Good afternoon.

My name is Tony. I am a first-year PhD candidate under Rolf’s supervision.

I wish to contribute to today’s session with some preliminary results from my first PhD project that asks whether GPA subjects differ in their difficulties.

This question is important because assessment practices have significant impact on students’ lives, particularly when learners approach the end of their grunnskole (Year 10), where decisions must be made over vocational or academic trajectories.

In Norway and most Nordic countries, such high-stake decisions are made almost exclusively based on a single criterion: grunnskolepoeng, a Norwegian term equivalent to grade point averages (GPA) in English.

Although details vary across jurisdictions, GPA is largely a sum-score measure.

Ensuring each component entering GPA computation is comparable in difficulties is, therefore, important not only for upholding assessment fairness, but also for enhancing measurement validity.

Surprisingly few studies, however, have tried to answering this fundamental question in the Nordic context.

Studies that have looked into this question in different countries, such as the UK and the Netherlands, find evidence challenging the assumption that subjects have equal difficulty.

This study therefore wishes to contribute to the academic and policy debates by examining the inter-subject difficulties in Norway’s GPA computation.

Under the Norwegian system, both teacher-assigned grades (standpunkt) and exam grades are included in the GPA calculation.

Both teacher-assigned and exam grades are integers between 1 and 6, with 6 being the top grade.

Students receive grades from their teachers on 13 compulsory subjects, such as Norwegian, English, mathematics, natural sciences and social sciences, etc, as well as one grade from a wide selection of electives.

Exams consist of both written and oral forms.

The written form covers Mathematics, Norwegian and English, and,

the oral form covers the same subjects, as well as other subjects.

While every student should receive teacher-assigned grades (that is, 100% sampling), the student cohort is divided evenly between participating in the written exam in mathematics, Norwegian and English, giving exam grades a 33% sampling probability.

Although 2/3 of the exam grades are missing for each exam subject, this can be safely modelled under the missing completely at random (MCAR) assumptions due to random assignment.

GPA is then computed as the unweighted sum divided by the number of subjects (ie, a simple average), multiplied by 10, then rounded to two decimal points.

With such background in mind, I now describe the current study.

This study draws its data from Norway’s national register.

This data source is unique such that it is the population, not samples, that is the subject of analyses.

The targeted population is the Year 10 cohort graduating in 2019.

I excluded students without valid GPAs, and the subject “Sidemål” from my analyses, leading to a dataset of 60,618 observations and 12 teacher-assigned grades, 3 written- and 2 oral-exam grades.

I employed partial credit models (PCMs) for my analyses. PCMs are the polytomous analogous of Rasch models.

It is particularly suited for the current study because GPAs are constructed as unweighted sums, therefore requiring the same discrimination parameters.

A PCM generates a series of probability curves, as shown in this diagram.

The horizontal axis represents students’ competency (usually represented by the Greek letter θ), with low competency on the left and high on the right.

The vertical axis represents probabilities, ranging from 0 to 1.

Taking the red curve “P4” as an example:

Students with a competency score of θ = 0 is most likely to receive a grade of 4.

As the competency increases to, say θ = 3, the probability of receiving 4 drops, while the probability of receiving a grade of 5 increases.

At certain point, the red curve “P4” crosses the yellow curve “P5”, signalling that students at this point is switching from being “more likely to be a 4” to “more likely to be a 5”. This switching point is marked as “b\_4” in this study.

Geometrically, these “b” parameters mark where the probability curves cross each other.

This slide presents visual summaries of the “b” parameters derived from PCMs.

Difficulty parameters of the 12 compulsory subjects (ie, teacher-assigned grades) are grouped on the left panel, in descending order.

For the purpose of enhancing comparability, difficulty parameters of the exam grades are presented in pairs with their teacher-assigned counterparts in the right panel.

Let’s first of all examine the teacher-assigned grades:

The b\_5 line on the top is relative flat, while the b\_4, b\_3, down to b\_1 lines are increasingly downward-sloping.

This “fanning out“ effect suggests a partial answer to the research question:

“Yes, subjects difficulties did differ—more so for the lower grades.”

In fact, a Grade 2 in the easiest subject (Food and Health) shared similar “b” parameter with a Grade 1 in the hardest subject (Mathematics).

Such differences, however, were minimal among the top grades 5 and 6.

From the right panel, differences between teacher- and examiner-assigned marks also showed interesting patterns.

It is first of all noticeable that written exams all had upward-sloping curves, while oral exams had downward-sloping ones.

On the surface, this pattern suggests that examiners were stricter than teachers in marking written tests, but more lenient in oral exams.

The second strand of the answer to the research question therefore can be:

“Yes, exam grades differ from teacher-assigned grades, depending on the form of examination.”

Contrary to the previous observation form the left panel, slopes in the right panel were the steepest on the top end, suggesting larger disagreement between teachers and examiners for awarding grade 5s and 6s while disagreement remained minimum for lower grades.

Amongst the 5 exams, Written Norwegian stood out as the subject with the largest teacher-examiner disagreement.

These preliminary results suggest great nuance among the GPA debates.

If divergence in grade difficulties signals potential unfairness or threats to measurement validity, it is the lower end in teacher-assigned grades, and higher end in exam grades where we see the greatest differences. SAY WHAT YOU FOUND AGAIN.

The causes and practical implications of difficulty parameters remain open to interpretation:

If mathematics carries higher difficulty parameters, does it suggest

learners are inherently less capable in this subject, or

the measurement device for mathematics is less sensitive than that of other subjects?

This person-centric vs subject-centric interpretations demand vastly different policy responses.

These results are also aggregates at geography level, sex level and SES level.

It would be fruitful to see whether grade difficulties diverge more severely for, say rural schools, females and the left tails of the SES distribution.

Towards the overarching goal of promoting educational fairness, this study may lend itself towards identifying the most vulnerable, who carry the heaviest burden of grade difficulty differentials.

(last slide should be who carries and change sexes to gender?)