Math 1206 Spring 2006 after-action report
Frank Quinn May 20 2006

Math 1206 is the second semester of the first-year science and engineering calculus course at Virginia Tech. In Spring 2006 there were 851 students enrolled in 18 sections. All students took a common final exam. About half the students were enrolled in six computer-tested sections.

The objective in the first part is an overall analysis to assess the quality of the exam; to compare outcomes in different sections; and to assess outcomes on common finals as a measure of teaching effectiveness.

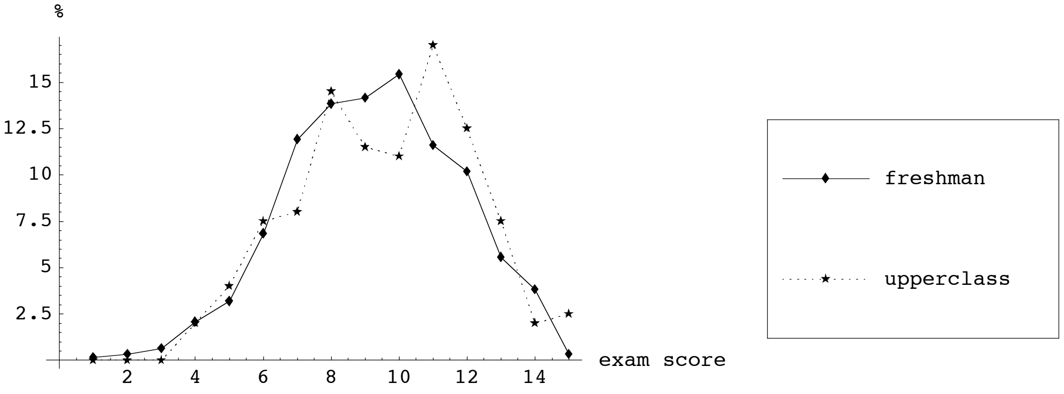
The second part concerns the computer-tested sections: first to compare exam outcomes with traditional sections, then to assess policies and factors that relate to student success.

Exam and dropout analysis

■ Exam overall

■ Outcomes

The following graph shows average grades on the common final (out of 15 possible), with freshman and upper-class separated.



■ Specific problems

Overall the exam was well written and response patterns suggested most problems did probe the target subject as well as can be expected with a multiple-choice format. There were two exceptions:

Form A Problem 5: Find $\int_{-2}^{0} | x + 1 | dx$.

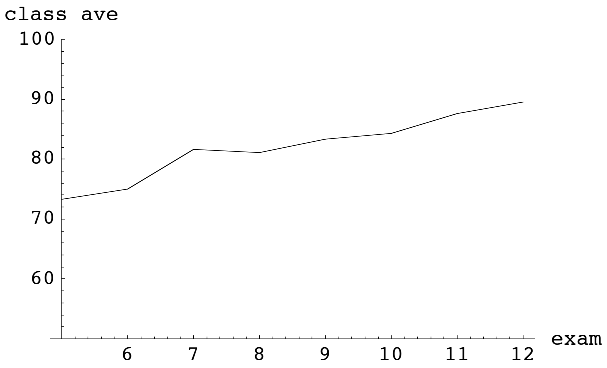
60% of the students apparently misread the absolute value signs as parenthesis or square brackets, perhaps believing this was a symmetry problem, and selected 0 as the answer. 29% selected the correct answer 1, while the other two answers were selected by 4% and 7% respectively. The other piecewise-defined problem, Form A problem 8, had an 88% success rate so students were clearly capable of doing the problem if they read it correctly.

Form A Problem 2: If $F(x) = \int_{0}^{\sqrt{x}} \frac{1}{\sqrt{1+t^4}} dt$, find the value of $F'(1)$.

First, the "prime" on F was unclear and prompted lots of questions. $\frac{dF}{dx}(1)$ might have been clearer. Second 58% chose the incorrect answer $\frac{\sqrt{2}}{2}$ as opposed to 29% the correct answer $\frac{\sqrt{2}}{4}$. It is likely that the mistake was not misuse of the fundamental theorem, but missing the $\frac{1}{2}$ factor from the chain rule. The problem was therefore probably not successful as a test of understanding of the fundamental theorem.

■ Correlation with class averages

The graph below compares exam scores with class averages in two classes totalling 140 students. The two are poorly correlated, and in the interval [7, 10] around the exam average they are nearly independent. Also the distribution of class grades for a fixed exam score is much like the overall distribution shifted a little (ie. not localized near a class grade). Similarly the exam distribution for a fixed class grade is not localized near a corresponding exam score. The weak correlation shown in the graph is not evident in scatter plots, but only shows up at the level of averages.



There are two contributors to this poor correlation: first, cumulative learning at exam time is not well correlated with class scores; and second, a single-administration multiple-choice test is not a great way to measure learning in a calculus course. The weak correlation of learning with class scores is well known; shows up on any kind of exam; and is a major reason to have final exams. However it means class scores cannot be used to evaluate the exam. It seems to be impossible to do this with the data available.

■ Exam by major

Exam performance and dropout rates varied significantly by student major and year. This data can be used to find expected outcomes for each class depending on the major/year profile. Different classes had significantly different major/year profiles so there is significant variation in expected outcomes. Regression analysis compares expected and actual outcomes, and reveals variation not explained by average tendencies.

The Majors subsection describes how majors were collected into groups and gives outcomes for the groups. Other subsections compare expected and observed outcomes in detail.

■ Majors

University Studies and General Engineering provided the bulk of students. Other majors were grouped topically, but performance patterns within a group were reasonably uniform.

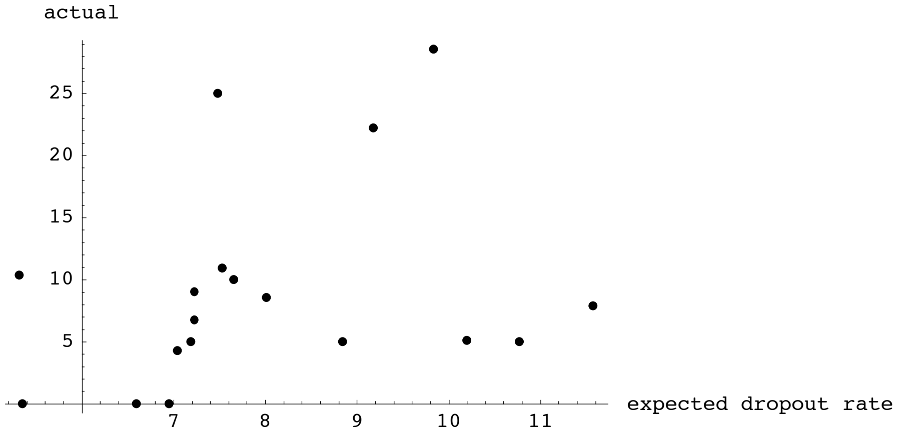
```
Nontech                          Univ Studies                       General Eng
---------                        ---------                          ---------
Business                         University Studies                 General Engineeri
Business Information Tech        Theatre Arts
International Studies            History
Management                       English                            Sciences
Agricultural & Applied Econ                                         ---------
Animal and Poultry Sciences                                         Biochemistry
ApprlHousing                     Other Eng                          Biological Scienc
Hmn Nutrtn Foods & Exercise      ---------                          Chemistry
Wood Sci & Forest Products       Aerospace Engineering              Economics
Architecture                     Biological Systems Engineering     Environmental Sci
Building Construction            Chemical Engineering               Geosciences
Environmental Science            Civil Engineering                  Physics
Enviro Policy & Planning         Computer Engineering               Political Science
Industrial Design                Electrical Engineering             Psychology
Public and Urban Affairs         Engineering Sci & Mechanics        Computer Science
                                 Industrial&Systems Engineering     Mathematics
                                 Mechanical Engineering             Statistics
                                 Mining Engineering
```

Enrollment, dropout rates, and exam averages for groups:

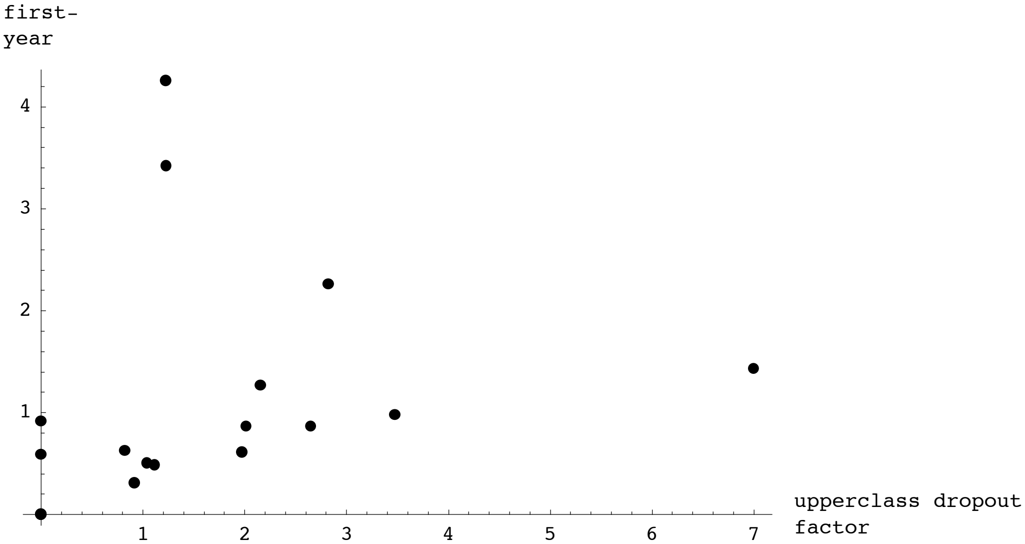
group	frsh	drop	ave	uppr	drop	ave
	enrl	%		enrl	%	
Nontech	33	3.03	8.53	22	18.2	9.11
Univ Studies	163	16.0	9.07	28	7.14	8.54
General Eng	377	7.96	9.18	93	3.23	9.20
Sciences	63	1.59	9.15	36	8.33	10.0
Other Eng	51	0	10.6	33	0	11.3
All	687	8.44	9.24	212	5.66	9.59

■ Dropout rates by class

The following plot shows expected vs actual dropout rates for the 18 classes:

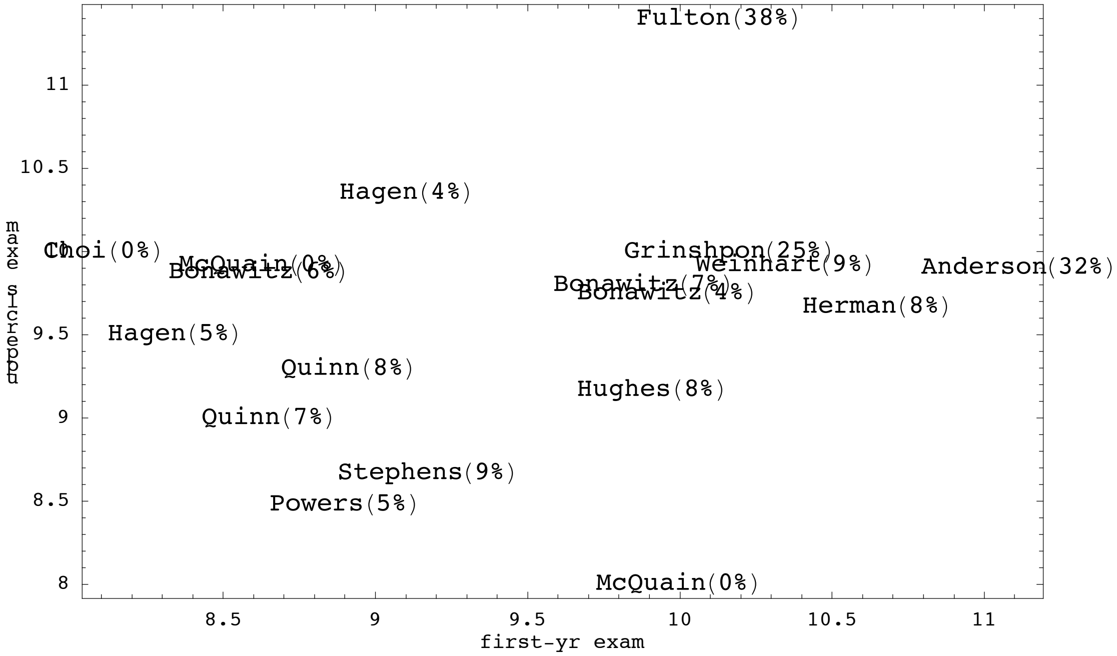


The next plot shows quotients expected/expected, so a quotient 2 corresponds to a dropout rate twice the expected rate. It shows that in classes with high dropout rates the dropouts tended to be concentrated among either first-year students or among upper-class students. In extreme cases it seems reasonable to think this is a consequence of teaching styles.



■ Exam by class

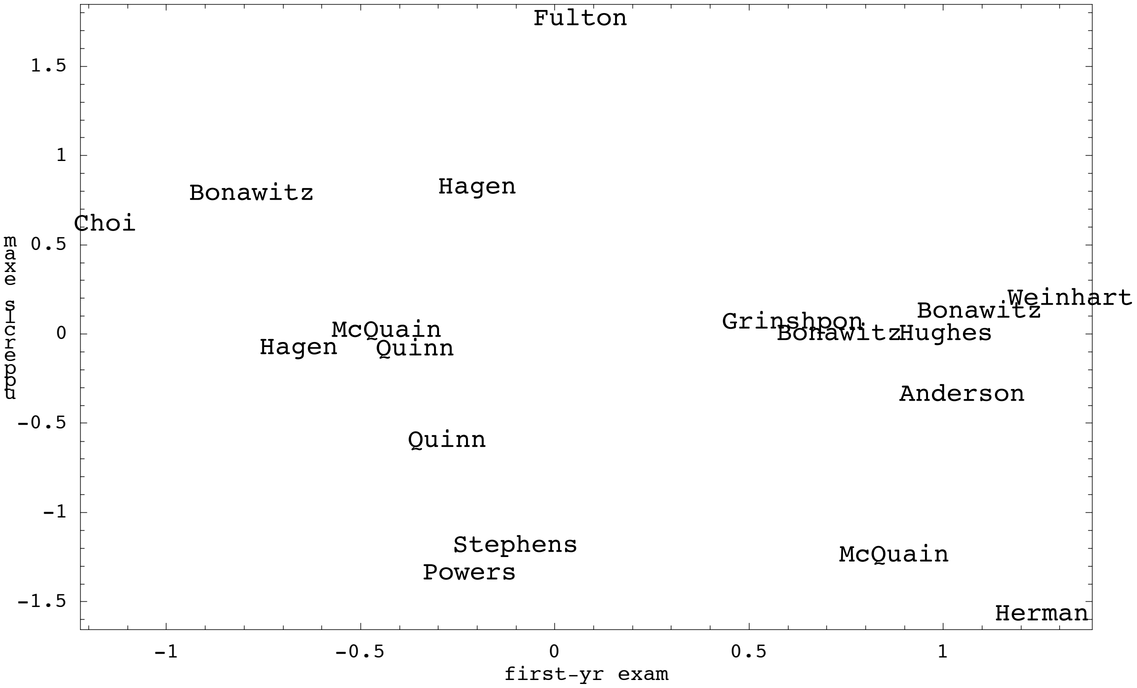
The following plot shows teacher and first-year dropout rate located by first-year exam average (horizontal axis) and upper-class exam average (vertical axis). Overall dropout rates were 8.4% for first-year and 5.7% for upper-class students. A regression-analysis version is given below.



The following table gives the data used in the plot above.

1206	0=comp	firstyear	total	firstyear	upperclass
CRN	1=trad	dropout %	dropout %	exam ave	examave
13519	1	25	22	10.1667	10.
13520	0	5	7	8.90323	8.47619
13522	1	6	5	8.62069	9.875
13523	0	9	11	9.17582	8.66667
13527	0	7	5	8.65385	9.
13530	1	4	5	9.96154	9.75
13531	0	8	9	8.91667	9.29412
13532	1	8	8	9.91304	9.16667
13533	1	37	29	10.1333	11.4
13534	1	9	9	10.35	9.91667
13535	1	7	10	9.88462	9.8
13536	1	8	10	10.6522	9.66667
13538	1	0	0	8.62963	9.91667
13539	1	0	0	8.11111	10.
13540	1	0	0	10.	8.
13541	0	4	4	9.10638	10.35
13542	0	5	5	8.34483	9.5
13543	1	32	25	11.1176	9.9

The next plot locates teachers according to comparison of actual exam outcomes with expected outcomes predicted from major profiles and dropout rates.



Qualitative differences from the raw-data version above are mainly that teachers with high dropout rates come out less well, and there is greater spread between classes with the same teacher. The latter seems larger than might be expected from randomness due to small sample sizes, suggesting that there are systematic effects that are not accessible with this data.

In particular we conclude that common-time exam data is not a reliable indicator of teaching except perhaps in really extreme cases or when there is a consistent pattern in five or more courses.

Computer-tested sections were taught by Hagen, Powers, Quinn, and Stephens. The plot does seem to justify conclusions about the computer sections, and these will be discussed below.

Computer-tested sections

■ Exam outcomes

The final plot in the "Exams by course" section above shows a clear enough pattern to justify conclusions about the computer-tested courses: they seem to work as well as traditional courses for upper-class students, but are less effective for first-year students. All of the computer sections have lower than expected outcomes for freshmen while all but three of the traditional sections have higher than expected outcomes. The erratic nature of the data is emphasized by the fact that two of the lower-than-expected traditional sections were taught by teachers who also had sections with higher-than expected outcomes. Nonetheless it seems reasonable to conclude that, for first-year students at least, the computer-course format is less effective as preparation for the exam.

The testing system design restricted tests to multiple-choice format. There is a general distrust of the effectiveness of such tests at this level and exam data presented above supports this. Further the tests and the course are still very much under development, not finished products. The surprise is not that they are less effective than traditional classes with experienced teachers, but that they are only slightly less effective.

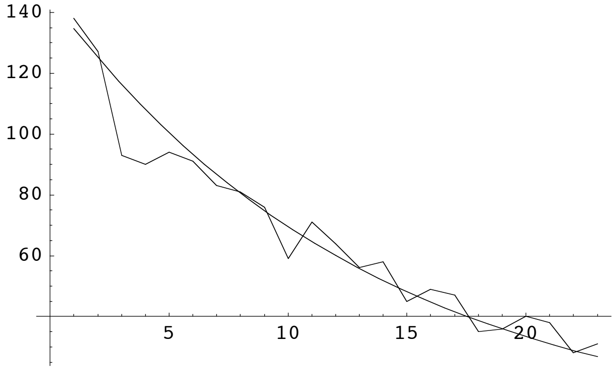
■ Conclusion

The tests currently used are not fully satisfactory but they are close enough that another round of development and refinement should bring them to parity with traditional classrooms.

■ Attendance

The data in this section comes from two classes with no attendance incentives: no in-class tests or quizzes, no homework collected, and while attendance was recorded it did not count in the final grade. The point of this policy was that the tests have feedback and web links so practice versions serve as a crude approximation to an on-line course. There is no point in requiring students who can use them this way to attend classes. The obvious concern is that large numbers of students who can't learn from non-class resources will skip classes anyway. Past experience indicates this does not happen, in the sense that outcomes are almost completely independent of attendance, and this pattern held in this data.

Attendance declined roughly exponentially, to about 20% by the end of the course. (Half-life 9.5 class meetings). The graph shows attendance vs class meeting number.

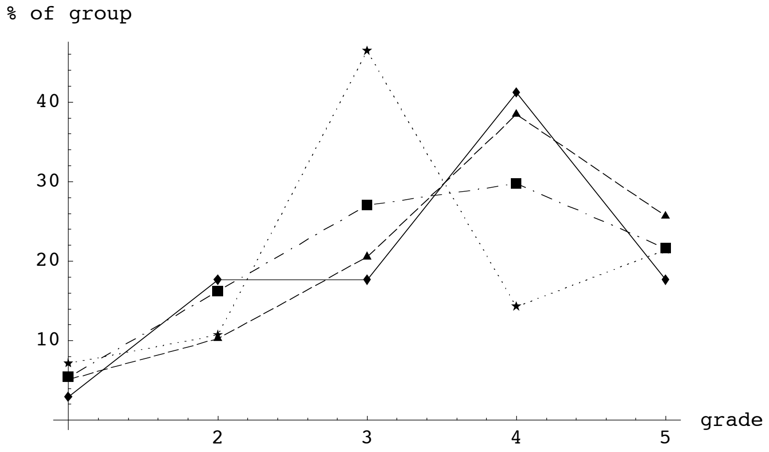


Students were divided into four groups of roughly equal size depending on attendance.

The dropout rate was higher for those attending fewer than 2/3 of classes. However the overall dropout rate was typical for the course; dropouts quit coming just as they do even when there are attendance incentives. Course averages were almost the same for the four groups: the group attending more than 2/3 of classes had averages 2-3% higher. Curiously this group did better in course tests and less well on the final. If the final had been weighted slightly higher (ie not so much curve) then the difference would have disappeared.

attended	1-3	4-7	8-15	16-23
enrolled	40	31	44	40
took exam	34	28	37	39
drop %	15	10	16	2
course average	79.0235	78.2786	78.6865	81.5538

Grade distributions in the very-short and long groups was nearly identical. The 8-15 group has a modest shift from B to C, and the 4-7 group had a pronounced shift. The pronounced shift in 4-7 is largely an artifact due to averages lying close to the B/C cutoff. The graph overstates differences in GPA as well because the scale does not incorporate +/- data; In principle this distortion could be overcome by using finer subdivisions of the data, but in practice the sample size is too small.



■ Conclusion

Attendance requirement or incentives would not improve outcomes in this course. Optional attendance is beneficial to students because they can use resources (including lectures) in ways most efficient for their individual learning styles.

■ Procrastination

Class attendance does not correlate with outcomes, but data mining turned up one thing that does: procrastination. The course is based on multiple-try tests with deadlines, and best score used in the final grade. Scores decline as the first for-credit attempt approaches the deadline. Students can also take or download practice versions for study purposes. Scores do not correlate with when or how many practice versions are used.

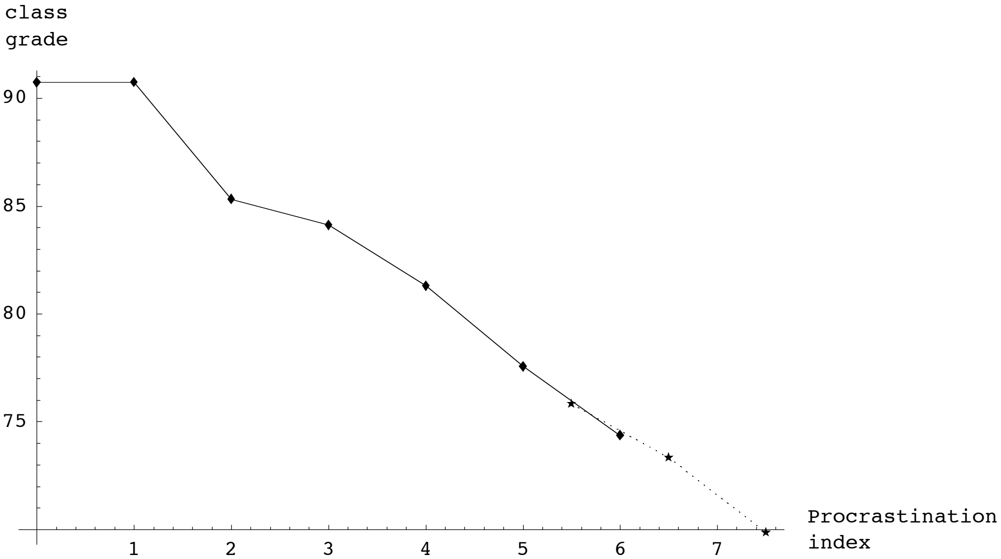
Since procrastination is known to correlate with low scores the test policy was designed to discourage it. Each test could be taken up to five times, but only once on the last day (i.e. people waiting to the last day lost four of their chances).

Average max scores by people waiting until the next-to-last day were 92% of scores of people starting earlier. Scores on the last day were 71% of earlier-takers. Data uses six tests, 470 students. The data uses test taken during the semester. A significant number improved their final score with makeup tests, as discussed in the next section. A regression analysis was done to sharpen the picture.

■ Procrastination index

This is a scale chosen to linearize the score decline. Let p for early, next-to-last day, last day, be 0, 1, 3.5 respectively, then scores are approximated by (early score)(1 - 0.78 p). This is derived using averages over all tests, and individual tests are reasonably close to this pattern. A total procrastination index is obtained by adding procrastination indices for individual tests. For instance a student who waited to the last day once and next-to-last day twice had a total index of 3.5 + 1 + 1 = 5.5.

The data is consistent with the following picture: A subpopulation of about 10% worked effectively under pressure. They went down to the wire on all or all but one test and did well anyway. The other 90% had scores very nearly linear (slope -0.022) in the total procrastination index. The main deviation from this is that there was no consequence for waiting to the next-to-last day *once* but significant consequences for waiting twice or more. The following graph shows the pattern in the 90% subpopulation:



Linearity in the total index means that the procrastination effect is additive: the hit associated with putting off a test does not depend on how many tests were put off. It can therefore be associated with individual tests rather than some sort of behavior pattern.

We caution that this division into subpopulations is somewhat speculative since the 10% that worked well under pressure were not distinguished in any other way in this data set. They had typical patterns in attendance, exam outcomes, and major subject. They were nearly all first-year students but the sample size is too small for this to be significant.

■ Conclusion

For most students procrastination is the only clear correlate with lower performance and it is appropriate to design test policy to discourage it. Having all but one test opportunity expire on the last day is new this semester and has partially alleviated the last-minute problems seen in the past. The data presented here suggests it would appropriate to go further: offer five total tries before the next-to-last day, decreasing to three on the next-to-last day, and to one on the last day. The objective is to communicate roughly what the dangers are in putting it off.

There is a significant population that either already knows the material or works well under pressure, and does not show much disadvantage from late tests. This means it is not appropriate to have penalties or interventions for students who show a procrastination-type pattern early in the semester.

■ Makeup tests

It is necessary to have a mechanism for students to make up missed tests. It is also reasonable to offer an opportunity to fix a bad test score. However satisfactory outcomes seem to depend on the policy.

■ Makeup policy

In previous semesters the policy was that tests would reopen for one additional try during the last week of class. The best of the previous scores and the makeup was used for the grade.

The restriction to one try was intended to discourage extreme procrastination: skipping a test and counting on the makeup. It was successful in this. However the use of the highest of all scores undercut much of the intent. A large number took makeups but most had no improvement and nearly half got lower scores than before. Since it was without risk most students used it as a "free shot" without doing the preparation that would ensure better scores or improve learning.

This semester the policy in two sections (total 138 students) was: they could take up to four makeup tests. These could be multiple tries at a single test, or different tests. However taking a makeup deleted scores from earlier in the semester, so the best *makeup* score was used in the grade. This meant taking a makeup test put them at risk of *lowering* their class averages, so it should only be done if they were pretty certain they could do at least as well as before. This was quite successful: it weeded out frivolous retakes and resulted in a greater improvement in overall class performance than did the old policy.

■ Outcomes

47 students (about 1/3 of the group) took at least one makeup test.

28 of these took one makeup (possibly multiple times). All but one raised their score, with average class grade increasing from 74.5% to 79.0%. About 2/3 of these retakes were of tests originally taken late (last or next-to-last day).

Outcomes for students taking two or three makeups did not depend on how many they took. There were 18 of these (about 13% of the group) ant they took a total of 42 tests. 32 raised the score, 10 did not change the score, and 2 lowered the score. Average class grades increased from 63.9% to 72.8%. Nearly all these retakes were of tests originally taken late.

■ Conclusion

A carefully designed makeup test policy can enable students to fix isolated problems and encourage them to fill gaps in their learning. However the policy must avoid encouraging procrastination or frivolous use. The policy described above seemed to satisfy these criteria.

We remark on allowing multiple retakes of a single test, within the rigid overall limit. These tests were designed to play a large role in the learning process (via practice tests, feedback, etc) rather than only as assessment instruments. This means they have to be harder than traditional tests. Further there is a significant random element in any test at this level from arithmetic errors etc. The intent is that the difficulty and random element should offset by allowing multiple tries. If makeup tests delete the earlier scores then these considerations apply to them as well.