**Fairness-issues related to exams and teacher-made grade in the Norwegian context**

This note is to elaborate on some issues of fairness that I have wanted to explore for quite some time. Given that data exist and will be augmented with new data in relation to Jose’s project, I found it useful to pin down some of the ideas I have for how such fairness issues could be studied, given appropriate data. The ideas that I formulate are not directly related to Jose’s project, but hopefully there is still some interest for you to establish cooperation with me on this. Happy to work with either one or all of you on this. I understand there may be some issues on access to data, but hopefully this can be resolved if this is something we would like to continue exploring. In general, I am looking for data for grade 10 students, including detailed grades from both exams and teachers (standpunkt). I would also very much like to have a detailed set of grades for students in upper secondary school general track. My hope is that relevant data exist in, or may be added to, your project.

In brief, I have made observations (quite anecdotal evidence) that makes it interesting to explore some of the following issues. I briefly explain the issues, suggesting also some approaches to how the issue possibly could be studied. I would also like to add that some of these issues relate to work done in the working group that Sigrid steered on the future of exams in Norway. Their first report, summarizing the current knowledge, partly also identifies these or closely related issues, as issues where we currently do not have enough knowledge. I guess it is fair to say that some of these issues could have been elaborated on even further by this group, but that due to various backgrounds/interest/motivations among group members, some of them were not given a high priority.

1. **Is “trekkordningen” a truly randomized process?**

We know that in the logistics around exams, principals and others in schools with responsibilities to organise exams, have to take several conditions into consideration when student groups are identified for the various exams. Hence, there is likely not procedures in place, at least not in all schools, that resembles anything like drawing numbered balls from a jar. Augmented by convinced reasoning that I have heard from students waiting for the exam to be “drawn”, I also suspect that in addition to practical logistic constrains, there may be a selection process where schools to some extent seem to somehow maximise their student samples for certain exams.

**How to study:** In it’s simplest form, this is an ANOVA study where students are grouped according to their written exam (mathematics, English, Norwegian) where the mean teacher-made grade for students are compared across groups. If my suspicion is true, students who are “drawn” for the exam in mathematics, will have a slightly higher mean score than students who are drawn for another subject. I suspect that if there is any effect, this would have to be rather small, since whole classes are selected within schools. Similar designs can obviously also be used for the oral exams, and for these exams, the likelihood would be that the effect is even larger, given that these are smaller groups picked within each classroom (I have to check, but I believe the groups are typically less than 10 students)

1. **Uneven practice of grading across subjects.**

I believe this is a much larger issue than the one above, and maybe the one I am most concerned with. It is evident that there exist stable differences in both mean and standard deviations across subjects. For some reason this seems to be silently accepted by teachers and the system. In some sense, if this difference in mean grading is constant across schools and over time, the fairness aspect is not severe, given that this would be a constant shift in grades across subjects which is more or less the same for all students. However, there are several sub-issues relating to this phenomenon that deserves attention:

1. For me, such a difference in grading practice is intuitively unacceptable, so I have to ask: Why is this phenomenon accepted? What could be the possible reasons for accepting a practice where for instance mathematics is constantly graded lower than English? Is there an underlying and sub-conscious idea that mathematics is a more difficult subject than English, or equivalent, that kids are better in English than mathematics? Is this a phenomenon that exist across countries (I suspect, it is).
2. How does this affect students and future choices? What would happen if grading across subjects where better calibrated?

Obviously both a and b above would require unique and new data that I do not expect exist in your project. Partly, this would also be qualitative data. However, I include them here in order to present the totality of the issues with grading practices

1. Given that some subjects receives lower grades, students selected in grade 10 to sit an exam in mathematics are de-facto penalised (by the system) as compared to a student being selected for English. And IF students are not drawn to subjects randomly, the fairness issue becomes even more of an issue, where school administrators actually select students (although a penalization of students is a highly unlikely part of their motivation). In any case, this is clearly a fairness issue. In order to discuss this phenomenon with support in empirical data some of the following, again, simple descriptive analyses would be informative:
   * Overview of mean teacher-made grades across subjects, compared with exam grades for the same students (already reported, elsewhere, I am sure). This may possibly be transformed to an estimate of the effect this potentially has on the GPA used for the selection process.
   * For how many students does this have a direct effect? Is there for instance a variable identifying if students were offered their first choice for upper-secondary school? If so, a (weak) relationship between written exam subject and “first-choice-offered” may exist. It may be that this issue mainly is an effect in larger cities where selection of upper secondary school is markedly more competitive.
   * Using subject specific grades as “items” in an IRT model. The advantage of this would be to place subjects on a continuum, representing students’ general ability (as quantified by GPA). Using a two-parameter model we could furthermore study how some subjects possibly also discriminates harder than others. I will get back to this idea of running IRT on grades in subjects, as this opens possibilities for studying other fairness issues.
2. Continuation of c) above with specific issues relating to upper-secondary general (academic) track. All the issues mentioned above in the context of grade 10 exams are relevant to study also for upper-secondary. In addition, issues are a bit different given that students in upper-secondary general track have very different choices of subjects. However, in the end, for most tertiary studies, they will be competing with each other, independent of choice of subjects.
   * Carl Angell & Svein Lie did a study in the 80/90-ies where they demonstrated quite clearly that students with a specialisation in the “hard” sciences (physics and maths) are severely penalised in their GPA. Nothing really happened with their report (not available right now, but I will try to dig it up). What they did was the following: They calculated the mean scores in the compulsory and joint subjects (such as Norwegian language, English, history, etc) for student groups with different specialisations. Physics students were top performers with an average grade in the general subjects of approximately 5, while they received closer to 3 in Physics/Maths – both by their teachers and in their exams. For students with for instance a language specialisation this relationship between general subjects and their grades in the specialised subject were opposite. A similar description is still relevant to include, but augmented with even more sophisticated analyses (to be returned to). Some would argue that this is now (as opposed to when Angell & Lie did their analysis) compensated for by “realfagspoeng”. I would like to investigate this argument further by looking into the parliamentary debates and documents circulating at that time. I am pretty convinced that “realfagspoeng” where not introduced with this rationale. Rather this was introduced to motivate more students to select these subjects. If the situation still is similar to the one described by Angell & Lie I would argue that the motivational aspect of “realfagspoeng” is balanced out by the relatively lower grades you would receive if choosing these subjects in your specialisation.
   * To further study the possible effect of this issue, I suggest to do an IRT analysis where all students grades in all subjects in the general track are placed into one scale representing general ability (as described briefly above). The great advantage of IRT with these data, is that you do not need complete datasets. You can actually work quite efficiently on very sparce data with large chunks of missing by design. In this case, all the general subjects will work as an anchor. At least two immediate useful discussions can be brought up by this model: (i) again, a continuum of subjects from “easy” to hard” can be placed on a linear scale with approximate interval properties, and (ii) more importantly the IRT model would produce an alternative GPA estimate based on actually modelling general ability as a unidimensional scale. This scale could then be compared to students’ actual GPA. Furthermore, we could also discuss how subjects actually fits or contributes to this unidimensional scale
3. **Unequal grading practices across schools**

There is a widespread opposition to the idea that exam grades can be used to monitor unequal grading practices among teachers. The argument most often heard is that exams do not measure the complete curriculum, but rather represents a selection of it. This is, I assume, in itself a fair observation. However, what if the gaps in teacher-made and exam grades varies over school characteristics? What if there is a systematic discrepancy in the gap between teacher made and exam grades for the same schools over time? Even if accepting that exam and teacher-made grades do not represent measures of the same construct, I believe that such issues would represent a lack of fairness. One reasonable hypothesis is that, although grading is supposed to be criterion-based, it is hard for teachers to avoid using some kind of norming internally in their classrooms. The effect would then for instance be that students in a school with low-average exam grades, would receive upward biased teacher-made grades, and vice versa. Multilevel models would be relevant to study this issue, where some school characteristics are used as predictors

1. **Gender differences in grading**

By using the mentioned IRT modelling approach, the issue of gender differences could also be addressed as a DIF issue. In other words, are there subjects where girls/boys of the same general abilities receive systematically different grades?

I could probably have written down several other issues. Sigrid could also possibly add issues along the same lines (e.g. inequities relating to regions, municipal type, etc.). I still hope these notes are helpful for you. What I would like to discuss with you could be the following:

* + Do you have data already that can be used for these purposes?
  + Are these issues of interest for you to cooperate on?
  + Are there possible overlaps with work you have already started with?

I would also add that although this document represent a rather pragmatic and practically oriented perspective on issues of grading in Norway, I do believe they represent studies of international relevance. I know for instance that the IRT approach presented has been attempted for A-levels in the UK by a person in the research department at OFQAL, the regulating authority for exams. I also think that the issues of subject-related bias is not unique to Norway. Possibly, cooperation could be done with colleagues in other countries. Finally, I think the IRT approach could be potentially relevant for your work relating to predicting outcomes. If there are bias represented in the GPA’s used for selection into tertiary education, it is not unreasonable to assume that the IRT score for general ability, would represent a stronger predictor for future outcomes than the GPA.