**Comparing the Difficulty of Elective Subjects in Norwegian High Schools**

**with Non-ignorable Missing Data**

Grade point averages (GPAs) are widely used for tertiary admissions based on the implicit assumption that equal GPAs represent the same level of academic competency. Such presumption is further complicated by the introduction of unweighted marks from elective subjects such they are all equal measures of a student’s academic competency. Obviously, this assumption does not hold when there are differences in subject difficulty. In this case, students who enroll in the most difficult subjects will have their GPA punished and vice versa. Hence, inter-subject comparability is essentially an issue of fairness: when students take different subjects and these are equally weighted and used for high-stake purposes, inter-subject difficulty should be equal (Coe, 2008). Even though the importance of the GPA varies between and within countries, it is the most used measure for admission into tertiary education (Orr, 2019). Although some differences in subject difficulty is inevitable, assessing the degree to which subjects and therefore GPAs are comparable is of upmost importance. Yet, conceptual and methodological issues complicate this assessment.

In the past two decades, much of the literature on subject comparability has revolved around the difficulties of establishing justifiable criteria for comparison (Ofqual, 2015). Although there are many definitions of difficulty (see Newton, 2010), they generally involve assessing the competency, i.e., the skills, knowledge, and understanding required by a student to receive a specific grade. In a latent modeling framework, this can be done by identifying a common “linking construct” underlying all subjects through which they can be compared (Coe, 2008; Newton, 2005). If the conversion rate between the amount of the underlying construct and expected grade varies between subjects we can say that there are differences in difficulty. However, if subjects measure different constructs, factors such as student interest and motivation, the quality of teaching, and the utility value students see in the subject could explain differences in difficulty (Coe, 2008). Hence, if we cannot hold all these factors constant, multi-dimensionality makes comparisons hard. In these cases, perhaps the only justifiable choice is to revert back to what Newton (1997) calls the nominalist definition of comparability, where we compare the grade distribution of subjects.

The methodological issue of estimating inter-subject difficulty stems from the fact that whenever elective subjects are involved, missing data is present. Rubin (1976) showed that the missing at random (MAR) assumption is the weakest condition under which we can ignore the missing data without biasing full information maximum likelihood (FIML). In simple terms, a MAR mechanism posits that missingness is unrelated to unobserved variables, but possibly related to observed variables. Conversely, a missing not at random (MNAR) mechanism means that missingness is stochastically dependent on unobserved variables, something which can lead to considerable bias (Enders, 2022). Although a major methodological issue, surprisingly little attention has been given to the issue of missing data in the subject comparability literature. While the most comprehensive overviews on the extant literature make no mention of the issue (Coe et al. 2008; Ofqual, 2015), individual studies have often ignored the issue or taken for granted that student choice of electives can be explained by a MAR process (He et al. 2018, Coe et al. 2008).

**The current study**

In this study, we utilized register data to assess the difficulty of 11 elective and 4 mandatory electives in Norwegian high school. Utilizing a model proposed by Holman and Glas (2005), we also allowed for the possibility that student choice of electives can be explained by an MNAR process. Moreover, we also assess the dimensionality underlying the included subjects to arrive at a justifiable definition of difficulty. To reduce the influence of factors that are irrelevant to the study, we used data from 2018-2019, before the start of the COVID pandemic, a period marked by high levels of absence, digital schooling, and higher grades (The Norwegian Directorate for Education and Training [Udir], 2020a). Teacher-assigned grades were used instead of exam grades as they constitute 80-90% of a student’s final GPA (Udir, 2020b), and so are more important to the students’ future, while also providing a larger sample size.

**Research questions**

We posed two research questions:

**RQ1**: Is variation in Norwegian high school grades better explained by a uni-dimensional or multi-dimensional construct?

**RQ2**: To what degree is the difficulty of electives comparable?

To answer these, we modelled the subjects by item response theory (IRT) models that relate the underlying construct(s) of the subjects to the probability of obtaining a specific grade.

To answer RQ1, we compared models with different factor structures using nested model fit indices: a unidimensional model (Model 1) to a two- and three-dimensional model, whose structure was based on grouping of subjects available to the students. The two-dimensional model (Model 2) consisted of one dimension comprising science, technology, engineering, mathematics (STEM) subjects and one dimension comprising humanities subjects. The three-dimensional model (Model 3) also consisted of a STEM dimension, but the humanities dimension was further disaggregated into a language and social science dimension.

To answer RQ2, we added a selection model (Heckman, 1979) to the mode with the factor structure that provided the best representation of the data. In this way, we allowed for the possibility that s best fitting model of RQ1 to account for a potential MNAR mechanism. This model (Model 4) estimated the data-generating and missing data model simultaneously. Model 4 was compared to its baseline model without the selection model, and the best fitting model was used to assess relative difficulty in two ways: Firstly, difficulty was assessed through the convertion rate between amount of the underlying construct and expected grade—the linking construct definition. Secondly, since multi-dimensionality complicates this assessment, we also simulated a scenario where every student enrolled in every subject by computing expected grades based on the posterior expectation of the best fitting model.

**Literature review**

The issue of inter-subject difficulty comparability applies to all countries that utilize the GPA as a measure of academic competency. Although the instruments used to select students into tertiary education varies between and within countries, most universities and colleges use the GPA in some capacity for student selection (OECD, 2019). For instance, in countries such as the US and Italy, standardized tests play a huge role in admission processes, while countries such as Canada, Sweden, France, and Germany mostly rely on the GPA for selecting students (Orr et al., 2017). Moreover, in recent times, European countries have provided their national universities with increasing amount of autonomy in how the instruments they want to use for admission processes (Haj et al., 2018). Therefore, when universities are looking for the most appropriate selection instruments, it is important that they know the strengths and limits of using the GPA as a unitary measure of academic competency. Because even though hundreds of thousands of students are accepted into universities in the OECD each year, no publicly available research supports the comparability of subjects (Ofqual, 2015).

On the contrary, research shows that a host of different methodologies and study contexts support the notion that there are considerable inter-subject difficulty differences in high school subjects around the world. In England, Coe (2008) found a uni-dimensional model could describe variance in 34 grades sufficiently well after removing music, art and design, and fine art and performing studies. Comparing these subjects through their common linking construct, he found that physics and foreign language subjects were considerably more difficult than others. Also in England, the same subjects were found to be among the most difficult by He et al. (2018), who found a uni-dimensional construct adequate. The linking construct comparison was also made by Veas (2017) who found that mathematics and Valencian were the hardest subjects for Spanish high schoolers. One notion that Newton (1997) pointed out about this linking construct approach is that if the common factor is not strong enough, we risk ignoring the skills, knowledge, and understanding that is specific to each subject (Newton, 1997). Hence, these studies could exclude important and unique competencies that each individual subject measures.

Nevertheless, other researchers support these results as STEM and foreign language subjects have consistently been found to be more difficult compared to other subjects (Kuipianen et al., 2016; Ofqual, 2015).

Find more studies.  
Explanations for why these are more difficult?

Some countries acknowledge and adjust for the difference in subject difficulty. For instance, in Norway, students receive bonus points added to their GPA by enrolling in STEM subjects, something which was partly implemented to counteract the lower grades awarded in these electives (Tveitereid et al., 1997). Likewise, Sweden has a similar system where students can earn merit points if they take subjects that are perceived to be more difficult (Utbildningsdepartementet, 2018). However, the amount of points students receive in these countries seems to be determined rather heuristically and is not based on any quantitative analysis. Other countries, such as Singapore and Cyprus, and Finland have adopted norm-referenced grades to align grade distributions for electives (He et al., 2018; Kuipianen et al., 2016). These countries highlight the many problems around the distributional fairness of statistically adjusting grades as many of the best students avoid subjects that are consistently scaled down (Lamprianou, 2009). This is however also a problem when subject grades are unweightet as there is ample evidence across the world that students are highly aware of the discrepancy in difficulty among subjects and are driven away from the hardest ones (Bell et al., 2007; Lamprianou, 2009; Lødding et al., 2021). Strikingly, Kjærnsli & Lie, 2011 found that Scandinavian student’s expectations of succeeding in a subject is a greater predictor of enrolling in that subject than their actual grades.