# A

# *How to Pass an Exam*

## Background

Exam testing is a common means to assess students' knowledge and skills. Not only do different educational systems have different ways to hold and grade exams, standards change over time as well. With few students to test, oral exams are clearly the least intricate and least time consuming way to grade and rank students. With hundreds of students in a course — unknown in the U.S., but quite common in Western Europe during early semesters — written exams are the only solution.

Multiple choice tests consist of a list of possibly correct answers for each question and the student chooses (marks) one (or more) answers each. Such tests are increasingly attractive with more students to test as grading becomes a matter of comparing the choices made by a student to the 'true' set of answers. Multiple choice questions can cover a far broader range of topics than fewer free-response questions. On the other hand, multiple choice tests may overemphasize memorization and test processes and comprehension poorly. They usually leave no room for disagreement or alternate interpretation, which makes them particularly unsuitable for non-technical subjects.

## Goals of Study

Given that most university exams are not defined by a central instance, it is worthwhile to check how well an exam captured the knowledge of the students. In particular, has the preparation for the exam by homework

worked out well? Are certain questions too complex or too easy, or can the structure of the exam — multiple choice vs. free-response — be optimized? Depending on the results, the exam should be changed to better assess the student's skills.

## Description of Data

The data we look at in this case study reflects the results of a 3-hour written exam in probability theory, which is taken mainly by math students in their third semester of studies. The exam was conducted in the winter semester 2005/06 at Augsburg University. In addition to the lecture, students were prepared in 4 homework groups. The exam consists of a multiple choice part (weighted 1/6th) and 5 out of 8 questions (weighted 5/6th) free for the student to choose. For 62 students who took the test, the following data was recorded:

- **Gender**
- **Major Subject**
  Mainly Math or Business Math
- **Semester**
  Students should take the test in the third semester, but can also take the test in a later semester, especially if they failed their first attempt or want to improve their mark.
- **Homework Group**
  One out of four; student's choice according to their class schedule.
- **Pre-Score**
  The average mark on the 12 assignments ranged between 0 and 25 points.
- **Multiple Choice**
  Points achieved in the multiple choice part of the exam, 0 – 50.
- **Points in Question 1 – 8**
  0 – 50 points, missing, if question was not selected.
- **Question 1 – 8 Selected?**
  Binary variables (Yes/No) indicating whether a question was chosen or not.
- **Total Selected**
  Number of questions worked on.
- **Sum Points**
  Sum of the multiple choice part and the 5 best questions worked on.
- **Mark**
  According to the German educational system, ranging between 1 (best) and 5 (failed) with $\pm 0.3$ differentiations.
  (There is no "0.7" or "5.3" and any score lower than 4.0 is a failing grade.)

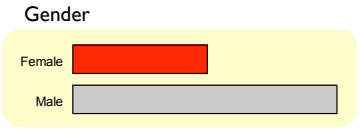
## Graphical Analysis

This analysis will only cover some of the most interesting of the numerous structural features of the dataset.

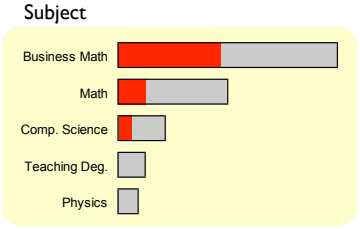
We start with the most important univariate features, which describe the sample structure. The barchart for *Gender* shows that about 1/3 of all students are female. Most of them (71.4%) study Business Math, which is the most popular subject with more than half of all students. Another quarter of all students registered for Math. The barchart for *Group* shows that the four homework groups differ strongly in the proportion of female students. Whereas group 1 has more than 50% female students, group 3 (the smallest group) has less than 20%. The fluctuation diagram for *Subject* and *Group* shows that Business Math students can be found primarily in groups 1 and 3 and Math students in groups 2 and 4. Such a structure is to be expected, because different subjects imply different class schedules, such that some of the homework groups overlap with other courses. The histogram of *Pre-Score* shows three groups of students: (1) those who regularly handed in their homework, (2) those who only rarely worked on the homework and (3) those who almost never handed in their homework. The corresponding spinogram reveals no relevant structure, given the small sample size. Only half of all students actually took the test in the scheduled semester as indicated in the barchart for *Semester*.

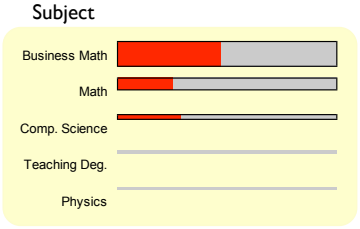
The missing value plot immediately reveals the popularity of the 8 questions. Questions 5 and 6 are by far the least popular choices, whereas almost every student worked on question 7. The parallel boxplots for the results of the individual questions show median results between 27 and 30 points for questions 2, 4, 6, 7 and 8. The three remaining questions have median results of 20 (question 3), 5 (question 1) and 0 points (question 5), which matches quite well with the insights from the missing value plot. Both distributions of the multiple choice part as well as of the total points achieved are left skewed. Comparing the overall results between the groups in a boxplot of *Total Points* by *Group* shows the best results for group 2 and the worst results for groups 1 and 4.

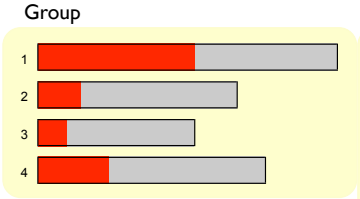
We use a scatterplot to test how well *Pre-Score* can predict the total points in the exam. Adding a linear regression gives a relatively poor $R^2$ of 20%. This is mainly due to the three groups that can be found in the distribution of the pre-score. Selecting group 2 shows a far stronger association for this group with an $R^2$ of 73.9%. The number of semesters has a surprisingly strong effect on the overall result. The boxplot *Total Points* by *Semester* shows a decline of roughly 30 points per year and a penalty of almost 50 points for students who started in the summer term and not in the winter term which is the default. To check whether the multiple choice part of the exam adds a discrimination of the result which is not already captured by the 5 questions, we derive the variable *Total Points - Multiple Choice*.

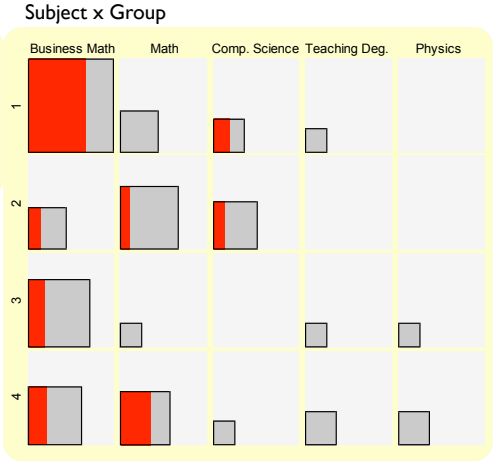



About 1/3 of the students are female. Business Math makes up for 50% of all students and has an almost equal share of male and female students. Only a few students come from other subjects.

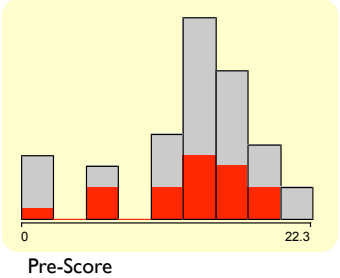


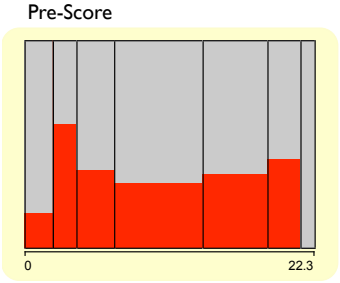



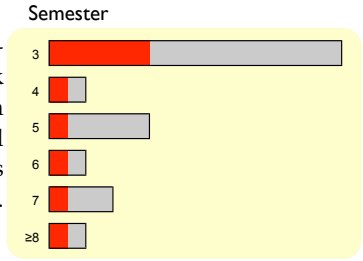

Group 1 is the largest of all four groups. Groups 1 and 4 have a higher rate of female students. Group 1 is dominated by female Business Math students. Physicists can only be found in Groups 3 and 4.

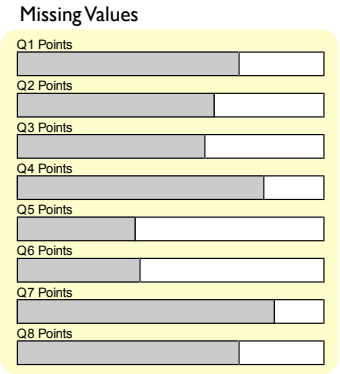



There are three groups in the distribution of the Pre-Score. The majority worked on their homework regularly and reached an average score of 15 points on average. Others handed in only a few assignments ending up with no more than 6 points. A third group almost never got any points on their homeworks.

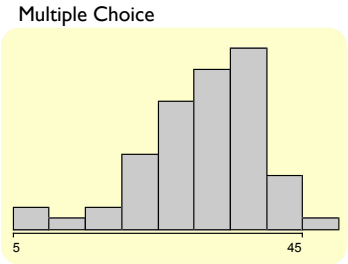



50% of all students took the exam in the 3rd semester as planned.



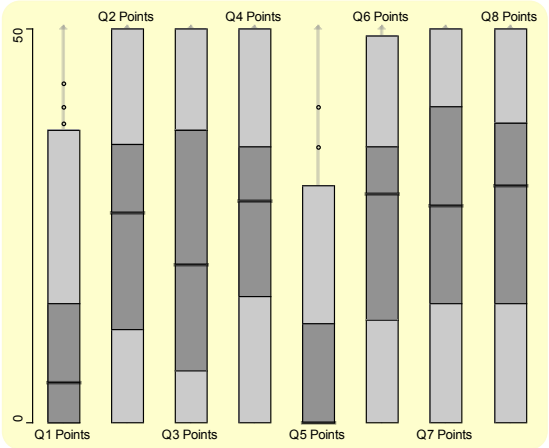


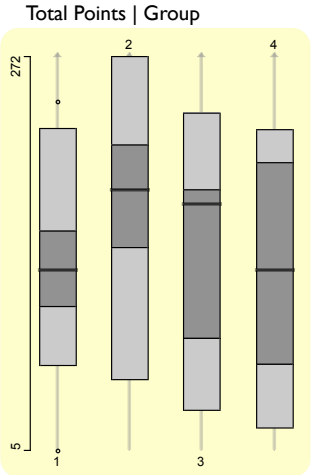

The missing value plot shows the popularity of the questions. Q7 was most often selected, whereas Q5 was least popular.



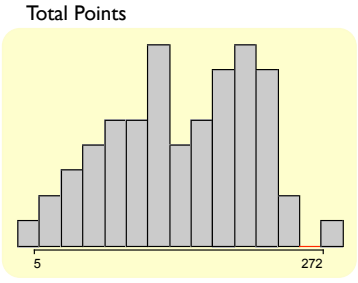

The distribution of the multiple choice part is left skewed with a mode at 35 points.

The distributions of the eight questions can be compared in parallel boxplots.
Q1 and Q5 show very poor results, whereas the other six questions have median results between 20 and 30 points.
None of the students got the full number of points for Q1, Q5 and Q6.

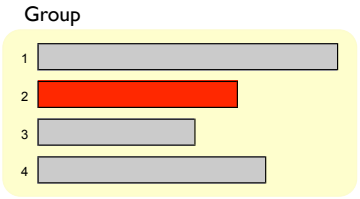


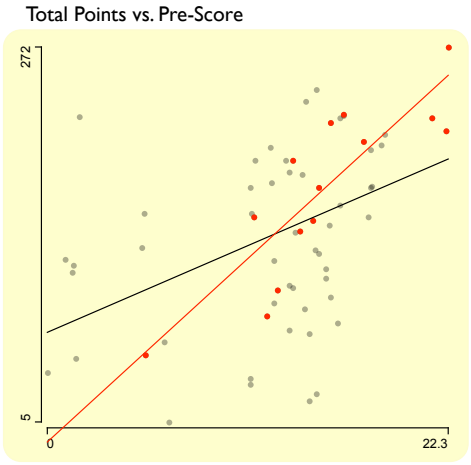

The four groups show quite different results. The median result of groups 1 and 4 is about 50 points below the median result in groups 2 and 3.

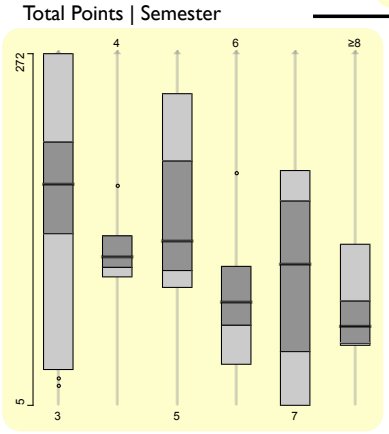



The distribution of Total Points is also left skewed with a mode at 200 points but has an extra mode at 120 points, which corresponds to "barely passed."



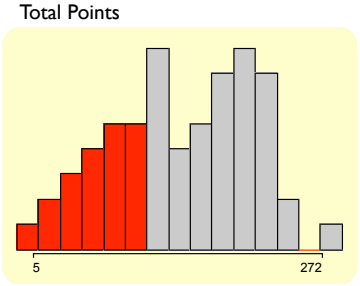

There is a moderate association between the Pre-Score and the Total Points ($R^2$=20%).
For group 2 the association is far tighter with an $R^2$ of 73.9%.

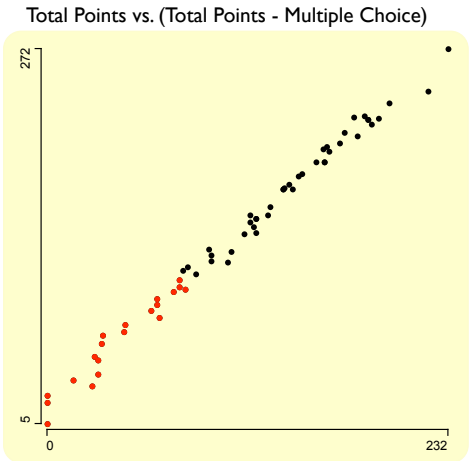



The results depend strongly on the semester in which the test was taken.
There is a decline by roughly 30 points per year, and students who took the test in an even semester performed worse by a margin of about 50 points.

Removing the multiple choice part from the exam does not alter the result.
Students who failed are selected.



In the scatterplot of *Total Points* vs. *Total Points - Multiple Choice* we see a very strong correlation; which is to be expected. Only one student who failed would rank before a student who passed, when using the result without the multiple choice points.

## Further Analysis

In the graphical analysis, we found that students in later semesters and those who started in summer term have on average inferior results compared to students in their third semester. We set up a simple linear model in R (assuming the data to be in the dataframe called `PTE`) to estimate the effect, which we already read roughly from the boxplots.

```
# Make all >=8 to 8 and the factor numeric
#
> levels(PTE$Semester)[6] <- "8"
> PTE$Semester <- as.numeric(as.character(PTE$Semester))

# Dummy for even years
#
> even <- 1 - (PTE$Semester %% 2)

> l1 <- lm(Total.Points ~ Semester + even, data=PTE)
> summary(l1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  200.273     21.927   9.133 6.85e-13 ***
Semester     -12.623      5.063  -2.493   0.0155 *
even         -23.909     21.633  -1.105   0.2736
...
> plot(PTE$Semester, PTE$Total.Points)
> abline(l1$coeff[1:2])
> abline(l1$coeff[1]+l1$coeff[3], l1$coeff[2])
```

The model estimates a decline of 12.6 points per semester, i.e., about 25 points per year. Students who are in an even semester — who started in summer term — have an estimated extra penalty of 23.9 points. These estimates are slightly smaller than what we read from the medians of the boxplots, which might be explained by the left-skewed distribution.

## Summary

Graphical methods are well suited to explore the structure of the sample. Using boxplots $y$ by $x$ and parallel boxplots the result of the exam can be conditioned on the various factors. Furthermore, it turned out that the

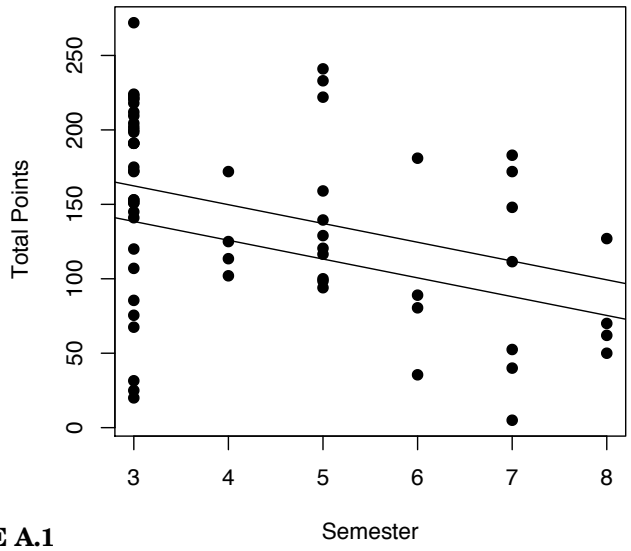



**FIGURE A.1**
The linear model estimate.

pre-score is a good indicator for the exam result for those students who take the homework seriously. The multiple choice part of the exam might also be neglected, as it does not contribute much to the discrimination of the students' skills.

The examination of the popularity of certain combinations of questions is left as an exercise.

## Exercises

1. Is *Pre-Score* correlated with *Semester*? Test graphically with linked barchart and histogram/boxplot and a boxplot $y$ by $x$.
2. What can be said about the number of questions a student worked on? Five questions are the norm — does it help to try out more questions or focus on fewer?
3. Analyze the structure of combinations of questions using the binary variables in a mosaic plot. Which combinations are most popular; which are most successful?
4. Compare the four homework groups regarding their average result and the number of students passed (students needed at least 110 points to pass).
5. Investigate and describe the influence of *Gender* and *Subject* on the students' results. Can these factors be added to the model set up above?