by each instructor, aggregated over all courses taught by that instructor, were then generated and compared with other instructors. As illustrated in Figure 1, there is a significant variation between instructors: some instructors give primarily high grades, others give primarily low grades, while the vast majority give a range of grades somewhere in-between.

Urban wisdom would say that the high grades (near the top) would correspond primarily to upper division classes for math majors, such as capstone classes, where students typically do well, whereas the lower grades would be expected to correspond to road-blocking courses such as algebra and pre-calculus. However, this conclusion could only be made if the instructors were limited to these specific courses. The data in Figure 1 is aggregated by instructor, and not by course, and has all types of classes mixed together within each box.

To quantify this distribution across the department, we computed the mean of the grades given by each instructor, aggregated across all course taught (the same data plotted in the boxes in Figure 1). We then binned the means into grade point ranges of 0.1 and plotted the histogram shown in Figure 3. The mean (standard deviation) of the instructor averages was 2.24 (0.59). By comparison, the mean (standard deviation) of all grades given to students was 1.97 (1.39) (Figure 2). The corresponding class size distribution for the same data set is shown in Figure 6.
Figure 1: Box plots for each instructor who taught courses with letter grades in the math department during the reporting period. The boxes are sorted by average grade given by instructor. Each box represents one instructor and extends from the 25th to the 75th percentile of grades given. The black line gives the median, and the whiskers extruding beyond each box extend to the highest and lowest grades given. Grades indicated by plus symbols (+) were identified by software as possible statistical outliers.

The best fit normal distribution is also illustrated; this was computed using \texttt{scipy.stats.norm.fit}. The fact that the distribution resembles a normal fairly closely is probably because we are plotting a histogram of sample means (mean instructor grade) and a normal distribution is not unexpected for a large number of sample means (central limit theorem, CLT). If all samples were of identical size then the CLT predicts that the standard deviation of the sample means would approach $\sigma / \sqrt{n} \approx 1.39 / \sqrt{34.5} \approx 0.24$. We are seeing a somewhat wider spread but this could be attributed to the fact that there are only 213 instructors. Furthermore, the histogram does not actually show the distribution of the means of the course sections, but the means of the instructors after they have been aggregated from their individual course sections. Each instructor has an average of $2922/213=13$ course sections in this data.

Table 2 gives a limit on the impact on students as it tells us the number of students from which the averages were drawn. However, care should be taken in interpreting this table. Just because there are 1,726 students in the $<1.0$ grade row, all this means is that the average grade of the classes that those 1,726 students was in was less than 1.0. That does not mean that every student received less than an 1.0.

Table 2: Upper limit of impact on students of variation of mean grade by instructor. The middle column gives the number of students in the sample from which the instructor means in the first column were generated.

<table>
<thead>
<tr>
<th>Range of Grades</th>
<th>Number of Students</th>
<th>Number of Instructors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;1.0$</td>
<td>1,726</td>
<td>4</td>
</tr>
<tr>
<td>1.0 – 1.5</td>
<td>14,646</td>
<td>33</td>
</tr>
<tr>
<td>1.5 – 2.0</td>
<td>23,924</td>
<td>43</td>
</tr>
<tr>
<td>2.0 – 2.5</td>
<td>45,518</td>
<td>71</td>
</tr>
<tr>
<td>2.5 – 3.0</td>
<td>13,947</td>
<td>44</td>
</tr>
<tr>
<td>3.0 – 3.5</td>
<td>1,185</td>
<td>7</td>
</tr>
<tr>
<td>3.5 – 4.0</td>
<td>31</td>
<td>2</td>
</tr>
</tbody>
</table>

Credit/no-credit courses, on the other hand, have only two possible outcomes, and the distribution looks quite different (Table 3). To visualize the variation, grades were assigned a value of either 1 (pass) or 0 (did not pass) and the range of grades was plotted for each instructor. Because of the small number of values, a box and whiskers plot is less meaningful here (the quartiles will all be either 0 or 1), and so the mean standard deviation was used to plot the spread (Figure 4). To visualize the variation over the department, the average for each instructor was computed and histogram
Figure 3: Top: Histogram of average grade given in all class sections over the reporting period. The curve is the equivalent normal distribution. Middle: Actual distribution of grades given in math grades. Bottom: Distribution of sample sizes by instructor. They range in size from 6 to 3911 students, with mean (standard deviation) of 471 (653) and quartiles 65 (25th percentile), 183 (median), 596 (75th percentile).

Figure 4: Distribution of grades in Credit/No Credit classes by instructor. Each vertical bar represents one instructor, showing the average and standard deviation. The ranges are capped at zero and one, where zero represents no credit and one represents credit. The instructors are ordered by average grade given.

1% of the total students failed these classes. It means that fewer than 1% of these students failed these classes in sections that were taught by the 17% of instructors who were especially tough graders. The remainder of students who failed did so in classes led by instructors who were not as tough at grading.)

Table 3: Summary statistics, grades given by instructor for credit/no-credit classes. The first row gives the distribution of the instructor mean grades; the second row summarizes the distribution of the actual grades received by students over the entire sample.

<table>
<thead>
<tr>
<th>Graded Given in CR/NC Classes</th>
<th>mean</th>
<th>std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Grade Given By Instructor</td>
<td>0.70</td>
<td>0.25</td>
</tr>
<tr>
<td>Range of Grades Received By Students</td>
<td>0.71</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Does Instructor Matter? We compared the grading patterns between tenured/tenure-track faculty and non-tenure-track (adjunct) faculty in all class that were taught to aggregated cohorts over 100 students over the reporting period. The grading patterns were compared using the Kolmogorov-Smirnov two-sample statistic. A number of classes showed a significant difference in grading pattern between the two categories of instructor. The results for all classes are summarized in Table 4.
Grade Variation Among Instructors. To visualize grade variation between the different instructors in each class, we constructed histograms of the grade distributions for each class and for each instructor. We then display these histograms together as a heat map, one heat map for each class. A single heat map therefore gives a visual display of the variation of grading style for a given course. We repeated this for all class sections taught by the Math Department (Figures 7 through 10; classes taught by fewer that five instructors are not shown). The names of the corresponding courses are given in Table 5.

In each heat map, the color displays the number of students who were given a particular grade, ranging from low (in red) to high (in green). Since some instructors choose not to use the +(plus) or -(minus) grades, two forms of the heat maps are shown. In Figures 7 and 8, all grades are shown; in Figures 9 and 10, the same data is shown collected into A, B, C, D and F grades.

The histogram for each instructor is then plotted as a single horizontal bar on the plot, with either five (A to F) or twelve (including plus/minus grades) rectangles, each rectangle colored according to the number of A, B, C, D or F grades. The vertical height of the bar is then proportional to the aggregate number of students taught by that instructor in all sections of the given class. The bars are then stacked vertically top to bottom in order of average GPA by instructor.

It can be seen that in some classes there is a significant variation in grading pattern between instructors. Instructors that show a lot of green toward the right and red toward the left are hard graders, while those who show green toward the left and red toward the right are easy graders. Instructors who ignore the “+” and “-” grades stand out clearly (for example, see the heat maps for Math 255A and 255B). Classes like Math 350 (Advanced Calculus) and Math 462 (Advanced Linear

Table 4: Comparison of grading patterns between (tenured & tenure-track) faculty and (adjunct lecturers). a p-value close to 1 indicates a similar grading pattern; a p-value close to 0 indicates a different grading pattern. Courses that were only taught tenured/tenure-track or only taught by adjunct faculty are not shown.
Table 5: Courses examined in heat maps of instructor grade variation in Figures 7 and 8.

Algebra) are perplexing: some instructors have very high failure rates while others have very high pass rates.

The message here is that grading patterns vary significantly by instructor. The impact of each of these instructors on the overall student population is seen by observing the height of the bar. Thicker bars correspond to more students; thus the Math 102 instructor with the large green block on the right has a very heavy impact as he or she had a large number of students.

Does class size matter? The multi-modal plot in Figure 6 raises the intriguing question of whether or not class size is important. While a complete analysis of this question is beyond the scope of the present report, we did examine this question by plotting the grade of all students in the data set (letter-grade classes) against course size and found only a weak negative correlation between grade and course size (correlation coefficient -.095) (Figure 11).

Lessons Learned and Recommendations

We have observed that there is a significant and observable variation between grading of individual instructors even in a single class. We cannot determine from the
present data set if this is due to variability in academic performance among students or variability is pedagogical skills among teachers. While there is a statistically significant difference in grading pattern between instructors of different rank in some classes, it does not occur in other classes. In most cases, however, the difference in grading pattern between individuals, rather than between rank of instructor, is more significant.

High-performing students tend to get priority registration and pick their favorite instructors (e.g., those with a reputation as easy graders). Lower-performing students, such as those on academic probation or who need to repeat a course for credit, must wait until late to enroll in a class, so they get their second or third choice professors. They either end up taking classes with instructors who have a reputation as hard graders, or with “Professor Staff,” the just-in-time instructor assigned to the class (or hired) the week before the semester begins. The lowest-performing students, or the ones who are not cognizant of how the American university system functions (e.g., they don’t know about placement exams) end up going to many classes on the first day and begging instructors to be added. The result is an instruction gap: the best teachers get the best students, the most poorly prepared instructors get the neediest students. The rich get richer and poor get poorer. Is this any way to run a college?

Very little assessment across multiple sections is performed in CSUN math classes. There are common final exams in some of these classes (e.g., Business Math (103), Calculus I (150A), Calculus II (150B), Calculus III (250), and starting in Fall 2015, Introductory statistics (Math 140)). However, common grading is not consistently performed. Even in classes where common grading is performed, there are frequently too many problems for a single grader to grade all of them (sometimes several hundred), and so they are divided up. For example, graders A and B grade problem 1; graders C, D, and grade problem 2; and so forth. This can lead to inconsistent grading styles across the population. Even when a single grader is grading the entire set, it is no inconceivable that the first paper graded may have been graded differently than the 400th exam.

On suggestion to alleviate grading inconsistencies is contextual grading. This technique is most often suggested in the context of grade inflation and not the context of grader inconsistency. The overall level of grade inflation across campus from 2004 to 2013 was 0.007 (s.e. 0.001; correlation 0.85); grade points per year. For math it has been 0.013 (s.e. 0.003; correlation 0.71). In the past, long term grade inflation had accumulated over several decades causing the average student grade nearly a full grade point from a C average in the 1950’s (before CSUN was founded) to a 2.83 (B) average in 2013. If the current linear trend holds, the mean CSUN GPA will reach 3.0 in 2043 and 3.5 in 2114, so while grade inflation still exists, it has mostly self-corrected itself to a new level and (more-or-less) stabilized. It is not possible to determine, in the context of this analysis, whether the faster rate of increase of math grades, is due to better instruction or easier grading. However, it would be of more concern if the trend were in the other direction, i.e., if the math grades were not approaching the campus average.

What is of more concern is the large gap between math grades and other grades on campus. It seem indicative of an institutional failure to provide students with the skills they need to succeed in mathematics. However, it could also indicate overly harsh grading by the math department.

Educational research suggests that mere fact of making this information available and encouraging sharing and collaboration between educators is often enough to improve student performance [Shaun Harper, personal communication, 9 March 2015]. Perhaps we should take some advice from the experts: share statistical summaries with faculty in a non-threatening manner. Individuals many not even know they are grading differently. First time faculty may feel a need to prove that they are “good enough” and push their students too hard. Others may find the transition from dissertation research to teaching “easy” undergraduate material psychologically jarring to the point of boredom. Regular round-table meetings where experienced faculty can share their own pedagogical lessons learned and individual war stories in a comfortable setting may also be a way to help newer faculty adjust and to campus life.

Before we consider forcing specific syllabi, collective examinations, grade scales and teaching patterns on instructors, we should at least share the data (with names withheld) and discuss teaching strategies. Many instructors think there is only one way to teach, the way that they were taught themselves, and were never exposed to other paradigms. Perhaps instead it is better to learn together through growth and collaboration.

Acknowledgment

We would like to acknowledge the help of Harry Hellenbrand; the Office of the Provost at CSUN; the Office of institutional Research at CSUN; and the following individuals who made valuable suggestions during this anal-
IC Incomplete Charged. Reported when a faculty member does not report a grade for an incomplete (I). Counts as an F (failing) grade in the student’s GPA.

RP Report in Progress. Used for multi-term projects, such as a thesis, where a grade has not yet been determined. Does not affect student GPA.

SP Satisfactory Progress. Used for multi-term projects, such as a thesis, where satisfactory progress is being made. Does not affect student GPA.

W Medical Withdrawal. Does not affect student GPA.

WU Unauthorized Withdrawal. Reported when a student stops coming to class or misses a significant amount of material. This grade is completely arbitrary; some instructors choose to give an F, an Incomplete, or a grade for partial work completed in this situation. The grade of WU counts as an F in the computation of the student’s GPA.

CR Credit. Assigned for A, A-, B+, B, B-, C+ or C for undergraduate classes and A, A-, B+ or B for graduated classes. Not included in the GPA.

NC No Credit. Assigned for C-, D+, D, D- or F for undergraduate courses, and to B-, C+, C, C-, D+, D, D- or F graduate classes.

Table 6: Description of administrative grades at CSUN.

Literature Citations


Figure 7: Heat maps show the variation in grading pattern for different classes in the Math department. Each horizontal bar corresponds to a single instructor. The vertical thickness is proportional to the number of students. The color gives the number of grades given. Plus and minus grades are aggregated into the central grade, e.g., B+ and B- are included in the sum for B. The names of the corresponding courses are given in Table 5. Lower division courses are shown here; selected upper division courses are shown in Figure 8. The data plotted here is also shown with the plus and minus grades aggregated into the central grades in Figure 9.
Figure 8: Continuation of heat maps shown in Figure 7, for selected upper division courses offered in the department. Courses taught by fewer than five instructors over the reporting period are not shown. The data plotted here is also shown with the plus and minus grades aggregated into the central grades in Figure 10.
Figure 9: The same data as shown in Figure 7 is shown here, but the plus and minus grades are aggregated into the central grade, e.g., the A also includes the A- grades, the B includes both B- and B+, and so forth.
Figure 10: The same data as shown in Figure 8 is shown here, but the plus and minus grades are aggregated into the central grade, e.g., the A also includes the A- grades, the B includes both B- and B+, and so forth.