UNIVERSITY
OF OSLO

Master thesis

Comparing the Difficulty of Elective Subjects in Norwegian High School with Non-ignorable Missing Data

Sverre Ofstad

Assessment, Measurement and Evaluation

30 Credits

Centre for Educational Measurement

Spring, 2023

Popular Abstract

When high school students can choose their own subjects, differences in the difficulty of these subjects can lead to varying GPAs, even if students have the same academic skills. Therefore, it is important to establish subject difficulty to ensure fairness to students. However, statistical estimation of subject difficulty is complex because of selection bias: students in different elective subjects differ in their academic competency. In this study, we compared the difficulty of subjects by controlling for this selection bias. The study included 11 elective subjects in Norwegian high school, with a sample of 21,832 third year students. We found that natural science and math (STEM) subjects require different skills and are much harder than other subjects. When correcting for selection bias, the mean grade of STEM and non-STEM electives was 3.25, and 4.38 respectively. Hence, the GPA of Norwegian high school students does not only reflect their academic skills, but also the decisions they make when choosing elective subjects.

Acknowledgments

I want to thank my wonderful supervisors, Björn Andersson and Tony Tanwho have supported me throughout this process. Thank you, Tony, for your eagerness to help and letting me steal the idea for this thesis. Thank you, Björn, for your immense methodological knowledge and your encouragement throughout this process.

I also want to thank intelligent and kind staff at CEMO. I’ve had an amazing time here, something which you all contributed to. Lastly, I also want to thank my classmates, and especially the *absolute* madlads in the study room at CEMO for the good conversation and fun we’ve had.

Abstract

When grade point averages (GPAs) are used for admission into tertiary education, there is an implicit assumption that the GPA of any two students relate to the same level of academic competency. However, this assumption does not hold when students are free to choose electives, and there are differences in their difficulty. Assessing the degree to which difficulty differs is complicated by two factors. Firstly, the sense in which we can compare subjects relies on the degree to which they measure the same construct. Secondly, if students' choice of elective is related to their academic competency, selection bias could distort difficulty estimates. In this study, we utilized Item Response Theory (IRT) to examine dimensionality and difficulty of 11 electives in Norwegian high school with a sample of 21,832 third year students. Dimensionality was assessed by comparing models with different factor structures, and selection bias was accounted for when estimating difficulty by incorporating student choice via a selection model. Our results found that natural science and math (STEM) subjects measure something distinct from other subjects. Furthermore, we found that ignoring student choice of electives when estimating difficulty introduced considerable bias in difficulty estimation. When correcting for selection bias, the mean grade of STEM and non-STEM electives was 3.25, and 4.38 respectively. Hence, the GPA of Norwegian high school students does not only reflect their academic competencies, but also the decisions they make when choosing elective subjects.

Keywords: inter-subject difficulty, missing data, MNAR, selection model, dimensionality

Comparing the Difficulty of Elective Subjects in Norwegian High School with Non-ignorable Missing Data

Comparing the difficulty of school subjects has been a controversial and long-standing issue in educational assessment. Already a century ago, Crofts and Caradog Jones (1928) used a form of equating to quantify the relative difficulty of examination subjects in English secondary schools. Since then, the statistical difficulties of any such analysis have become increasingly apparent. Additionally, although seemingly simple, conceptualizing inter-subject difficulty has shown to be an abstract and complex task. Poor or undeclared conceptualizations of comparability and statistical issues have interweaved, leading to a research field riddled with disagreement (Newton, 2011). As a result, some have argued that it is an impossible task, and that "we should learn to accept and adapt to the unintelligible enigma of comparability between subjects" (Newton, 1997, p. 448). Yet, inter-subject comparability is difficult to ignore as it 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). When grade point averages (GPAs) are used for admission into tertiary education, there is an implicit assumption that the GPA of any two students relate to the same level of academic competency. However, this is not necessarily the case if students enroll in different subjects.

In Norway, although the GPA of a student is an unweighted measure, the point system used for admission into tertiary education is not. University and college admission is based on study points that comprise a student's GPA in addition to their bonus points. High school students can receive bonus points for enrolling in either language or science, technology, engineering, and mathematics (STEM) electives. While the general purpose of introducing these STEM points was to increase recruitment into STEM subjects, it was also explicitly argued that they are meant to counteract the generally lower grades awarded in these subjects

(Tveitereid et al., 1997). Since their introduction in 1998, The Norwegian Ministry of Education and Research (KD) has repeatedly emphasized the importance of STEM points to maintain recruitment into these subjects despite generally awarding lower grades (KD, 2005; KD 2010). However, recently, a public inquiry commissioned by KD proposed to remove STEM and other bonus points to streamline university admission. Although the report recognizes subjective differences in how teachers grade, it also argues that GPAs are objective and comparable measures of academic competency (Official Norwegian Reports, 2022, p. 67). However, as noted earlier, the comparability of subjects is a complex task, and does not only depend on inter-subject difficulty, but also the comparability of the underlying constructs that subjects measure.

The potential multi-dimensionality of subjects does not only limit our ability to compare any two GPAs, but also complicates the basis from which we can compare subject difficulty. Although there are many definitions of inter-subject difficulty (Newton, 2010), they generally involve comparing the competency, i.e., skills, knowledge, and understanding, required by a student to receive a specific grade. If subject A and B measure the same competencies, but subject A requires higher levels of this competency to award the same grade, subject A is more stringently graded and therefore more difficult. 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). Unless it can be shown that these factors are equal across groups of students in subjects that measure different constructs, our basis for comparison is undermined. Therefore, some argue that inter-subject difficulty comparisons are only defensible when variation in all subject grades can be explained by a common underlying construct (Coe, 2008, Newton, 2010). Research shows conflicting results regarding the dimensionality of school subjects, depending on the methodology used to assess it, and the context of the study (e.g., Bowers, 2011; Coe, 2008; He et al., 2018; Korobko et al., 2008).

Item response theory (IRT), a group of latent trait models, can be used to assess dimensionality, and has been extensively used in comparability studies of school subjects (e.g., Coe, 2008; He, 2018; Korobko et al. 2018; Veas et al., 2017). IRT models can also handle some of the issues traditional approaches to subject comparability are faced with. In IRT, the difficulty of each subject is estimated independently, and so there is no need to assume grade intervals to be equal within or across subjects. Moreover, IRT accounts for measurement error, which can be substantial in educational assessment (Brookhart et al., 2016). Finally, IRT provides unbiased parameter estimation when students have taken a different number or different set of subjects. This is an important reason for the popularity of using IRT in comparability studies, because when elective subjects are involved, there will be missing data. Yet, if the probability that students enroll in specific electives depends on their proficiency level, a self-selection bias could distort IRT estimates.

Since IRT estimators do no utilize closed-form expressions, it is impossible to analytically examine the extent to which selection bias leads to parameter bias (Rose, 2013). However, IRT estimation on simulated and real data shows that when choice is involved and subsequently ignored, likelihood estimation can produce bias in item and person parameters (Finch, 2008; Korobko, et al., 2008; Rose et al., 2010; Rose 2013). The severity of this bias largely relies on two factors: the strength of the association between proficiency and missing propensity and the amount of missing data (Pohl & Becker, 2020; Rose, 2013). Therefore, in some contexts, IRT models that ignore selection bias have shown to be robust for their intended purposes (Korobko et al., 2008; Pohl et al., 2014). The missing data is then said to be ignorable: information about the missing data does not need to be included in the model. In other studies, selection bias has produced severe parameter bias (Finch 2008; Rose et al., 2010). In these cases, the missing data is said to be non-ignorable: to obtain unbiased estimates, information about the missing data must be incorporated in the model.

The Current Study

In this study, we used Norwegian register data on seniors in high school to estimate inter-subject difficulty with non-ignorable missing data. 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. The difficulty of 11 electives and 4 mandatory subjects was estimated by IRT models. To examine dimensionality and the impact of selection bias on parameter estimation, we compared models that make different assumptions regarding these factors. The baseline model was a unidimensional IRT model which assumes that variation in grades can be explained by a single underlying trait. This model was expanded to a two- and three-dimensional simple structure model. The two-dimensional model consisted of one dimension comprising STEM subjects and one dimension comprising humanities subjects. The three-dimensional model also consisted of a STEM dimension, but the humanities dimension was further disaggregated into a language and social science dimension. Lastly, to account for the hypothesized self-selection bias that occurs when students freely choose subjects, we utilized a joint model proposed by Holman and Glas (2005) that estimates the data-generating model and missing data model simultaneously. These models were used to answer three 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?
- RQ3: How much does selection bias distort the comparability of difficulty between subjects?

Research questions 1 and 2 pertain to the degree to which subjects, and therefore GPAs, in Norwegian high school are comparable, while research question 3 addresses whether IRT models that ignore the missing data can be used to make these comparisons. Hence, the results of this study have implications regarding the use of the GPA as a summary measure of academic competency and the considerations needed when using teacher-assigned grades from electives in statistical modelling.

Conceptual Framework

The Norwegian School System

Any Norwegian student that completes grade 10 has the right to attend high school, and most students begin the year they turn 16 (Udir, 2022). The Norwegian high school system is structured in a way that provides much freedom in what students can study, with 15 different study programs in either vocational or general studies. This study focuses on the most popular of these programs—specialization in general studies—which almost half of all students attend (Udir, n.d). In their second year, students enroll in electives and must specialize in one of two areas: languages, social science, and economics studies or natural science and mathematics. The latter is often referred to as science, technology, engineering, and mathematics (STEM) subjects, a term that is also used in this study. Specialization means that the student must pick at least two electives from one specialization, and two of these subjects must be continued into their third year. When graduating, a typical student's GPA consists of a handful of exam grades, and teacher-assigned grades from a little over 20 subjects. Six of the teacher-assigned grades and roughly half of the exam grades are from electives (Udir, 2023a). After the national educational reform in 2006, teacher-assigned grades are stipulated to reflect only the overall competency the student has attained in a subject (Udir 2023b). The basis teachers use for this assessment are the learning standards stipulated for each subject in its curriculum.

The GPA is used as a summary measure of grades from subjects with varied learning standards, and is defined as a measure of overall academic competency (Udir, 2023b). When students take electives, the validity of this definition rests upon the assumption that overall academic competency is a unidimensional construct. In this study, confirmatory IRT models were used to assess the dimensionality of academic competency as measured by subjects. Unlike exploratory factor analysis, we therefore compared a priori specified model structures. Groupings were identified based on the two specializations students can choose from: STEM, and languages, social science, and economics studies. Since we did not include enough economics electives in the analysis, these subjects were subsumed under social sciences. We therefore identified three possible dimensions: STEM, languages, and social sciences.

Defining Difficulty

Varying definitions of comparability, or lack thereof, have caused much confusion in subject comparability research. Additionally, when defining comparability, researchers have often neglected to separate between definitions of comparability and the methods used to assess it. In recent times, there has therefore been a call for comparability studies to explicitly define the concepts that are compared (Newton, 2010; Ofqual, 2015). In IRT models, the difficulty of subjects is defined in relation to their underlying latent traits, and this relationship was used as the basis for our two methods and definitions of difficulty. The first method involves comparing the item characteristic curves (ICCs) of subjects, which graphically represent the probability of observing a specific grade given a student’s location on the latent trait. If a common “academic competency” construct can be identified across all subjects, we can compare the conversion rate between levels of academic competency and grades. This difficulty definition is akin to Coe's (2008) and Newton's (2010) concept of a "linking construct". However, when subjects are multi-dimensional, there is not one unifying linking construct, and therefore hard to find a common basis for comparison. Therefore, we utilized a second definition which Newton (1997) calls the nominalist approach. Here, missing grades were calculated based on the posterior distribution of the best-fitting model to simulate a scenario where every student took every subject. The accompanied definition of difficulty is that a subject is more difficult than another if every student took the subject and received lower grades. This definition does not say anything about grading leniency/stringency, but since we calculated grades for every student it does provide a comparison of grades unaffected by selection bias.

Missing Data Theory

Two concepts within missing data theory elucidate how students' self-selection into electives could bias difficulty estimates. These concepts are explained based on Rubin's (1976) typology of missing data, which is still widely employed today. Let **Y** be the vector of subjects in Norwegian high school, which can be partitioned into the observed and missing grades, $\mathbf{Y}_{\mathbf{obs}}$ and $\mathbf{Y}_{\mathbf{miss}}$**,** so that $\mathbf{Y} = (\mathbf{Y}_{\mathbf{obs}}, \mathbf{Y}_{\mathbf{miss}})$. Additionally, let D be a random variable that indicates whether Y is observed (D = 1) or unobserved (D = 0). The configuration of Y and D describe the pattern of missing grades in the data matrix, and is called the missing data pattern (Enders, 2022). The missing data pattern may take many forms. For instance, in large-scale assessments such as PISA, the pattern can be identified as planned missingness, with large amounts of missing data in some variables. In the context of this study, however, there are missing grades scattered throughout the entire data matrix.

While the missing data pattern describe where missing grades are, the missing data mechanism describes different ways which D relates to $\mathbf{Y}_{\mathbf{obs}}$ and $\mathbf{Y}_{\mathbf{miss}}$. We assume that Y and D are random variables with a joint distribution, and so we can theoretically compare the conditional distributions g(Y | D = 1) and g(Y | D = 0). Even though the latter is unobservable, we still assume that such a distribution exists, and use it to distinguish between different missing data mechanisms. Rose (2013) provides a frequentist definition of Rubin's three missing data mechanisms by further separating $\mathbf{Y}_{\mathbf{obs}}$ and $\mathbf{Y}_{\mathbf{miss}}$ into $\mathbf{Y}^{\mathbf{-n}}{}_{\mathbf{obs}}$ $\mathbf{Y}^{\mathbf{-n}}{}_{\mathbf{miss}}$, which denote the observed and missing grades of the subject vectors without subject *n*. With respect to a single subject Y_n, the observational status of which is denoted by D_n, the first missing data mechanism is defined as

$$P(D_n = 0 \mid \boldsymbol{Y}) = P\ (D_n = 0). \tag{2.1}$$

In this case, the probability that grades are missing is independent of any missing or observable grades, and so they are said to be missing completely at random (MCAR). The second missing data mechanism is called missing at random (MAR) and is defined as

$$P(D_n = 0 \mid \boldsymbol{Y}) = P(D_n = 0\ |\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{n}}). \tag{2.2}$$

Here, the probability that grades are missing is *not* stochastically independent from the grades in the data matrix. However, once conditioned on the observable grades, missingness becomes a purely random event. Therefore, MAR is often termed conditionally missing at random (Enders, 2022). In IRT, maximum likelihood estimation provides unbiased estimates when the missing data mechanism is MAR or MCAR (Rose, 2013). Therefore, even though there is selection bias when students choose electives, as long students' selection of subjects can be described by Equation 2.2, difficulty estimates are not biased. However, this is not the case for grades that are missing not at random (MNAR), which definition is:

$$P(D_n = 0 \mid \boldsymbol{Y})\ \neq\ P(D_n = 0\ |\boldsymbol{Y}_{\boldsymbol{obs}}^{-\boldsymbol{n}}). \tag{2.3}$$

Here, the probability that grades are missing is related to what they would have been if they were observed. An MNAR mechanism implies that even after controlling for the observed grades, the distributions g(Y | D = 0) and g(Y | D = 1) differ, thus violating the distributional assumptions of maximum likelihood estimation. If information about the selection of students into electives is not included in the model, parameter estimation can be considerably biased.

Selection Models

One method of including information about the selection process is through a selection model, which models the probability of a value being observed through a set of parameters, $\boldsymbol{\phi}$ (Heckman, 1979). Selection models are often called nuisance models because $\boldsymbol{\phi}$ are not really of interest; they just describe the missing data. The parameters of interest are called the focal parameters, denoted by $\boldsymbol{\beta}$. In this study, these include item and person parameters describing how teacher-assigned subject grades relate to their underlying construct. The focal model and nuisance model can be jointly estimated to assess the degree to which the missing data is ignorable. The missing data is said to be ignorable if it is MAR and the parameters of the nuisance model do not convey any information about the focal parameters. On the contrary, if the missing data is MNAR or the nuisance model provides information in the estimation of the focal parameters, the missing data mechanism is non-ignorable (Enders, 2022). If the selection model is properly specified, information from the nuisance parameters is utilized in the estimation of the focal parameters, and the bias introduced by the MNAR data is eliminated (Korobko et al. 2008). The selection model utilized in this study, proposed by Holman and Glas (2005), models the propensity of students to enroll in electives as a manifestation of a latent choice propensity. In other words, we assume that there exists a latent trait which governs student choice of electives.

Eccles et al.’s (1983) value-expectancy theory has been used as a framework in studies that examine why students choose electives in Norwegian high school (Holmseth, 2013; Lødding et al. 2021; Ramberg, 2006). The theory posits that students enroll in subjects based on value—interest and utility—and their expectations of success. Although external factors, such as guidance from parents and counselors, are important considerations for students when choosing electives, intrinsic factors generally agree with the tenets of value-expectancy theory (Lødding et al., 2021). Interest in, and perceived utility value of subjects, as well as expectations of success are important determinants for choice of electives (Holmseth, 2013). In fact, student's expectations of succeeding in a subject is a greater predictor of enrolling in that subject than their actual grades (Kjærnsli & Lie, 2011). According to value-expectancy theory, value and expectancy of success are reciprocal in that interest and competency are mutually reinforcing: an activity becomes more valued when the expectation for success is higher and vice versa. This is supported in the Norwegian context where students who choose electives based on interest tend to have an innate joy and confidence in their own abilities to do well in the subject (Ramberg, 2006). Further, value-expectancy theory posits that interest is maximized when students are optimally challenged. This is also supported in Norway, where the ability to demonstrate one's abilities and to be challenged are important factors for students when choosing electives (Holmseth, 2013; Lødding et al., 2021). To summarize, value-expectancy theory proposes that students will choose subjects that are interesting and allows them to demonstrate their competencies, and that are optimally challenging. Thus, we assume that students enroll in subjects in large parts because their level of academic competency aligns with the difficulty of subjects.

Method

Data and Sample

The data were obtained by Udir and contains the grades of every high school student in Norway. A data protection impact assessment for using the data was evaluated and approved by the Norwegian Agency for Shared Services in Education and Research (Appendix I). The target population for this study was third year students enrolled in the specialization in general studies program in the year 2018-2019. Since the dataset did not show the year or program of all students, the target population was identified by including students who took all five mandatory subjects for third year students in the general studies program. This led to a sample of 28,553 students. Two other study programs share the same mandatory subjects as general studies, and therefore we identified and removed students who took mandatory subjects offered exclusively by these programs. This led to a sample of 25,505.

The model specification used to model student choice assumes that students are free to choose subjects, something that required further exclusions. Firstly, students who repeated year three were excluded to avoid including those forced to retake subjects and therefore did not have a choice. This reduced the sample to 23,457. Secondly, only students who took at least three electives, which is required to finish year three, were included. The electives available to students were imported from www.vilbli.no, an information service for high school students established by local municipalities and Udir (Vilbli, n.d.). This led to a final sample of 21,832 students.

Subjects

All mandatory subjects except physical education were included in the analysis. This subject has highly unique grading criteria and practices and competence goals (Vinje, 2021). Due to its idiosyncratic nature, physical education does not fit into either a uni- or multi-dimensional confirmatory framework of academic competency and was therefore excluded. The number of electives to include was decided upon by plotting subjects against the number of students enrolled in them to see where there was a considerable drop-off (Appendix 3A). This resulted in the inclusion of 11 electives in addition to 4 mandatory subjects (Table 1). The mandatory subject Norwegian and the elective subject English provide students with both oral and written grades, and so 17 different grade variables were used in the analysis. 73% of the 25,505 students identified as enrolled in the general studies program took two or

Table 1

Subjects Included in the Study

Elective subjects	Mandatory subjects
Psychology	Religion
Law	History
Politics	Norwegian, written
Marketing	Norwegian, oral
Sociology	Norwegian, second-choice
English, written	
English, oral	
Biology	
Math, STEM	
Math, sociological	
Chemistry	
Physics	

Note. The difference between STEM and sociological math is that the former focuses more on the theoretical aspects of mathematics, while the latter involves more practical applications of mathematics in a social science research context. Second-choice Norwegian teaches students how to write in the dialect of Norwegian that differs from the main one spoken in their municipality. The full name of each subject is found in Appendix 3B.

more of the included electives, so the included subjects are a fairly representative sample of the most frequently taken electives.

Coding

Norwegian high school students receive a number grade from 1-6, where 1 indicates a failing grade. However, some of the grades in the dataset were non-numeric, indicating some special grading procedure. About 2% of grades were marked «exempted from grading», "participated", and "approved", all of which mean that the student enrolled in the subject but was exempted from being graded (see Appendix 3C for a detailed description). Virtually all these cases related to the subject second-choice Norwegian from which non-native students or those with language issues can be exempted from being graded (Udir, 2023b). Since these grades mean that students enrolled in the subject, but did not receive a number grade, they were coded as NA. Additionally, about 2% of the grades were marked "no basis for assessment", indicating that the teacher did not have enough information about the student to set a grade, something which usually signals large amounts of absenteeism. As this is treated as a failing grade, they were recoded to 1.

Analysis

The following models, with different assumptions about dimensionality and the missing data mechanism were compared in how well they represent the relationship between subjects and the underlying construct that they measure.

Models 0 and 1

In Models 0 and 1 it is assumed that academic competency, as measured by teacher-assigned grades, is a unidimensional construct. Two unidimensional IRT models were compared: the generalized partial credit model (GPCM; Muraki, 1992) and the graded response model (GRM; Samejima, 1969). These models are highly similar and include the same number of parameters but are conceptualized differently. In the GPCM, the probability that a student gets a specific grade is modelled directly, whereas in the GRM, this relationship is modelled in a cumulative fashion (Dai et al., 2021). The GRM specifies the cumulative probability of student *i* obtaining grade *j* (*j* = 1,…,6) or higher in subject *n* as

$$P^{*}(X_{ni} = j \mid d_{ni} = 1;\ \theta) = \frac{e^{a_n(\theta-\delta_{nj})}}{1+e^{a_n(\theta-\delta_{nj})}}, \qquad (3.1)$$

so that$P^{*}(X_{ni} = 1 \mid d_{ni} = 1;\ \theta) = 1$. θ is the latent trait assumed to explain variation in grades, α_n indicates how strongly grades vary for subjects as ability level changes, and δ_{nj} are the category boundary locations on the latent scale, where the probability of obtaining grade *j* or higher is .50. Marginal probabilities are given by the difference between the cumulative probabilities,

$$P(X_{ni} = j \mid d_{ni} = 1;\ \theta) = P^{*}(X_{ni} = j \mid d_{ni} = 1;\ \theta) - P^{*}(X_{ni} = j+1 \mid d_{ni} = 1;\ \theta). \qquad (3.2)$$

Model 0 (GPCM) and Model 1 (GRM) make two assumptions about the data that are tested in the subsequent models. Firstly, they assume that the missing data mechanism is ignorable, and therefore does not need to be included in the estimation of focal parameters. Secondly, they assume that the dimensionality of the model is properly specified: differences in grades can be explained by a unidimensional construct.

Models 2 and 3

in Models 2 and 3, the dimensionality assumption is tested, and the GRM is expanded to allow academic competency to be represented by more than one latent trait. As outlined in the conceptual framework, we hypothesized two multi-dimensional models (Table 2). Model 2 is a two-dimensional model consisting of a STEM dimension, and a humanities dimension that contains both social science and language subjects. Model 3 is a three-dimensional model where the humanities dimension is further divided into a social science and language dimension, in addition to the STEM dimension. In these models, Equation 3.1 is expanded so that the probability function of the models is now with respect to multiple latent traits, θ_q,

$$P^*\left(X_{ni} = j \mid d_{ni} = 1;\ \theta_q\right) = \frac{e^{[\Sigma\, a_{qn}(\theta_q - \delta_{nj})]}}{1 + e^{[\Sigma\, a_{qn}(\theta_q - \delta_{nj})]}}. \qquad (3.3)$$

However, Models 2 and 3 are simple-structure models where each subject only loads on one dimension, so α_{qn} is fixed to zero to all but one dimension.

Table 2

Structure of Models 2 and 3

Humanities		
Social science	Language	STEM
Religion	Norwegian, written	Biology
History	Norwegian, oral	Math, sociological
Psychology	Norwegian, second-choice	Math, STEM
Sociology	English, written	Chemistry
Marketing	English, oral	Physics
Law		
Politics		

Note. STEM = science, technology, engineering, mathematics

Model 4: Modelling academic competency with non-ignorable missing data

In Model 4, the assumption that the missing data mechanism is ignorable is tested. Here, following the model proposed by Holman and Glas (2005), a selection model is introduced that estimates the probability that a student enrolls in a specific subject. The selection model relates the observational status of subject *n* and student *i* to a latent choice propensity through a choice variable,

$$d_{ni} = \begin{cases} 0 \text{ if student } i \text{ did not choose subject } n \\ 1 \text{ if student } i \text{ chose subject } n \\ \text{NA if student } i \text{ could not choose subject } n. \end{cases} \tag{3.4}$$

Our approach differed from the model of Holman and Glas (2005) by including the possibility that the choice variable can be missing for some students. Not all subjects are offered at every Norwegian high school, and therefore students at these schools did not have a choice in enrolling in those subjects. The introduction of the NA response category of the choice variable is conceptually similar to the model utilized by Glas and Pimentel (2008) in their IRT estimation with missing data.

The dichotomous GGUM (Roberts et al., 2000) was the choice for the selection model. The GGUM specifies the relationship between the observed choices, *dni*, and a latent choice propensity, assumed to govern student choice, as

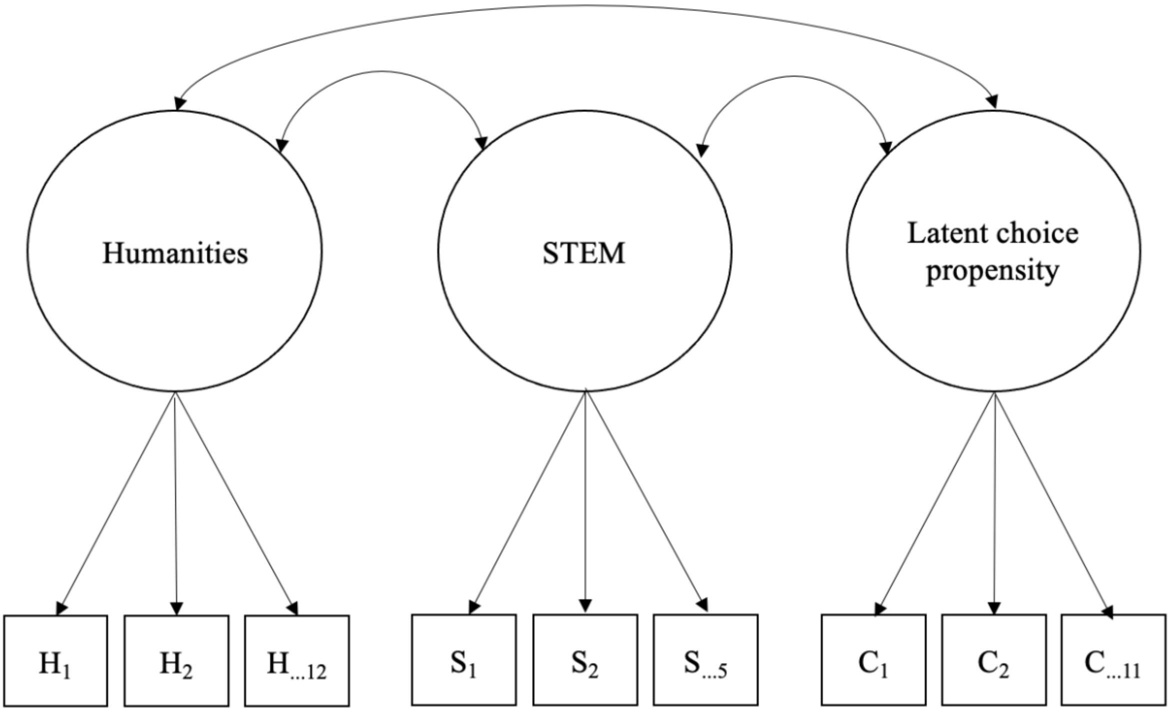
$$\mathrm{P(d_{ni}} = 1\,|\theta_i) = \frac{e^{(a_n[(\theta_i - \delta_n) - \tau_n])} + e^{(\alpha_n[2(\theta_i - \delta_n) - \tau_n])}}{1 + e^{(a_n[3(\theta_i - \delta_n)\,])} + e^{(a_n[2(\theta_i - \delta_n) - \tau_n])} + e^{(a_n[(\theta_i - \delta_n) - \tau_n])}}, \tag{3.5}$$

where a_n, δ_n, and τ_n refer to the *n*ths subject's discrimination, location, and subjective threshold parameter. In contrast to dominance models, the item response function of Equation 3.5 is single-peaked, meaning that the probability that a student endorses a subject increase as the distance between θ_i and δ_n approaches zero. Additionally, τ_n is symmetric around δ_n, resulting in an item characteristic curve shaped like an inverted u where students with levels θ_i either -h or +h units away from δ_n are equally likely to select subject *n*. Following Holman and Glas (2005), a_n was constrained to be equal across all subjects in the selection model. As outlined in the conceptual framework, we assume students enroll in subjects they expect to perform well in, and therefore the latent choice propensity can be thought of as an overall proficiency-dimension.

Model 4 is an extension of Model 2, where the parameters in humanities and STEM subjects are estimated simultaneously with parameters of the selection model. Although other structures for IRT model-based approaches to missing data have been proposed (see Holman & Glas, 2005) a simple structure offers two advantages (Figure 1). Firstly, the factor loadings indicate the degree to which the observed choice-variables are related to the latent choice propensity. Secondly, the correlation between the proficiency dimensions and the choice dimension allows us to directly gauge the degree to which choice relates to competency (Pohl et al., 2014). Due to convergence problems, Model 4 was estimated by using parameter

Figure 1

Model 4 Factor Structure



Note. STEM = Science, technology, engineering, mathematics.

starting values from a GGUM model where τ_n was constrained to be equal across all groups. In the final model, τ_n was freely estimated in each subject.

Data cleaning and analysis was carried out with R, a programming language for statistical computing (R core Team, 2022), and the models were estimated with the mirt package (Chalmers, 2012). The full coding script can be found in Appendix 2. All models were identified by fixing the mean and variance of the latent traits to 0 and 1 respectively. Therefore, we also assume that the latent choice propensity is normally distributed. The models were estimated by marginal maximum likelihood estimation using 20^q quadrature points, where *q* refers to the number of dimensions, and the likelihood function was maximized using the expectation-maximization algorithm (Bock & Aitkin, 1981).

Expected grades

As previously mentioned, our second definition of difficulty is based on a hypothetical scenario where every student enrolled in all subjects. This scenario was simulated by computing expected grades, for each missing grade. The expectation that student *i* would receive grade *j* (*j* = 1,…,m) in subject *n*, was estimated by the posterior expectation of the best fitting model, calculated by

$$E(X_{ni}|\boldsymbol{x}_i, \boldsymbol{d}_i) = \sum_{j=1}^{m_n} j \int p(X_{ni} = j \mid d_{ni} = 1; \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_i, \boldsymbol{d}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i, \boldsymbol{d}_i)} \phi(\boldsymbol{\theta}; \mathbf{u}, \Sigma) d\boldsymbol{\theta}, \qquad (3.6)$$

where $p(\boldsymbol{x}_i, \boldsymbol{d}_i)$ is given by

$$p(\boldsymbol{x}_i, \boldsymbol{d}_i) = \int p(\boldsymbol{x}_i, \boldsymbol{d}_i|\boldsymbol{\theta})\ \phi(\boldsymbol{\theta}; \Sigma) d\boldsymbol{\theta}. \qquad (3.7)$$

θ refers to the joint distribution of the STEM and humanities dimensions for Model 2 and also includes the latent choice propensity for Model 4, $p(X_{ni} = j \mid d_{ni} = 1; \boldsymbol{\theta})$ denotes the conditional grade distributions of the models, and $\phi(\boldsymbol{\theta}; \Sigma)$ is the multivariate normal density function with a mean vector 0 and covariance matrix Σ. As the integrals in Equations 3.6 and 3.7 do not have closed-form solutions, they were numerically approximated using Gauss-

Hermite quadrature rules with 15 quadrature points.

Model Fit

To evaluate the relative fit of Models 0 through 3, we used the Bayesian information criterion (BIC). Compared to other information criteria, the BIC favors more parsimonious models, although they all tend to select the same model when the sample size is large (Dziak et al., 2019). To assess the degree to which the selection model was acceptable, we evaluated the fit of the GGUM model by itself. Two absolute fit measures, based on the M2 statistic (Maydeu-Olivares & Joe 2006), were used: the root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR). Thresholds of acceptable fit were based on the recommendations by Hu and Bentler (1999) and are .08 for the SRMR and .06 for the RMSEA. However, as we allowed for the choice variable to be NA, there were only 7,314 complete rows which these fit statistics were based on. As such, the fit statistics are only approximations of what they would have been with a complete data matrix.

Since no well-established absolute fit statistic exists for Model 4, we followed two relative fit statistics proposed by Korobko et al. (2008). Since Model 2 was used as the basis for Model 4, we compared these models relative to each other. Global fit was assessed through a likelihood ratio test, and for this, both models must refer to the same data. Therefore, the likelihood of Model 4 was compared to the product of the likelihood of Model 2 and the likelihood of the GGUM selection model by itself. This is equivalent to testing whether the covariances between the latent choice propensity and the competency dimensions are zero. We also utilized a local fit statistic, which is a modified version of the item fit statistic proposed by Korobko et al. (2008, p.146). The method used here is more descriptive and evaluates how well model-expected grades match the observed grades for groups with different proficiency distributions. These groups were identified by defining a splitter variable, which splits the sample into those that received a grade in subject *s* (d_{si} = 1) and those who did not (d_{si} = 0). The observed average grade for each group was calculated for all subject grades *n* (n = 1,…,17) as

$$S_{n0} = \left[\sum_i (1 - d_{si}) d_{ni} x_{ni}\right] \Big/ \left[\sum_i (1 - d_{si}) d_{ni} m\right]$$

and

$$S_{n1} = \left[\sum_i d_{si} d_{ni} x_{ni}\right] \Big/ \left[\sum_i d_{si} d_{ni} m\right],$$

where x_{ni} refers to the observed grade of student *i* on subject *n*. S_{n0} and S_{n1} were compared against the model-expected averages

$$E_{n0} = \left[\sum_i (1 - d_{si}) d_{ni} E(X_{ni}|\boldsymbol{x_i})\right] \Big/ \left[\sum_i (1 - d_{si}) d_{ni} m\right]$$

and

$$E_{n1} = \left[\sum_i d_{si} d_{ni} E(X_{ni}|\boldsymbol{x_i})\right] \Big/ \left[\sum_i d_{si} d_{ni} m\right],$$

where $E(X_{ni}|\boldsymbol{x_i})$ is given by Equation 3.6, which was approximated again, but now also for grades that were already observed in order to facilitate the comparison between observed and expected grades. For each subject, a summary measure of the deviations between observed and expected grades was calculated as $(E_{n0} - S_{n0})^{2} + (E_{n1} - S_{n1})^{2}$.

Results

Descriptive Statistics

The total sample consisted of 21,832 students from 308 schools. Although every student in the sample took all mandatory subjects for third year students, some were exempted from being graded, leading to NAs for these subjects as well (Table 3). In total, roughly 60% of grades in the sample were NA, which is higher than in other studies that have utilized similar methods (e.g., Korobko et al. 2008; Pohl et al. 2014; Rose et al., 2010). The table also shows the correlation between GPA and enrolling in a subject. Here we see

Table 3

Descriptive Statistics

Subject name	n	NA (%)	mean	GPA-choice correlation	SD	skew
History†	21,199	2.9	4.51	-	1.12	-0.66
Religion†	21,191	2.94	4.52	-	1.11	-0.65
Psychology	5,240	76	4.38	-.04	1.13	-0.54
Law	3,949	81.91	4.14	.06	1.15	-0.35
Politics	3,368	84.57	4.13	-.05	1.12	-0.39
Marketing	2,458	88.74	4.11	-.17	1.18	-0.41
Sociology	3,421	84.33	4.13	-.09	1.12	-0.47
Norwegian, written†	21,089	3.4	4.05	-	1	-0.26
Norwegian second-choice†	19,347	11.38	3.86	-	0.99	-0.14
Norwegian, oral †	21,111	3.3	4.54	-	1.05	-0.66
English, written	3,223	85.24	4.12	-.05	1.09	-0.39
English, oral	3,230	85.21	4.6	-.05	1.08	-0.81
Biology	3,003	86.24	4.29	.07	1.17	-0.47
Math, STEM	4,401	79.84	4.17	.22	1.32	-0.36
Math, sociological	4,378	79.95	3.75	.06	1.26	-0.13
Chemistry	3,840	82.41	4.11	.20	1.29	-0.34
Physics	2,562	88.26	4.1	.15	1.2	-0.22
Total	147,010	59.55	4.26	-	1.13	-0.57

Note. † Signals that the subject is mandatory. GPA = grade point average. STEM = Science, technology, engineering, mathematics. GPA-choice correlation indicates the correlation between d_n and GPA. All correlations were significant using a Bonferroni-adjusted alpha level of .05/11.

that students with higher GPAs are more likely to choose STEM subjects, signaling that the missing data is not MCAR. Moreover, these correlations do not take difficulty into account and will underestimate this relationship if STEM subjects are more difficult than other subjects. The mean grade in each subject does not signal that this is the case, but it is also a poor estimate of subject difficulty as it is potentially skewed by selection bias.

Dimensionality

Comparing Models 1 and 3 indicates that despite a strong common factor, academic competency is better conceived of as a multi-dimensional construct. First, Table 4 shows that the GRM provided a better representation of the data and was therefore used in the subsequent models. Table 4 also shows a reduction in the BIC, and therefore an increasingly

Table 4

Model Fit Comparison

Model (nr.)	Dimensions	Log likelihood	BIC	BIC difference
Uni-dimensional GPCM (0)	Academic competency	-163355	327728	-
Uni-dimensional GRM (1)	Academic competency	-163277	327571	-156
Two-dimensional GRM (2)	STEM, humanities	-161956	324940	-2631
Three-dimensional GRM (3)	STEM, social science, languages	-161002	323051	-1888

Note. BIC difference refers to the difference in BIC compared to the previous model. BIC = Bayesian information criterion. GPCM = generalized partial credit model. GRM = graded response model.

better fit, with more dimensions. However, even though dividing humanities into social science and languages led to a lower BIC, it is still questionable to perceive them as two separate dimensions because they have a factor correlation of .93 (Table 5). Because of this strong correlation, the Norwegian subjects correlate stronger with many social science subjects than with English, despite loading on the same factor. Hence, the common variance between the language subjects is largely indistinguishable from that of the social science subjects.

Despite the strong common factor found in Models 2 and 3, the results indicate that STEM subjects measure something distinct from humanities subjects. Although factor loadings are virtually equal for non-STEM subjects under all models, there is a clear difference for STEM subjects under Models 2 and 3. Here, all STEM subjects have very high factor loadings, ranging from .91 to .93. The squared standardized factor loading—or communality—in each STEM subject is therefore considerably higher when they are modelled as measuring a separate trait. For instance, the uni-dimensional construct of Model 1 explains 60% of the variance in physics, while this increases to 86% under the multi-dimensional models. More generally, the mean communality for STEM subjects is .64 for

Table 5

Parameters of Models 1,2, and 3

Dimension	Subject	Factor loadings		
		Model 1	Model 2	Model 3
Social sciences	History	.86	.86	.88
	Religion	.87	.87	.90
	Psychology	.86	.86	.87
	Law	.83	.83	.84
	Politics	.85	.85	.87
	Marketing	.85	.85	.85
	Sociology	.87	.87	.89
Language	Norwegian, written	.89	.89	.91
	Norwegian, second-choice	.91	.91	.93
	Norwegian, oral	.91	.91	.91
	English, written	.85	.84	.84
	English, oral	.82	.81	.80
STEM	Biology	.85	.92	.93
	Math, STEM	.78	.92	.93
	Math, sociological	.79	.91	.91
	Chemistry	.81	.93	.93
	Physics	.78	.93	.93

Correlation matrix

Model 2			Model 3			
	STEM	Humanities		STEM	SS	language
STEM	1		STEM	1		
Humanities	.82	1	SS	.84	1	
			Language	.79	.93	1

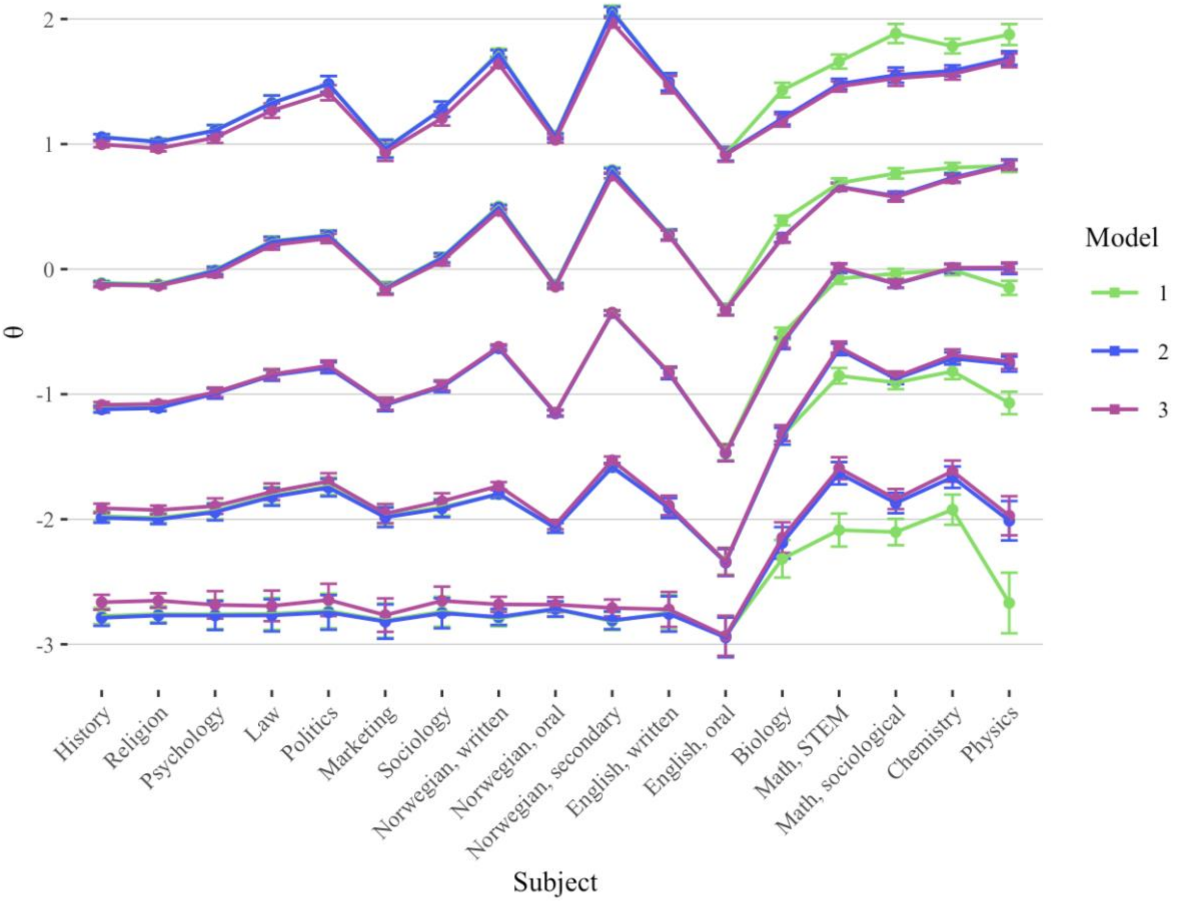
Note. STEM = science, technology, engineering, mathematics. SS = social science.

model 1, while it is .85 and .86 for model 2 and 3 respectively. Moreover, from a modelling-perspective, the difficulty point estimates change considerably for STEM subjects under the multi-dimensional models (Figure 1). While the point estimates for humanities subjects remain largely unchanged across all models, the difference in difficulty for STEM subjects is apparent.

Even though the BIC was lower for Model 3, Model 2 was augmented with the selection model for Model 4. A joint model of the missing data and a three-dimensional

Figure 2

IRT Difficulty Parameters for Models 1,2, and 3



Note. Each line represents IRT parameter category boundary locations on the latent scale, where the probability of getting grade *j* or higher is .50. Confidence intervals were calculated using the delta method.

structure of academic competency did not converge. The consequences of this might, however, be minor because the factor loadings and difficulty estimates between Models 2 and 3 are virtually identical. Moreover, the language dimension of Model 3 only contains grades from one elective subject—English—and therefore the factor correlation between languages and latent choice propensity would have been difficult to interpret.

Estimation Bias Due to Ignoring the Missing Data Mechanism

Comparing the likelihood of Models 2 and 4 shows a significantly smaller likelihood for Model 4 (χ^2 = 1524.12, df = 2, p < .001). Hence, the factor correlation between the latent choice propensity and proficiency dimensions are significantly different from zero. The model fit statistics for the selection model in isolation showed an RMSEA of .09 and SRMR of .10. According to the cutoff values of Hu and Bentler (1999), this constitutes less than acceptable fit. However, as Rose (2013) notes, model fit statistics are not the most important criteria for choosing the appropriate selection model, because the nuisance parameters are not of primary concern. The most important criterion is the degree to which the joint modelling of nuisance and focal parameters reduces bias in the latter. Although only more of a descriptive statistic, the item fit statistic (Table 6) shows that Model 4 generally predicts observed grades better than Model 2. Besides sociological math, the discrepancy between expected and observed mean grades is either lower or equal for all elective subjects under Model 4. The starkest difference is found for STEM subjects where Model 4 predicts observed grade means considerably better.

Table 6

Item Fit Statistic for Models 2 and 4

Subject	Model 2	Model 4
History	1.08	1.25
Religion	1.01	1.01
Psychology*	0.13	0.02
Law*	0.09	0.06
Politics*	0.43	0.01
Marketing*	0.30	0.01
Sociology*	0.43	0.01
Norwegian, written	0.57	0.64
Norwegian, oral	0.03	0.80
Norwegian, second-choice*	0.65	0.03
English, written*	0.57	0.19
English, oral*	0.60	0.18
Biology*	0.16	0.11
Math, STEM*	1.67	0.50
Math, sociological	0.10	0.24
Chemistry*	1.76	0.86
Physics*	2.09	0.36

Note. * Indicates a better fit for Model 4. STEM = science, technology, engineering, mathematics.

Comparing the parameters between Models 2 and 4 shows that the missing data mechanism led to parameter bias when ignored. To understand these results, it is instructive to first interpret the parameters of the selection model (Table 7). Firstly, the factor loading, which was constrained to be equal across all subjects, was .87 for the choice variables. In other words, the choices students make when enrolling in electives are highly related to the latent choice propensity. The choice location parameters in Table 7 indicate the peak of the ICC for each subject. For instance, the location of marketing is -1.46, meaning that the probability of choosing this subject increases as student's position on the latent trait

Table 7

Model 2 and Model 4 Parameters

Dimension	Subject	Factor loadings model 2	Factor loadings model 4	Choice location parameters
Humanities	History	.86	.85	-
	Religion	.87	.87	-
	Psychology	.86	.85	-0.62
	Law	.83	.83	-0.70
	Politics	.85	.85	-0.60
	Marketing	.85	.85	-1.46
	Sociology	.87	.87	-0.76
	Norwegian, written	.89	.91	-
	Norwegian, oral	.91	.90	-
	Norwegian, second-choice	.91	.90	-
	English, written	.84	.84	-.39
	English, oral	.81	.81	-.39
STEM	Biology	.92	.94	0.69
	Math, STEM	.92	.95	2.33
	Math, sociological	.91	.93	0.33
	Chemistry	.93	.95	1.12
	Physics	.93	.94	2.56

Correlation matrix

	Choice	STEM	Humanities
Choice	1		
STEM	0.60	1	
Humanities	0.37	0.82	1

Note. STEM = science, technology, engineering, mathematics.

continuum approaches this point. The general trend is clear: STEM electives are grouped towards the higher end of the latent trait continuum. Finally, the correlation between the latent choice propensity and humanities and STEM dimension was .34 and .60 respectively. This gives a direct indication of the degree to which students choice of elective subjects is related to competency. As the correlation is stronger for the STEM dimension, the choices students make are more related to their competencies in this dimension.

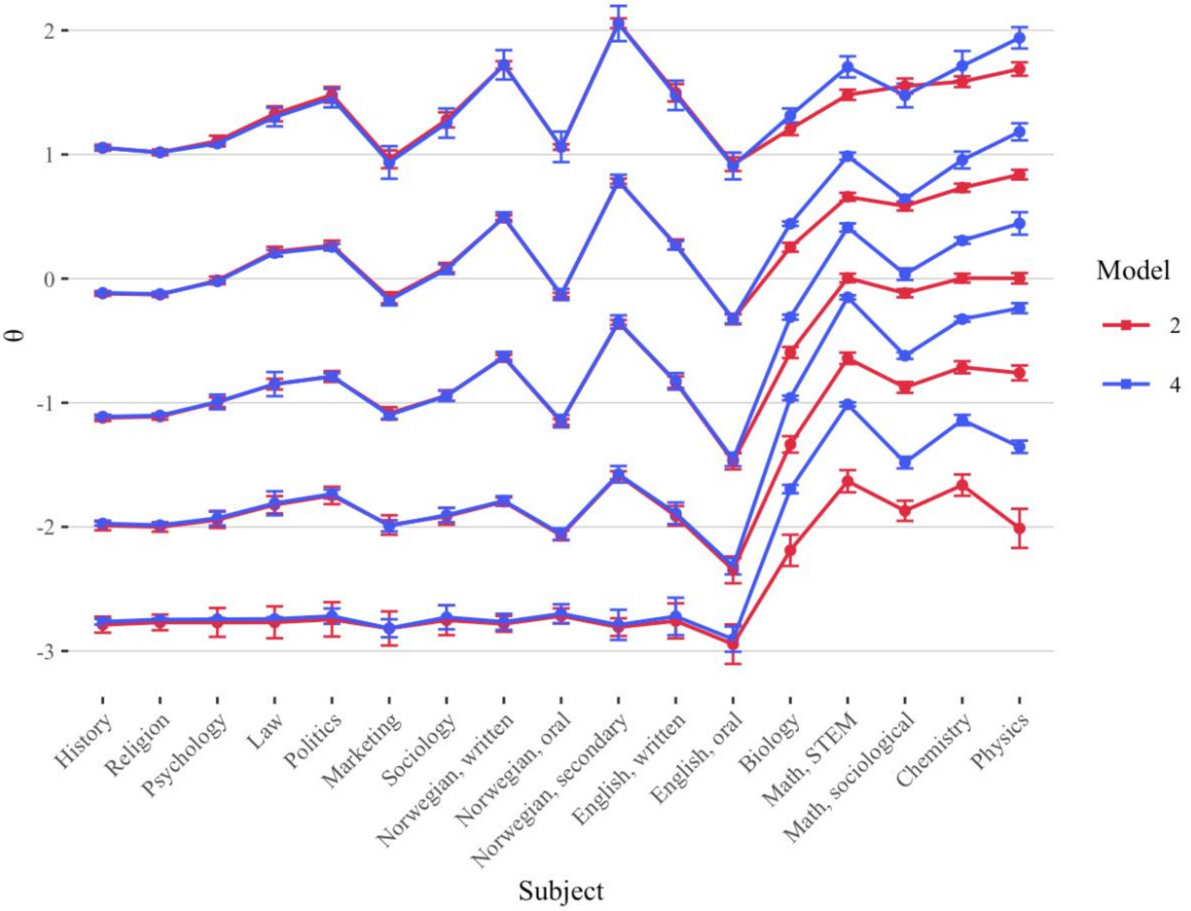
Comparing the item parameters of Models 2 and 4 shows that ignoring the missing data led to parameter bias. On one hand, the factor loadings between the models are largely comparable (Table 7), although there is a slight underestimation of factor loadings for STEM subjects and written Norwegian, which are 01-.03 higher under Model 4. On the other hand, the difficulty parameters differ substantially (Figure 3). Interestingly, while STEM subjects are noticeably more difficult under Model 4, the difficulty thresholds are virtually identical for humanities subjects. To better visualize the discrepancy between Models 2 and 4 for STEM subjects, Figure 4 shows the expected grade given the student's location on the STEM dimension. Here, we see that the discrepancy in difficulty between Models 2 and 4 is especially pronounced in the lower levels. For sociological math, there is virtually no difference in difficulty between the two models for achieving a grade of 4 or higher. Although the model fit comparisons indicated a better fit for Model 4, the fit statistics used to assess this were subject to some degree of uncertainty and should be interpreted with caution. Therefore, although the following section emphasizes the difficulty estimates of Model 4, we also provide results from Model 2.

Difficulty

Figure 3 corresponds to our first definition of difficulty: the conversion rate between levels of the latent trait and grades. The graph clearly shows that there are vast differences in inter-subject difficulty, and that STEM subjects require higher amounts of the underlying

Figure 3

Comparison of IRT difficulty parameters for Models 2 and 4

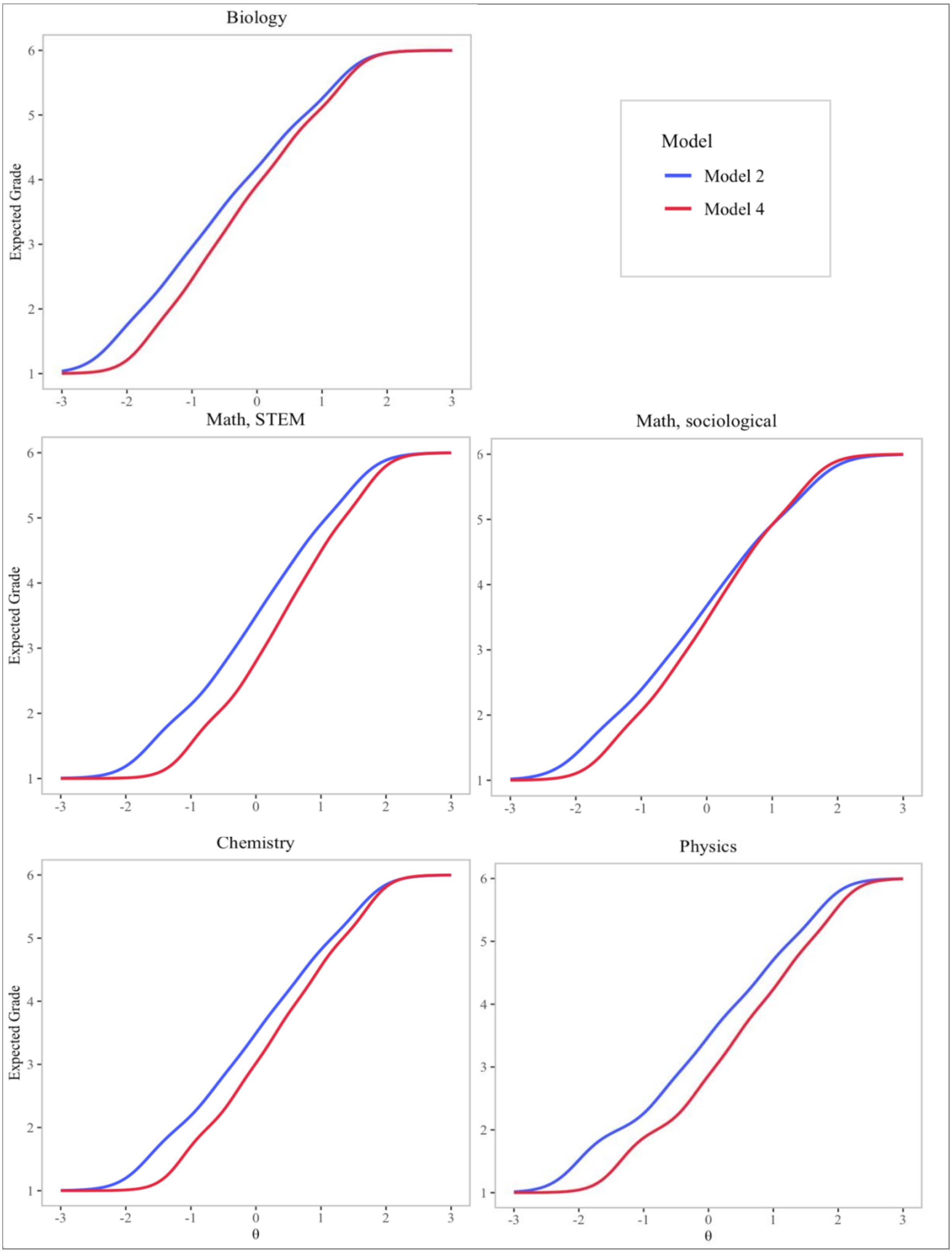


Note. STEM = science, technology, engineering, math. Each line represents IRT parameter category boundary locations on the latent scale, where the probability of getting grade *j* or higher is .50. Confidence intervals were calculated using the delta method.

construct to receive an equivalent grade in humanities subjects. For instance, under Model 4, a student with a latent trait value of -1 on both constructs is expected to receive a passing grade of 2 in STEM math but 4 on many of the humanities subjects. The discrepancy in difficulty is especially pronounced in the lower ability levels, but evens out for the highest levels of the latent trait continuum. Although not an elective subject, it is interesting to note that this relationship is reversed for second-choice Norwegian. Receiving a passing grade of 2 in this subject is as difficult as in the other humanities subjects, but it is the most difficult

Figure 4

Expected Subject Grade for Models 2 and 4



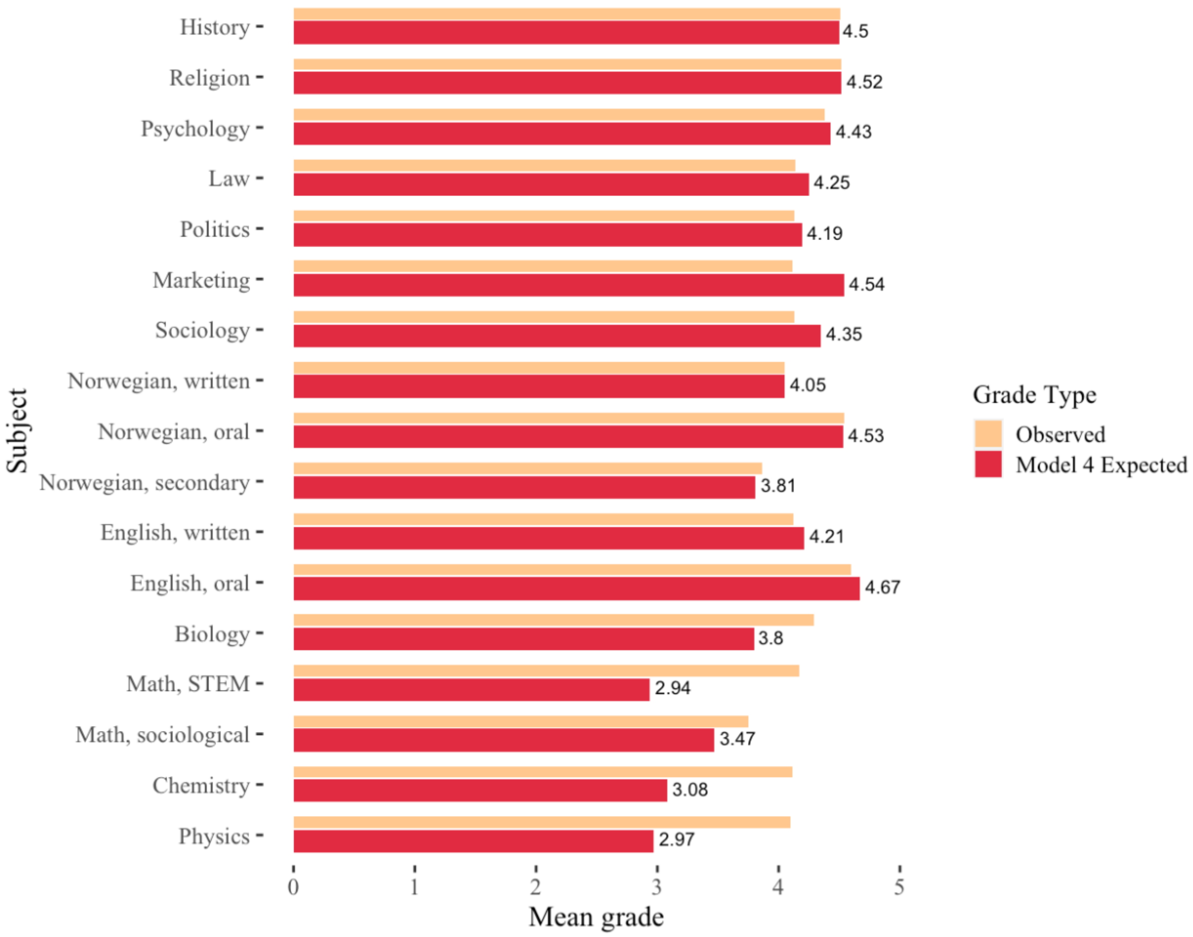
Note. STEM = science, technology, engineering, math. The plots show the expected grade given a student's location on the STEM construct.

subject to receive a grade of 6 in.

Since multi-dimensionality complicates the interpretability of our first definition, we provided a second definition of difficulty based on a hypothetical scenario where every student took all subjects. This scenario, based on Model 4, is shown in Figure 5, whereas the scenario based on Model 2 is shown in Appendix 3D. Even though the observed mean subject grades are relatively comparable, once selection bias is accounted for, the discrepancy in inter-subject difficulty is apparent. Most drastic is the mean difference of 1.73 between the easiest subject, oral English, and the hardest subject, STEM math. More generally, STEM subjects have grand mean of 3.25, while humanities electives have a grand mean of 4.38.

Figure 5

Expected subject grade means for Model 4



In Table 3, we provided the correlation between enrolling in a subject and GPA. However, as noted, these correlations were likely underestimated due to selection bias. Performing the same correlation, but now using model-expected grades from Model 4 instead of observed grades, provides a more accurate depiction of the degree to which choice of electives relates to academic competency (Table 8).

Discussion

At the outset of this study, we discussed potential issues regarding the assumption of the GPA—that subjects are equally difficult and measure the same construct. We further argued that these two aspects of the GPA—difficulty and dimensionality— are highly intertwined: if subjects measure different constructs, it is hard to provide any justifiable definition of difficulty. The results in this study showed that Norwegian high school subjects reflect performance in at least two constructs which are distinct for STEM and humanities subjects. At the same time, the correlation between these constructs is strong, signaling that variance in grades can largely be explained by an overarching construct of academic competency. To assess the degree to which a hypothesized self-selection bias distorts

Table 8

Correlation between GPA and subject enrollment

Subject name	GPA-enrollment correlation
Psychology	-.12
Law	-.13
Politics	-.11
Marketing	-.29
Sociology	-.17
English	-.09
Biology	.15
Math, STEM	.40
Math, sociological	.11
Chemistry	.34
Physics	.30

Note. All correlations were significant using a Bonferroni-adjusted alpha level of .05/11.

difficulty estimates, we jointly modelled academic competency and student choice of electives. The correlation between the latent choice propensity and competency traits showed that the choices student make is related to their academic competency, and especially their competency in STEM subjects. As a result, the difficulty estimates of Model 2 that ignored the missing data mechanism showed substantial bias in the STEM subjects. When this bias was accounted for in Model 4, we found considerable discrepancies in the mean grade of elective subjects.

IRT Difficulty Estimation with Non-ignorable Missing data

Our results provide insight into when the missing data mechanism must be accounted for in IRT parameter estimation, something which is not yet fully understood as model-based approaches to missing data are fairly new (Pohl & Becker, 2020). The correlation between the latent choice propensity and the STEM and humanities dimensions was .60 and .37 for respectively, something which signals a considerable degree of selection bias for both competency dimensions. However, when comparing the difficulty parameters of Models 2 and 4, only STEM subjects showed bias. One explanation could be the stronger correlation between choice and STEM as simulation studies have found that bias tends to increase with stronger correlation between the latent choice propensity and competency (Holman and Glas, 2005; Rose et al., 2010). However, the correlation of .37 between humanities and choice is still substantial. Moreover, the correlation between choice and STEM is lower than in some studies that have utilized similar methods and found that ignoring the missing data mechanism did not produce considerable bias (Glas & Pimentel 2005; Holman and Glas, 2005; Pohl et al. 2014; Rose et al. 2010). Another contributing factor could be the amount of missing data. While 51% of grades from humanities subjects were missing, the STEM subjects had a combined missing rate of 83%. However, 51% is far from negligible when selection bias is involved, and smaller amounts of missing data has led to considerable bias in other studies (Holman & Glas, 2005; Rose et al., 2010). Hence, the missing data mechanism and amount of missing data are necessary but not sufficient conditions to produce estimation bias.

The decisive factor that allowed for unbiased difficulty estimation of humanities subjects was most likely the missing data pattern. Whereas STEM subjects had large amounts of missing data for each subject, the humanities dimension included five mandatory subjects with few missing grades. Therefore, missing grades for these subjects constitute a case of MAR (Equation 2.2). The probability that these grades are missing is not stochastically independent from what they would have been if observed, but once conditioned on grades in the mandatory subjects, missingness becomes a purely random process which does not bias marginal maximum likelihood estimation. As there are no mandatory STEM subjects for third year high school students, the missing data cannot be ignored when estimating their difficulty. However, students do take mandatory STEM subjects in their first and second year. If these fit into the STEM construct proposed here, including them could allow the MAR assumption to hold for the this construct as well.

To our knowledge, Korobko et al. are the only other researchers who have examined subject comparability by utilizing IRT and a latent selection model, and so contrasting their results with ours is informative with regards to difficulty estimation with non-ignorable missing data. Contrary to our results, they found virtually no bias when ignoring the missing data, despite dealing with high amounts of missing data, no mandatory subjects, and a higher degree of selection bias than us. There are two potential explanations for this: firstly, they allowed subjects to load on multiple factors, and so the difficulty of subjects related to multiple latent traits. Hence, students were measured on multiple latent traits through a single subject, potentially leading to a MAR mechanism. They also utilized a multiple group model where groups were defined by students in similar clusters of subjects. If the missing data mechanism is conditional on some group-membership which can be observed and accounted for in estimation, the reduction in bias will be similar to that of the selection model utilized in this study (Demars, 2002; Rose et al., 2010; Rose, 2013). However, in the Norwegian high school context, such an approach is problematic. Following Korobko et al. (2008), we also tried to estimate a multi-dimensional, multiple group model where student groups were defined based on their specialization in either social science, languages, economic studies or STEM studies. However, since too few students in the former specialization also enroll in STEM subjects, the model would not converge. Ironically then, if students shy away from STEM subjects because they are more difficult, inter-subject difficulty estimation becomes more difficult.

The Multi-dimensionality of Teacher-assigned Grades

The results showed that although a general factor explains much of the variance in grades, STEM subjects measure something unique apart from other subjects. It is difficult to compare these findings with previous studies using teacher-assigned grades because results regarding the dimensionality of subjects are considerably influenced by the context and methodology of the study (Ofqual, 2015). Regardless, our results do mostly contrast with previous work. Both Coe et al. (2008) and Bowers (2011) found a two-dimensional structure of academic competency, but what separated these dimensions were academic and non-academic subjects such as arts and physical education. Coe et al. (2008) found an adequate model fit after removing non-academic subjects using a Rasch model, which is more restrictive than the GRM utilized in this study. The Rasch model was also used by He et al. (2015) and Veas et al. (2017) who found that a uni-dimensional model could describe differences in grades sufficiently well. These two studies were however not concerned about finding the best-fitting model, but used different criteria for assessing whether a uni-dimensional model fitted sufficiently well. Finally, Korobko et al. (2008) utilized an exploratory factor analysis, which resulted in a three-dimensional model consisting of a language, science, and economy dimension. However, as they allowed for cross-loadings, it is hard to interpret exactly what these dimensions mean.

One possible explanation for the two-dimensional structure found in this study is that the content domain of STEM subjects requires distinct academic skills. An expert panel commissioned by the department of education criticized the content domain of STEM subjects for overlapping too much compared to other subjects (KD, 2015). There is support for this under all the multi-dimensional models, where the factor loadings of STEM subjects were very high. In other words, there is not much unique variance associated with each subject beyond that of the common STEM construct. The distinction between STEM and humanities subjects could also be due to the nature of teacher-assigned grades. It is a well-established and replicated finding in educational research that teacher-assigned grades reflect a multi-dimensional construct consisting of both cognitive and behavioral factors (Brookhart et al., 2015; Bowers, 2011). These behavioral factors include aspects such as punctuality, effort, attention, and participation. Studies reveal that teachers in STEM use these factors less, findings which also have been replicated in Norwegian high school (Ofqual, 2015; Prøitz, 2013).

Differences in Difficulty

Our first, and strictest definition of difficulty stated that subjects can only be compared with regards to a common linking construct. However, since subject grades relate to different constructs, doing so is not possible. One possible way out, which is still consistent with this definition, but limits the number of possible comparisons, is to compare the difficulty of subjects within each construct. Yet, if this common factor is not strong enough, we risk ignoring the skills, knowledge, and understanding that is specific to each subject (Newton, 1997). For instance, the subjects law and marketing both have a factor loading of .85 under Models 2 and 4, and therefore 28% of their variance cannot be explained by the humanities construct. This residual variance is not necessarily irrelevant to the subjects' competency goals. When we then limit our comparison of the two subjects to their common construct, we could end up excluding important and unique competencies that they measure. Since STEM subjects have a very high factor loading under the multi-dimensional models, we can more validly compare them through our first definition of difficulty. For instance, under Model 4, biology and physics have factor loadings of .94. As Figure 4 shows, it requires a substantially higher ability levels in physics to receive equivalent grades in biology, indicating that physics is more stringently graded than biology.

Our second definition of difficulty, the nominal approach, compared the mean grades in a simulated scenario where every student took every subject. This scenario showed that STEM subjects are considerably more difficult than other subjects. This is not unique to Norwegian high school, as it has been found across the world in vastly different school systems (Ofqual, 2015). Potential explanations for this are manifold and could for instance be related to our findings that STEM and humanities subjects measure different constructs. If non-STEM teachers to a larger degree utilize behavioral factors when grading, and these generally push grades upwards, this could potentially explain the discrepancy. Another hypothesis, often supported by the Norwegian government, is that Norwegian students generally lack motivation and interest for STEM subjects (KD, 2010; KD, 2015). However, since this study examined STEM subjects freely chosen by students, they are most likely more motivated than the general student population. This is support by Holmseth (2013) who found that interest was equally important for STEM and non-STEM students when choosing electives.

If the multi-dimensional structure we found in our analysis stems from differential competencies demanded of STEM versus humanities subjects, the lower model-expected grades for STEM subjects could be due to non-STEM students having lower abilities in this domain. We earlier argued, based on value-expectancy theory, that students partly choose electives based on their interest and competencies. It can then be argued that our hypothetical scenario where non-STEM students enroll in STEM subjects is pointless, because students should be allowed to choose electives they find interesting, and therefore are more likely to do well in. However, when correcting for selection bias, the humanities electives became easier, signaling that STEM students are generally more academically adept, and that their GPA is punished by enrolling in STEM subjects.

Limitations and Future Research

There are a few limitations in this study worth noting. Firstly, the selection model is based on untestable assumptions, such that the latent choice propensity is uni-dimensional, normally distributed, and linearly related to the competency dimensions. These assumptions are problematic largely due to not being able to assess the model fit of Model 4. As most fit statistics are based on discrepancies between observed and expected scores, model fit procedures for selection models are generally problematic (Enders, 2022). Future research can therefore test the robustness of our findings through other approaches to dealing with missing data. Other types of selection models, latent regression models, pattern mixture models, or multiple imputation with auxiliary variables are possible avenues (See Holman & Glas 2005; Rose, 2013; Enders, 2022). The context of inter-subject comparability in Norwegian high school presents an atypical situation with large amounts of multi-dimensional missing data. Therefore, such studies could also more generally further our understanding of how to deal with missing data.

With regards to dimensionality, another limitation is that the factor structure found in this study is based on the inclusion of a limited set of electives. Although these were a representative selection of the most common electives, including further subjects could have provided evidence for a different factor structure. For instance, only one language elective was included, and so including more language electives could have helped clarify the validity of this construct. Future research that uses more sophisticated methods tailored specifically to assess the dimensionality of subjects could provide knowledge about the ways in which we can compare subjects in Norwegian high school. Lastly, the removal of a considerable portion of the sample limits the generalizability of our findings. Although IRT models assume item invariance, this is not necessarily the case if the items—or subjects—show differential item functioning (DIF) for students who were removed from this study. DIF is also a key concern in the subject comparability literature (See Newton, 2011; Ofqual, 2015) because differences in difficulty could potentially be due to different inclinations of demographic sub-groups to enroll in certain subjects. Checking for DIF across subjects was outside the scope of this study but could provide another fruitful avenue for future research.

Conclusion

In the introduction we discussed a report commissioned by the Ministry of Education that proposed to remove the bonus points awarded by STEM subjects on the basis that GPAs are comparable (Official Norwegian Reports, 2022). STEM points were implemented for two reasons: to increase recruitment into these subjects while also compensating for the lower grades they award. Our study has implications with regards to both these goals. Firstly, we have seen that more competent students are much more likely to choose STEM electives. Removing STEM points could therefore further dissuade less competent students from choosing them. This is worrying because STEM is not only important for students who need them as prerequisites for admission into certain university programs, but also for the scientific literacy of the population and consequently democracy (KD, 2010). Secondly, although there could be merits to removing STEM points to streamline university admission processes, the basis for such a decision should not be made on the grounds that GPAs are comparable. As we have demonstrated in this study, there are not only stark differences in the difficulty of subjects, but they also measure distinct traits. Consequently, the GPA is not a measure of a single construct, but rather of slightly different constructs whose meaning depends on the difficulty and dimensionality of the set of subjects that a student chooses.

References

Bock, D. R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *17*(3), 141–159. https://doi.org/10.7916/D8WM1QF6

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Coe, R. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch model. *Oxford Review of Education*, *34*(5), 609–636. https://doi.org/10.1080/03054980801970312.

Craig K. Enders. (2022). *Applied Missing Data Analysis (methodology in the social sciences)* (2nd ed.). The Guilford Press.

Crofts, J. M., & Jones, D. C. (1928). Review of Secondary School Examination Statistics. *The Mathematical Gazette*, *14*(197), 276–277. https://doi.org/10.2307/3607819

DeMars, C. (2002). Incomplete Data and Item Parameter Estimates Under JMLE and MML Estimation. *Applied Measurement in Education*, *15*(1), 15–31. https://doi.org/10.1207/S15324818AME1501_02

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2019). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, *21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Eccles, jacquelynne, Adler, T. E., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In *Achievement and achievement motives—Psychological and sociological approaches*. Freeman and Company.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing fata. *Journal of Educational Measurement*, *45*(3), 225–245. https://www.jstor.org/stable/20461894

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907–922. https://doi.org/10.1177/0013164408315262

He, Q., Stockford, I., & Meadows, M. (2018). Inter-subject comparability of examination standards in GCSE and GCE in England. *Oxford Review of Education*, *44*(4), 494–513. https://doi.org/10.1080/03054985.2018.1430562

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161. https://doi.org/10.2307/1912352

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Holmseth, S. (2013). *Utdanning 2013 – fra barnehage til doktorgrad* (No. 138; Utdanning, p. 190). Statistisk Sentralbyrå. https://www.ssb.no/utdanning/artikler-og-publikasjoner/_attachment/153399?_ts=144d059ffb8

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Inge Ramberg. (2006). *Realfag eller ikke? Elevers motivasjon for valg og bortvalg av realfag i videregående opplæring [STEM or not? Students' motivation for choosing and not choosing STEM subjects in upper secondary education]* (Working Paper No. 43). Nordisk institutt for studier av innovasjon, forskning og utdanning.

Kjærnsli, M., & Lie, S. (2011). Students' preference for science careers: International comparisons based on PISA 2006. *Internaational Journal of Science Education*, *33*(1), 121–144. https://doi.org/10.1080/09500693.2010.518642

Korobko, O. B., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the Difficulty of Examination Subjects with Item Response Theory. *Journal of Educational Measurement*, *45*(2), 139–157. https://doi.org/10.1111/j.1745-3984.2007.00057.x

Lødding, B., Daus, S., Reiling, R. B., Bungum, B., Vika, K. S., & Bergene, A. C. (2021). *Realistiske forventninger? Sluttrapport fra evalueringen av Tett på realfag. Nasjonal strategi for realfag i barnehagen og grunnopplæringen (2015–2019) [Realistic expectations? Final report from the evaluation of Getting closer to STEM. National strategy for science in kindergartens and primary education (2015-2019)]*. The Nordic Institute for Studies in Innovation, Research and Education. https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/realistiske-forventninger-sluttrapport-fra-evalueringen-av-tett-pa-realfag/

Maydeu-Olivares, A., & Joe, H. (2006). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Ministry of Education and Research. (2005). *Vilje til Forskning[The Will for Research]* (White Paper No. 20). Ministry of Education and Research of Norway. https://www.regjeringen.no/no/dokumenter/stmeld-nr-20-2004-2005-/id406791/

Ministry of Education and Research. (2010). *Realfag for framtida: Strategi for styrking av realfag og teknologi 2010–2014[STEM Subejcts for the Future: Strategy to Strengthen STEM and Technology 2010-2014]*. Ministry of Education and Research of Norway. https://www.regjeringen.no/globalassets/upload/kd/realfagstrategi.pdf

Ministry of Education and Research. (2022). *The education system in Norway*. Information for Newly Arrived Parents and Guardians: The Education System in Norway. https://www.udir.no/laring-og-trivsel/minoritetsspraklige-og-flyktninger/minoritetsspraklige/informasjon-til-nyankomne/information-for-newly-arrived/

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an Em Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Newton, P. E. (1997). Measuring Comparability of Standards between Subjects: Why our statistical techniques do not make the grade. *British Educational Research Journal*, *23*(4), 433–449. https://doi.org/10.1080/0141192970230404

Newton, P. E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, *25*(3), 285–292. https://doi.org/10.1080/02671522.2010.498144

Newton, P. E. (2011). *Full article: Making sense of decades of debate on inter-subject comparability in England*. *19*(2), 251–273. https://doi.org/10.1080/0969594X.2011.563357

Official Norwegian Reports. (2022). *Veier inn – ny modell for opptak til universiteter og høyskoler[Ways in—A New Model For Admission into Universities and Colleges]* (No.

17). Ministry of Education and Research of Norway. https://www.regjeringen.no/no/dokumenter/nou-2022-17/id2948927/

Pohl, S., & Becker, B. (2020). Performance of missing data approaches under nonignorable missing data conditions. *Methodology*, *16*, 147–165. https://doi.org/10.5964/meth.2805

Pohl, S., Gräfe, L., & Rose, N. (2014). *dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models*. *74*(3), 423–452. https://doi.org/10.1177/0013164413504926

Prøitz, T. S. (2013). Variations in grading practice – subjects matter. *Education Inquiry*, *4*(3), 555–575. https://doi.org/10.3402/edui.v4i3.22629

R Core Team (2022) *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses. *Applied Psychological Measurement*, *24*(1), 3–32. https://doi.org/10.1177/01466216000241001

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* [Doctoral dissertation, Friedrich Schiller University]. https://www.db-thueringen.de/receive/dbt_mods_00022476

Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), 1–53. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.2307/2335739

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. https://doi.org/10.1007/BF03372160

The Norwegian Directorate for Education and Training. (n.d.). *Elevtall i videregående skole – utdanningsprogram og trinn [Student numbers in high school—Educational program and grade]*. The Norwegian Directorate for Education and Training. Retrieved April 23, 2023, from https://www.udir.no/tall-og-forskning/statistikk/statistikk-videregaende-skole/elevtall-i-videregaende-skole/elevtall-vgo-utdanningsprogram/

The Norwegian Directorate for Education and Training. (2020a). *Utdanningsspeilet 2020 [The Education Mirror 2020]*. https://www.udir.no/tall-og-forskning/publikasjoner/utdanningsspeilet/utdanningsspeilet-2020/

The Norwegian Directorate for Education and Training. (2020b). *Vurderinger og anbefalinger om fremtidens eksamen [Evaluations and Recommendations of Future Exams]*. https://www.udir.no/eksamen-og-prover/eksamen/vurderinger-og-anbefalinger-fremtidens-eksamen/

The Norwegian Directorate for Education and Training. (2020c). *Individuell vurdering Udir-2-2020 [Individual evaluation Udir-2-2020]*. regelverkstolkninger frå Udir [regulatory interpretations from Udir]. https://www.udir.no/regelverkstolkninger/opplaring/Vurdering/udir-2-2020-individuell-vurdering/iv.-fritak-fra-vurdering-med-karakter/

The Norwegian Directorate for Education and Training. (2023a). *Trekkordning ved eksamen for grunnskole og videregående opplæring Udir-2-2018 [Draw system for exams for primary and secondary education Udir-2-2018]*. regelverkstolkninger frå Udir [regulatory interpretations from Udir]. https://www.udir.no/regelverkstolkninger/opplaring/eksamen/trekkordning-ved-eksamen-for-grunnskole-og-videregaende-opplaring-udir-2-2018/

The Norwegian Directorate for Education and Training. (2023b). *Føring av vitnemål og kompetansebevis for videregående opplæring i Kunnskapsløftet – 2023 [Recording of *transcripts and competency certificates for upper secondary education in the Knowledge Promotion Reform—2023]*. The Norwegian Directorate for Education and Training. https://www.udir.no/eksamen-og-prover/dokumentasjon/vitnemal-og-kompetansebevis/foring-vitnemal-kompetansebevis-vgs/1-dokumentasjon-i-videregaende-opplaring/

The Office of Qualifications and Examinations Regulation. (2015). *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper*. https://www.gov.uk/government/publications/inter-subject-comparability-a-review-of-the-technical-literature

Tveitereid, M., Dahl, M., Kaels, A. K., Pedersen, T., Sletbak, B., Vikhals, S., Ballestad, M., Flakstad, H., Kjelberg, A., Prestøy, L. A., Stavlund, E., Jorde, D., Lorentzen, L., Reikerås, E. K. L., & Tandberg, A. H. S. (1997). *Matematikk, naturvitenskap, teknologi:tiltak for å styrke disse fagområdene i norsk utdanning: Sluttrapport*. Kirke, utdannings- og forskningsdepartementet. https://www.nb.no/items/463e2c01e4ceabb866c5979bc6cb3237

Veas, A., Gilar, R., Miñano, P., & Castejón, J. L. (2017). Comparative analysis of academic grades in compulsory secondary education in Spain using statistical techniques. *Educational Studies*, *43*(5), 533–548. https://doi.org/10.1080/03055698.2017.1312287

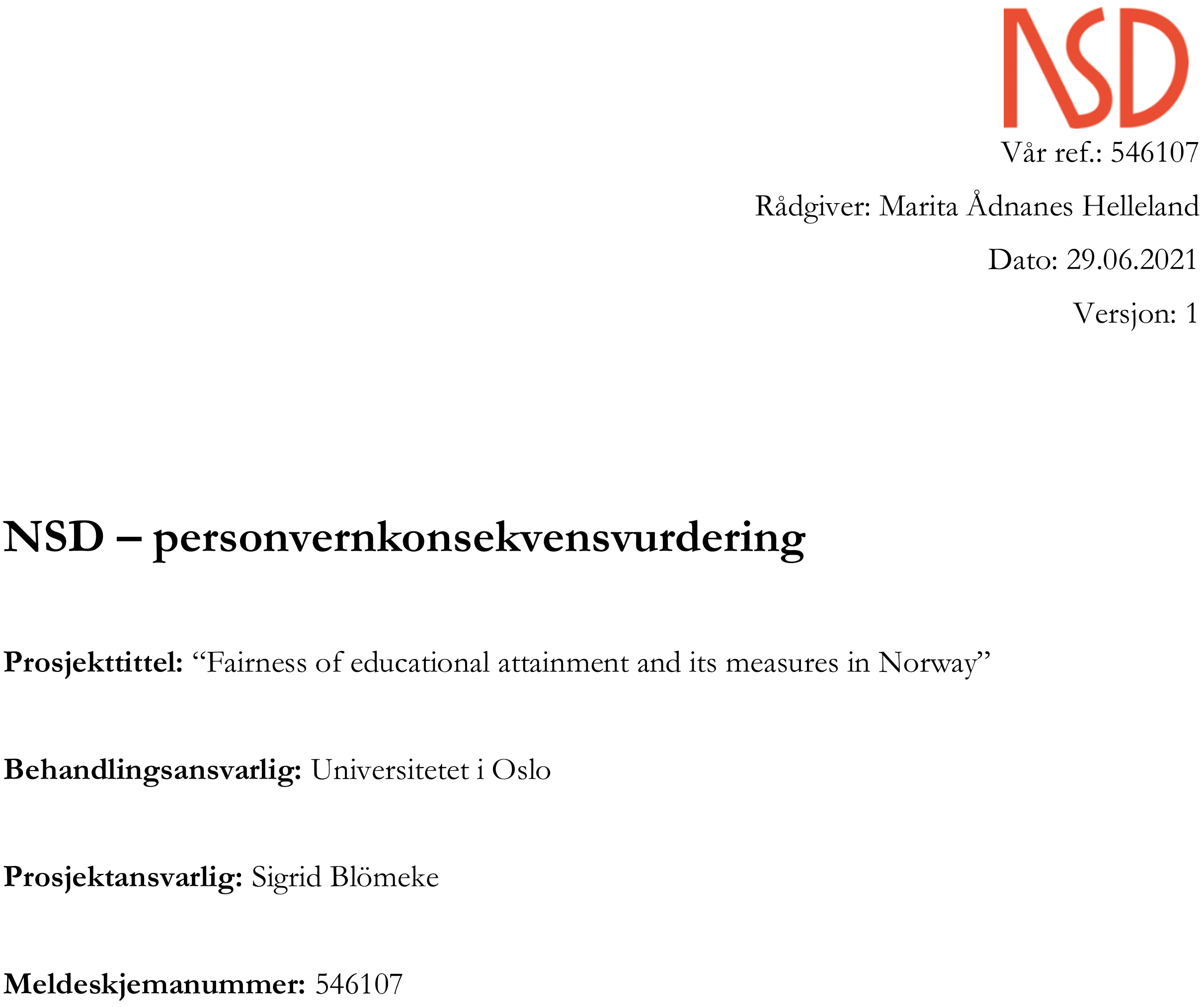
Vilbli. (n.d.). *Education programmes—Specialization in general studies*. Vilbli. Retrieved November 24, 2022, from https://www.vilbli.no/en/en/no/specialization-in-general-studies/program/v.st

Vinje, E. E. (2021). *Didaktiske utfordringer i kroppsøving* (1st ed.). Cappelen Damm Akademisk.

Appendix I – Data Management and Protection

A. Data Protection Impact Assessment

The data utilized in this study is part of a larger project, “fairness of educational attainment and its measures in Norway”, that utilize and manage the same data in an external environment. The following data protection impact assessment and its approval of this project was carried out by the Norwegian Agency for Shared Services in Education and Research.



Vår ref.: 546107

Rådgiver: Marita Ådnanes Helleland

Dato: 29.06.2021

Versjon: 1

NSD – personvernkonsekvensvurdering

Prosjekttittel: “Fairness of educational attainment and its measures in Norway”

Behandlingsansvarlig: Universitetet i Oslo

Prosjektansvarlig: Sigrid Blömeke

Meldeskjemanummer: 546107

Endring 08.07.2021

Det er i 1.3 Datakilder, type og omfang av personopplysninger tilføyd at prosjektet legger til en variabelliste fra SSB, som gjelder søknadsdata til videregående skole fra søknadsportalen Vigo. Variabellisten er lastet opp i meldeskjemaet, under "Tilleggsopplysninger". Forskerne vil be om at indirekte identifiserende variabler som skolenavn, skolenummer og spesifikke datoer blir utelatt fra datasettet (se 2.2.2 Dataminimering).

NSD sin samlede vurdering av endring 08.07.2021

Det er NSD sin vurdering er at dette ikke endrer vesentlig på den personvernvurderingen UiO allerede har godkjent. Endringen er oversendt UiO til orientering.

Om konsekvensvurderingen (DPIA)

NSD har gjennomgått innholdet i meldeskjemaet. Det er vår vurdering at den planlagte behandlingen av personopplysninger vil innebære relativt høy risiko for de registrertes rettigheter og friheter, og dermed krever en personvernkonsekvensvurdering (DPIA) jf. personvernforordningen art. 35.

Dette fordi den planlagte behandlingen av personopplysninger innebærer

- behandling av særlige kategorier av personopplysninger (sensitive opplysninger) eller opplysninger av svært personlig karakter
- behandling av personopplysninger i stor skala, både med hensyn til utvalgsstørrelse, mengde opplysninger, varighet og regelmessighet.

På oppdrag fra ledelsen ved Universitetet i Oslo har NSD i samråd med prosjektansvarlig og rådgivere ved institusjonen laget utkast til en DPIA som inneholder

1. en systematisk beskrivelse av den planlagte behandlingen av personopplysninger
2. vurdering av om behandlingsaktivitetene er nødvendige og står i rimelig forhold til formålene
3. analyse av risiko for de registrertes rettigheter og friheter
4. planlagte tiltak for å håndtere risikoene.

Ved å følge de planlagte tiltakene, mener NSD at personvernrisikoen er redusert i en slik grad at behandlingen kan gjennomføres i samsvar med personvernforordningen, uten forhåndsdrøfting med Datatilsynet.

Behandlingsansvarlig institusjon (ved ledelsen) bestemmer om personvernkonsekvensvurderingen er tilfredsstillende utført, og om personvernrisikoen er redusert til et akseptabelt nivå slik at behandlingen kan gjennomføres, eller om det er nødvendig med forhåndsdrøfting. Dette gjøres etter rådføring med institusjonen sitt personvernombud. Vi oversender derfor vår vurdering til oppgitt kontaktperson ved institusjonen. NSD ber om at den godkjente versjonen av DPIA lastes opp til meldeskjema av prosjektansvarlig.

Dersom behandling av personopplysninger igangsettes på grunnlag av DPIA, og deretter endres, minner vi om at endringene kan medføre behov for ny/oppdatert DPIA. Prosjektansvarlig skal melde endringer til NSD, og institusjonen har ansvar for å påse at dette skjer. Ved melding om endringer i prosjektet, vil NSD bistå med denne.

Følgende personer har deltatt i personvernkonsekvensvurderingen:

Navn	Rolle/funksjon	Virksomhet
Sigrid Blömeke	Prosjektansvarlig	UiO
Astrid Marie Jorde Sandsør	Prosjektdeltaker	UiO
Marita Ådnanes Helleland	Seniorrådgiver	NSD
Siri Tenden	Seniorrådgiver	NSD

1. Systematisk beskrivelse av planlagte behandlingsaktiviteter og formål

Her følger en beskrivelse av den planlagte behandlingen av personopplysninger, slik den er oppgitt i meldeskjema med vedlegg og etter dialog med prosjektansvarlig. Vurdering av behandlingen følger i del 2 og 3.

1.1 Formål

Et grunnleggende mål med utdanningssystemet er å sikre at alle barn får like utdanningsmuligheter. Allikevel er det slik at det er store utdanningsforskjeller på tvers av kjønn, etnisitet, sosioøkonomisk bakgrunn og/eller kontekstuelle kjennetegn som nabolag og skolekvalitet. Dette prosjektet har som mål å skille de kausale mekanismene bak disse forskjellene fra hverandre ved å se individuelle kjennetegn og kontekst i sammenheng med ferdighetsutviklingen i løpet av utdanningsløpet og senere i livet. Det er særlig lite kunnskap om hvorvidt norske utdanningsreformer har påvirket utdanningsulikhet, som vil være et særskilt fokus i dette prosjektet. I tillegg har prosjektet som mål å kunne skille endringer i utdanningsulikhet fra endringer i ulikhet langs andre samfunnsmessige dimensjoner. Til slutt har prosjektet som mål å kunne undersøke hvor rettferdige utdanningsmål er ved å undersøke hvordan partiskhet (bias) og andre måleeffekter påvirker utdanningsulikhet.

1.2 Registrerte

Prosjektet er en registerdatastudie og består av et utvalg – hele Norges befolkning så langt bak dataene til Statistisk sentralbyrå (SSB) strekkes seg for de aktuelle variablene. Dette inkluderer alle personer bosatt i Norge fra og med 1970 og frem til i dag, kull på ca. 50 000-68 000 individer hvert år, ca. 9 millioner til sammen. Alle variabellister er knyttet til dette utvalget. Dataene i utvalget kobles sammen via avidentifiserte løpenummer slik at individer kan kobles til foreldre, husholdning og organisasjon (skole, bedrift).

1.3 Datakilder, type og omfang av personopplysninger

Det skal innhentes opplysninger fra Utdanningsdirektoratet (Udir), UNIT og SSB. Data kobles og leveres ut fra SSB.

Fra SSB hentes opplysninger om befolkning, utdanning, inntekt, kontantstøtte og sysselsetting. Forskerne får data om lærere fordi de finnes i utvalget, og man kan identifisere hvem som jobber på skoler og har lærerutdanning. Da kan forskerne si noe mer detaljert om lærere som jobber ved skolen enn det man får fra GSI som bare er antall lærere. Data fra SSB inkluderer også opplysninger om søknad og opptak til videregående utdanning fra 2002 og frem til i dag. Disse opplysningene kommer fra fylkeskommunenes administrative datasystem for opptak til videregående opplæring (VIGO) og forvaltes av SSB. Se variabellisten «Grensesnitt_UDIR_SSB» under Tilleggsopplysninger.

I tillegg skal det hentes opplysninger fra Grunnskolens informasjonssystem (GSI) via Udir. Her vil forskerne få informasjon om elevtall, årstimer, årsverk og ansatte, spesialundervisning, språklige minoriteter, fremmedspråk, fysisk aktivitet og leksehjelp, valgfag og skolebibliotek, SFO, korona og PPT siden 1992. Det er nødvendig å koble på opplysninger fra GSI fordi det her finnes informasjon om skolen som kan kobles på elevers skoletilhørighet, slik at forskerne kan si noe om f.eks. lærertetthet ved skolen elevene går på.

Dataene sendes til SSB og gis samme avidentifiserte organisasjonsnummer som de øvrige dataene slik at de kan kobles sammen.

Fra UNIT hentes opplysninger om høyere utdanning og søknad til høyere utdanning (kurs, karakterer, eksamener og grader fra høyere utdanning, samt søknad til høyere utdanning). Variablene i datasettet om søknad til høyere utdanning inkluderer beståttkoder eller resultatkoder for ulike fag som man trenger for å søke opptak til studier, detaljer om studieprogram/studiested og ulike typer kvoter søkere er tatt inn på, og informasjon om søkere har takket ja/nei til plass og venteliste. Dette er viktig å ha med for å vurdere om søkerne fikk avslag f.eks. pga. manglende poeng eller manglende fag. I datasettet om høyere utdanning hentes opplysninger om hvilke studenter som er tatt opp hvor og bakgrunnsopplysninger om studentene (karakter fra videregående skole, kjønn). Prosjektet trenger dataene for å følge studieløp på ulike studieprogram og fag med tilhørende karakterer, og for å kunne forstå hva som predikerer studieprogresjon og resultater på studiene.

Dataene sendes til SSB og gis samme avidentifiserte løpenummer som de øvrige dataene slik at de kan kobles sammen. UNIT fjerner eventuelle variabler som er mer identifiserende enn det forskerne får fra SSB (f.eks. fødselsdato og postnummer).

Det behandles alminnelige personopplysninger om f.eks. kommune, grunnkrets, utdanning, inntekt, yrke, familiestatus osv. Forskerne har behov for grunnkrets for å kunne identifisere sannsynlig skoletilhørighet i perioder hvor registerdataene ikke identifiserer dette (f.eks. i perioden før nasjonale prøver 5. trinn), og for å kunne se på betydning av nabolag i kontekstanalyser. Forskerne vil be SSB om at opplysninger om inntekt avrundes.

Det behandles særlige kategorier av personopplysninger om helseforhold (opplysninger om nedsatt arbeidsevne, sosialhjelp, sykepenger, uførestønad o.l.) i dataene fra lønnsstatistikk og fra FD-trygd.

Fullstendige variabellister er vedlagt meldeskjemaet og er også tilgjengelige i Meldingsarkivet.

1.4 Kontakt med de registrerte

Hovedprosjektet vil ikke være i direkte kontakt med utvalget. Grunnet utvalgets størrelse, vil det ikke være praktisk mulig å innhente samtykker eller å gi informasjon til den enkelte registrerte. Informasjon om prosjektet som helhet vil gjøres offentlig tilgjengelig på UiO sine nettsider.

For de registrertes øvrige rettigheter vises det til punkt 2.3.

1.5 Dataflyt – hvordan personopplysningene behandles

UNIT, Udir oversender data til kobling av SSB i henhold til sine og SSB sine rutiner. All kobling av data skal foretas av SSB, som også skal oppbevare koblingsnøkler i prosjektperioden. SSB skal gjøre uttrekk og overføre data uten direkte identifiserende opplysninger direkte til TSD eller til forskningsserver ved UiO dersom den direkte løsningen ikke enda er tilgjengelig.

Oversendelse av sensitive data fra/til SSB skjer via tunnelen fx.ssb.no og videre til TSD. Portalen fx.ssb.no vil være tilgjengelig for prosjektleder og krever brukernavn og passord. Passord og brukernavn vil bli tilsendt separat på SMS. Koblingsnøkkel vil lagres separat hos SSB og utleveres ikke til UiO. Ved sending av data vil disse krypteres med passord som sendes på SMS i tråd med SSB sine retningslinjer.

TSD er spesielt utviklet for sikring og prosessering av sensitive data. Serveren er passordbeskyttet og frikoblet fra nett. All tilgang krever to-faktor autentisering. Data vil etter kobling bli behandlet uten direkte identifiserende kjennetegn. Forskerne i prosjektet vil ikke ha tilgang til koblingsnøkler, og hvis det er behov for å gjøre nye koblinger med oppdaterte årganger er det SSB som må gjøre dette.

1.6 Tilgang til personopplysninger

Følgende personer vil ha tilgang til datamaterialet:

Virksomhet	Ca. antall medarbeidere	Rolle/funksjon	Tilgang til alle personopplysninger?	Hvordan får de tilgang?
Universitetet i Oslo	1	Prosjektansvarlig	Ja	TSD
Universitetet i Oslo	ca. 7	Prosjektmedarbeidere	Ja	TSD

I prosjektet vil det være noen variabler som kan være særlig egnet til å identifisere enkeltpersoner. Dette gjelder spesielt grunnkrets, inntekts- og arbeidsopplysninger, og fødselsdato (måned år). Prosjektmedarbeidere som meldes til prosjektet nå vil ha tilgang til alle variablene i og med at de jobber på tvers av delprosjektene. Dersom nye medarbeidere meldes til prosjektet vil det vurderes i hvert enkelt tilfelle om de skal ha tilgang til alle variablene eller kun deler av datamaterialet. Variabelen grunnkrets vil være forbeholdt forskerne i prosjektet, og det vil derfor ikke være aktuelt å gi denne ut i datasett som skal benyttes av masterstudenter som deltar i prosjektet.

1.6.1 Krav fra SSB om oppdatert oversikt over prosjektmedarbeidere

Universitetet i Oslo skal ha en oppdatert oversikt over personer som har tilgang til datasettet. Oversikten er lagt til som et eget vedlegg i meldeskjemaet, og er derfor tilgjengelig for Universitetet i Oslo via Meldingsarkivet. Ved endringer/utskiftninger skal prosjektansvarlig sørge for å oppdatere oversikten.

1.7 Varighet

Prosjektslutt er satt til 31.07.2031. Innen denne dato skal data slettes, etter avtale med SSB.

2 Vurdering av om behandlingsaktivitetene er nødvendige og står i rimelig forhold til formålene

2.1 Rettslig grunnlag

Utdanningssystemet har som mål å sikre at alle barn får like utdanningsmuligheter, men likevel finnes store utdanningsforskjeller på tvers av kjønn, etnisitet, sosioøkonomisk bakgrunn og/eller kontekstuelle kjennetegn som nabolag og skolekvalitet. Prosjektet har som mål å skille de kausale mekanismene bak disse forskjellene fra hverandre ved å se individuelle kjennetegn og kontekst i sammenheng med ferdighetsutviklingen i løpet av utdanningsløpet og senere i livet. Særlig fokus vil rettes mot hvorvidt norske utdanningsreformer har påvirket utdanningsulikhet. Prosjektet har som mål å kunne skille endringer i utdanningsulikhet fra endringer i ulikhet langs andre samfunnsmessige dimensjoner. Prosjektet vil dessuten undersøke hvor rettferdige utdanningsmål er, ved å undersøke hvordan partiskhet (bias) og andre måleeffekter påvirker utdanningsulikhet.

NSD vurderer at prosjektet vil ha høy samfunnsnytte, og at behandlingene av personopplysninger er nødvendige for å utføre en oppgave i allmennhetens interesse og for formål knyttet til vitenskapelig forskning.

Lovlig grunnlag for behandlingen av alminnelige personopplysninger er dermed at den er nødvendig for å utføre en oppgave i allmennhetens interesse, jf. personvernforordningen art. 6 nr. 1 bokstav e, samt for formål knyttet til vitenskapelig forskning, jf. personopplysningsloven § 8, jf. personvernforordningen art. 6 nr. 3 bokstav b.

Lovlig grunnlag for behandlingen av særlige kategorier av personopplysninger er at den er nødvendig for formål knyttet til vitenskapelig forskning, jf. personvernforordningen art. 9 nr. 2 bokstav j, jf. personopplysningsloven § 9.

Behandlingen er omfattet av nødvendige garantier for å sikre den registrertes rettigheter og friheter, jf. personvernforordningen art. 89 nr. 1.

Det skal foreligge dispensasjon fra taushetsplikten fra Udir og UNIT for utlevering av opplysninger fra deres registre, jf. forvaltningsloven § 13 bokstav d.

2.2 Sentrale prinsipper

2.2.1 Formålsbegrensning

Prosjektets formål er å skille kausale mekanismer bak utdanningsforskjeller fra hverandre, å skille endringer i utdanningsulikhet fra endringer i ulikhet langs andre samfunnsmessige dimensjoner, og å undersøke hvor rettferdige utdanningsmål er. NSD vurderer at formålet er klart definert, spesifikt, uttrykkelig angitt og fremstår som rimelig for en forskningsinstitusjon.

2.2.2 Dataminimering

Behovet for de enkelte variablene og nødvendigheten av disse for å gjennomføre de planlagte analysene er godt gjort rede for. Forskerne vil få tilgang til en rekke opplysninger, som sammenstilt kan være egnet til å si noe om enkeltpersoner. NSD vurderer at nødvendigheten av å innhente opplysninger om kommune og grunnkrets er godt begrunnet. Opplysninger om inntekt vil avrundes. Forskerne vil be om at skolenavn, skolenummer og spesifikke datoer utelates fra data om **søknad og opptak til videregående utdanning**. Forskerne vil også be om at alle direkte identifiserende opplysninger i GIS-dataene utelates før de oversendes SSB for kobling. Forskerne vil få utlevert samtlige opplysninger i datasettet som UNIT sitter på, med unntak av direkte identifiserende opplysninger. Dette fordi det er avgjørende for prosjektet å sitte med samme analysegrunnlag som det Samordna opptak trenger for å vurdere om hver søker er kvalifisert for studiet de søker på, og som gir grunnlag for å gjøre studieopptak.

NSD vurderer derfor at personopplysningene som skal behandles er adekvate, relevante, nødvendige og begrenset til det som er nødvendig for formålet.

2.2.3 Riktighet

Dataene innhentes fra SSB, UNIT og Udir, og skal kobles av SSB. Det er derfor liten grunn til å tro at opplysningene skal være uriktige.

2.2.4 Lagringsbegrensning

Prosjektet har en varighet på ti år. Prosjektet er omfattende og NSD vurderer at varigheten av behandlingen av personopplysninger står i et rimelig forhold til formålet.

2.2.5 Integritet og konfidensialitet (personopplysningssikkerhet)

Data skal lastes opp i TSD fra SSB, og kun behandles på TSD. Dette ivaretar grunnleggende krav til informasjonssikkerhet, som tilgangsstyring, logging og etterfølgende kontroll. I tillegg vil en oppdatert oversikt over hvem som skal ha tilgang til opplysningene oppbevares ved UiO og også være tilgjengelig for institusjonen via NSD sitt Meldingsarkiv. NSD vurderer at de tekniske og organisatoriske tiltakene beskrevet i del 1 gir tilstrekkelig vern mot uautorisert/ulovlig behandling av personopplysninger samt utilsiktet tap/ødeleggelse/skade av personopplysninger.

2.3 De registrertes rettigheter og friheter

2.3.1 Rett til informasjon

Utvalgets størrelse tilsier at det vil kreve en uforholdsmessig stor innsats å informere de registrerte sett opp mot nytten den enkelte vil ha av å få informasjon. Basert på en avveining mellom tiltakene som kreves for å informere og ulempen for den enkelte registrerte, vurderer NSD at det kan unntas fra informasjonsplikten på grunnlag av at det vil gjøre gjennomføringen av prosjektet umulig eller i alvorlig grad vil hindre oppfyllelsen av de spesifikke formålene, jf. personvernforordningen art. 14 nr. 5 bokstav b.

Prosjektet vil på egen nettside informere om formål og om hvordan registrerte kan utøve sine rettigheter. Informasjonen på nettsiden skal oppfylle kravene i personvernforordningen art. 14.

2.3.2 Rett til innsyn, retting, sletting, behandlingsbegrensning, protest, dataportabilitet

I den grad den registrerte tar kontakt med prosjektet og kan identifiseres i datamaterialet vil den registrerte ha rett til innsyn, retting, sletting og protest.

3 Vurdering av risiko for de registrertes rettigheter og friheter

NSD vil trekke frem følgende risikoer i prosjektet:

- Det behandles personopplysninger i stor skala.
- Personopplysningene behandles uten at det gis informasjon eller innhentes samtykke fra de registrerte. Dette utfordrer deres mulighet til ha reell bestemmelse over sine personopplysninger og det utfordrer prinsippet om åpenhet.

4 Planlagte tiltak for å håndtere risikoene

Følgende tiltak er igangsatt for å håndtere de identifiserte risikoene:

- Datamaterialet skal behandles sikkert på TSD.

- Særlig identifiserende variabler, som grunnkrets, vil kun være tilgjengelige for forskerne i prosjektet. Det skal vurderes i hvert enkelt tilfelle om nye medarbeidere skal ha tilgang til alle variablene eller kun deler av datamaterialet. Variabelen grunnkrets vil være forbeholdt forskerne i prosjektet, og ikke gjøres tilgjengelig for masterstudenter. Variabelen inntekt skal avrundes.
- Databehovet er godt begrunnet.
- Det skal foreligge dispensasjon fra taushetsplikten fra UNIT og Udir.

5 NSDs samlede vurdering av personvernet

NSD vurderer på grunnlag av ovennevnte tiltak at prosjektet håndterer de identifiserte risikoene på en akseptabel måte, og at personvernet således er tilstrekkelig ivaretatt.

Vi legger spesielt vekt på at det er lagt opp til god informasjonssikkerhet og databehovet er svært godt begrunnet. Samfunnsnytten ved prosjektet er høy.

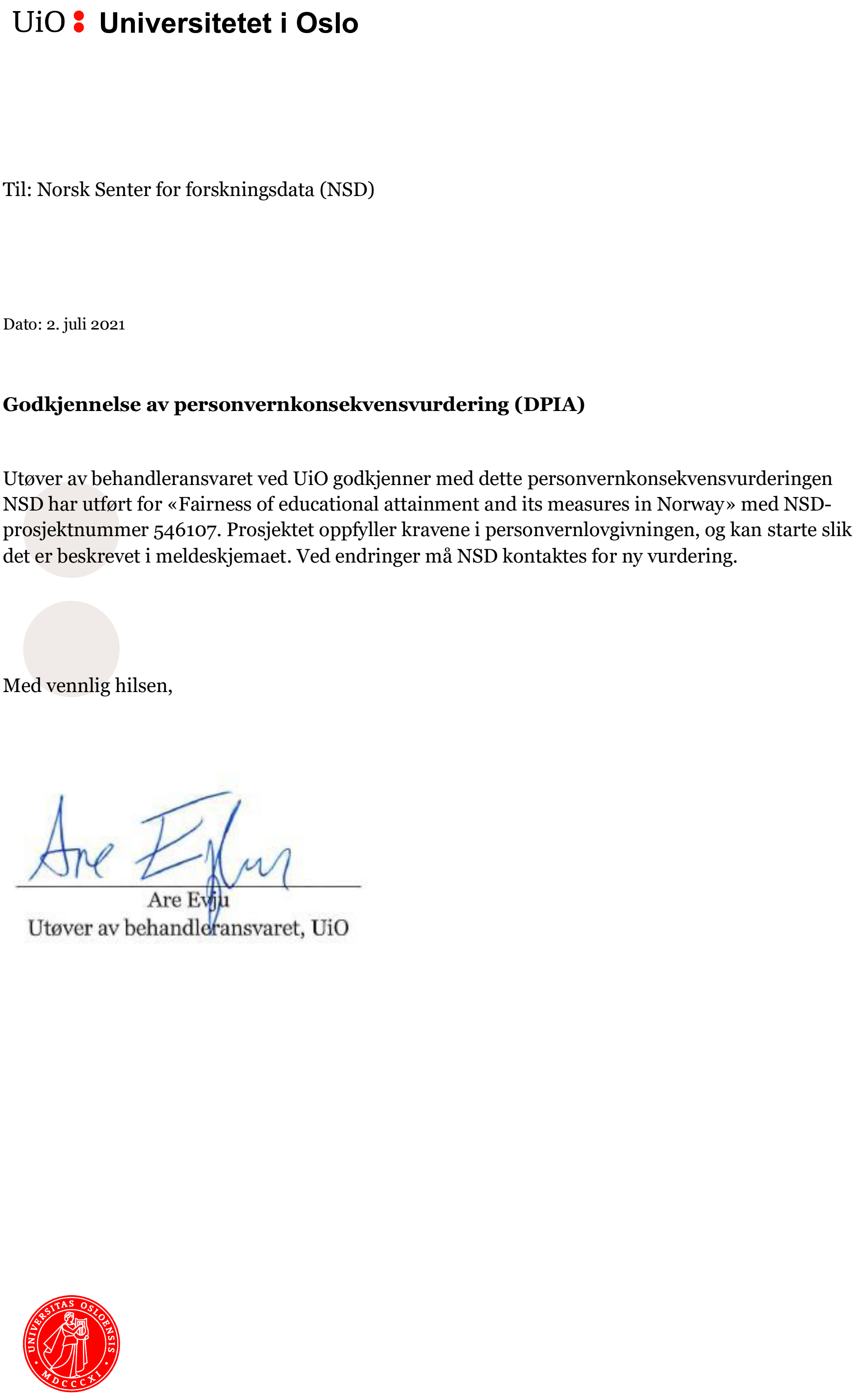
Vi legger til grunn at søknadene om dispensasjon fra taushetsplikten innvilges og at forsker etterfølger eventuelle vilkår. NSD ber om at vedtakene lastes opp i meldeskjema.

6 Godkjenning fra institusjonens ledelse

Utøver av behandleransvaret har gjennomgått DPIAen, og stiller seg bak NSDs vurdering. Personvernombudet har kontrollert gjennomføringen av personvernkonsekvensvurderinger ved UiO det og syns gjennomføringen er betryggende

Se vedlagt signert godkjennelse

B. Signed Approval of the Data Protection Impact Assessment



UiO : Universitetet i Oslo

Til: Norsk Senter for forskningsdata (NSD)

Dato: 2. juli 2021

Godkjennelse av personvernkonsekvensvurdering (DPIA)

Utøver av behandleransvaret ved UiO godkjenner med dette personvernkonsekvensvurderingen NSD har utført for «Fairness of educational attainment and its measures in Norway» med NSD-prosjektnummer 546107. Prosjektet oppfyller kravene i personvernlovgivningen, og kan starte slik det er beskrevet i meldeskjemaet. Ved endringer må NSD kontaktes for ny vurdering.

Med vennlig hilsen,

Are Evju

Are Evju
Utøver av behandleransvaret, UiO

UNIVERSITAS OSLOENSIS
MDCCCXI

Appendix II – Coding

The code used for data cleaning, analysis, and presentation of data can be found in the following GitHub repository: https://github.com/SverdoSverdo/Thesis-. The sections of the code are enumerated, and descriptions and further comments regarding each section is found in the table below. As the data is sensitive, it is not made available.

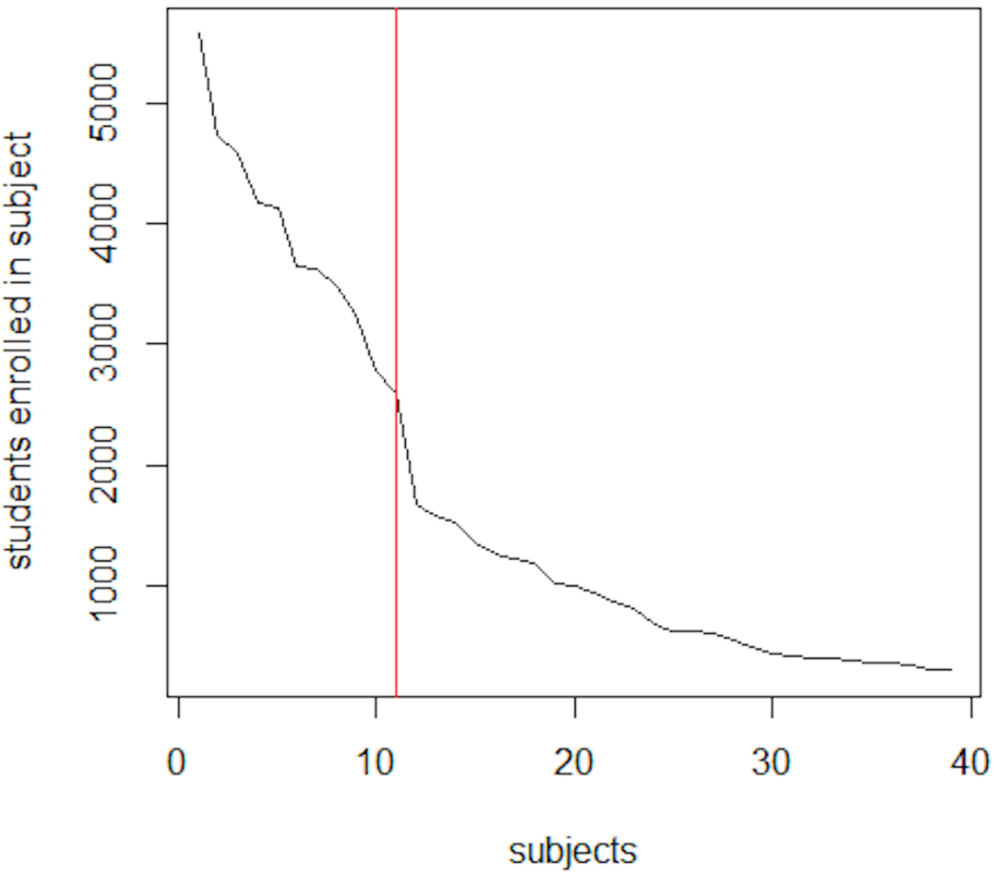
Section name	Description and comments
Loading data set and packages	Loading packages and the dataset containing grades and school and student ID.
1. Preparation grade dataset	This section cleans the dataset containing the grades of students and identifies the sample used for the study. We could not identify how many electives a student was enrolled in by looking at how many subjects the students took in addition to mandatory subjects because some students had to retake mandatory subjects from previous years. Language electives were identified differently from non-language electives because an exhaustive list of language electives available to third year students is not found at www.vilbli.no. Even though we had access to the program a student was enrolled in through the v_data file, this was only used to make sure that no one in the final sample was not in the specialization for general studies program. This is because v_data only contains information about those who have graduated, something which would have led to a lower sample size, and potentially a skewed proficiency distribution as it would have excluded students who did not graduate.
2. Preparing choice variable, d_{ni}	This section prepares the choice variable d_{ni} by using information about whether the student enrolled in a subject and whether the student's school offered the subject.
3. Preparing final df	This section aligns and combines the choice and grade datasets, and recodes letter grades
4. Descriptive Statistics	This section prepares descriptive statistics with observed grades

5. Models	This section specifies and estimates Models 0-4 as well as the selection model by itself.
6. Expected grades	This section calculates expected grades for Models 2 and 4. The computeExpected function calculates expected grades for missing values, while computeExpectedAll calculates grades for all cells.
7. Comparing Models 2 and 4	This section calculates the item fit statistic that compares grades from computeExpectedAll for Models 2 and 4 with observed grades.
8. Model Tables	Creates different tables used throughout the study
9. Preparing plots	Converts from slope/intercept parameteriztion to standardized parameterization for Models 0-4. The deltamethod for this conversion is not available for Models 4 through the mirt package, and had to be done manually.
10. Plotting	Creates different plots used through the study.

Appendix III – Supplementary Material

A. Choice of Elective to Include

The following figure was used to decide on how many electives to include. The number of students in each subject was summed up, and subjects in the figure were ordered by number of students enrolled in them. To increase readability, we chose to only include 40 subjects in the plot. The red line going through the 11th most popular elective was used as the cut-off as there was a considerable drop-off in students taking this subject compared to the 12th most popular subject (from 2,562 to 1,644).



B. Full name of included Subjects

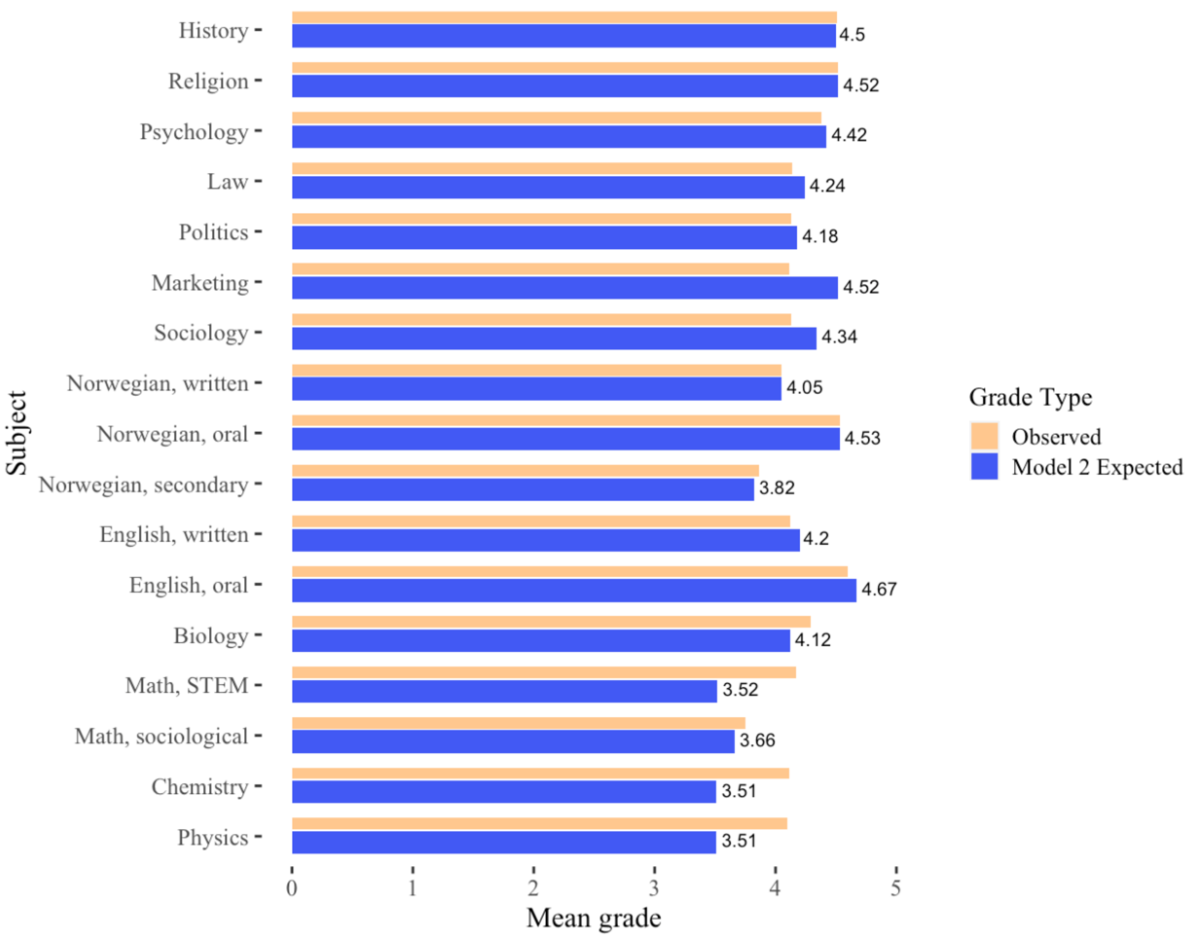
Name used in the study	Full name
History	History
Religion	Religion and ethics
Psychology	Psychology 2
Law	Law 2
Politics	Politics and human rights
Marketing	Marketing and management 2
Sociology	Sociology
Norwegian, written	Norwegian, first-choice form, written
Norwegian, oral	Norwegian, oral
Norwegian, second-choice	Norwegian, second-choice form of Norwegian, written
English, written	Social Studies English, written
English, oral	Social Studies English, oral
Biology	Biology 2
Mathematics, STEM	Mathematics R2
Mathematics, sociological	Mathematics S2
Chemistry	Chemistry 2
Physics	Physics 2

C. Detailed Description of Letter Grades

Letter grade	N (%)	Coded as	Explanation
Participated	2894 (1.97)	NA	Grade given to language minorities for Norwegian and English subjects. Is treated as a passing grade.
Approved	116 (0.08)	NA	Those with previous work experience that attend high school can receive this grade if they can demonstrate that the competencies, they have are equal to the one described in the subject’s competency goals. This is treated as a passing grade.
Exempted from grading	427 (0.29)	NA	Grade given to those that are exempted from being graded in Norwegian second-choice form. Is treated as a passing grade.
No basis for assessment	2914 (1.98)	1	Significant absences or other special reasons may result in the teacher not having a sufficient basis for providing a midterm assessment with a grade or final grade. This is treated as a failing grade.

Note. % refers to the percentage of the corresponding grade in relation to all grades in the sample. Source: Udir (2023b).

D. Model-Expected Grades under Model 2



Note. STEM = science, technology, engineering, math.