A quantitative enquiry into the fairness and equity in Norwegian secondary school assessment practices: Project proposal

Tony C. A. Tan

Centre for Educational Measurement

University of Oslo

UV9040A Research Seminar

Associate Prof Björn Andersson

14 September 2021

Introduction

Based on *Statistisk sentralbyrå* archive (SSB, 2021), 34,660 young Norwegians between 2014 and 2020 entrusted their future to the grade point average (GPA) system upon completion of their secondary schools (general studies). Although the public *assumes* axiomatically the GPA-based selection system to be fair, little theoretical and empirical efforts had been devoted to the detailed examination of such claim in the Norwegian context. The fairness of this trust-based selection system is further called into question by the increasing dissenting voices overseas. By analysing UK's 2004 General Certificate of Secondary Education data, for instance, Coe (2008) revealed significant differences in subject difficulties, with some being more than one grade harder than others. Similarly in the Netherlands, Korobko et al. (2008) reported comparable heterogeneity in subject difficulties using the Central Examinations in Secondary Education data. Cross country comparisons by Lamprianou (2009) also revealed significant differences in assessment practices for the purpose of enhancing test fairness. When test difficulties vary across subjects, winners and losers are created; delays in re-aligning subject difficulties during GPA calculations would erode public trust in the impartiality, therefore the legitimacy, of university entry selection process.

Enquiries into subject difficulty re-calibration have a long history. Tognolini and Andrich (1996) used extended logistic models to measure the latent trait of test candidates. In addition to their detailed mathematical derivations, theses authors made extensive justification for the viability of summarising learners' diverse capabilities into a single score, paving the way for subsequent research efforts using Rasch modelling. Parallel to empirical studies, advancement in estimation methods also enabled the proliferation of Rasch modellings. Three main approaches to parameter estimation have withstood the test of time: marginal maximum likelihood estimation (MML, Bock & Aitkin, 1981) where person parameters are treated as random effects and are integrated out of the likelihood function, joint maximum likelihood estimation (JML, Birnbaum, 1968; Lord, 1980; Mislevy & Stocking,

1987) where person parameters are treated as fixed effects and are retained in the likelihood function, and conditional maximum likelihood estimation (CML, Andersen, 1972), which takes advantage of the separability of the person and subject parameters in a Rasch model to condition the likelihood function on its sufficient statistics. Recent studies suggested CML over its competitors thanks to its computational efficiency (Christensen, 2013) and robustness to normality violation (Steinfeld & Robitzsch, 2021).

The current study responses to the under-research in Norway's GPA architecture. It is particularly interested in knowing whether the GPA in its present form is a "fair" measure of candidates' underlying capabilities. More specifically, fairness is conceptualised as the comparability in difficulty levels across subjects such that no candidate shall be favoured or penalised for their subject choices. Should significant differences in subject difficulties be found, research interests would be devoted to the spatial-temporal distribution (Which subjects are the easiest/hardest? Do these differences occur only in one particular year or in every year? Are these differences getting bigger, smaller, stable or fluctuate over the years?), causes (Why?), personal impact (Who are mostly (dis-)advantaged by such differences?) as well as policy actions (What can be done to address such unfairness?). On the contrary, should this study find no significant difference, it would also serve Norway's interest well by lending support and legitimacy to its current university entry selection mechanism. Results from this project directly benefit students and their parents through enhanced transparency and accountability of a system that determines their future and dreams. Educators and decision makers may also derive utility from this study for their evidence-based policy design.

With the aforementioned motivations, this study wishes to directly address the following research questions:

RQ1 To what extent can a Rasch model approximate Norway's GPA scores?

RQ2 Does the difficulty parameter differ significantly across subject?

RQ3 If the answer to RQ2 is "yes", then further RQs would involve:

RQ3.1 Which are the easiest and hardest subjects?

RQ3.2 Are there any systematic differences by socio-demographic variables such as sex, immigration history or socio-economic spectrum?

RQ3.3 Do such differences remain stable over time?

Methods

Sample

For this study, students' GPA records will be extracted from the Norwegian registry covering the period between 2006 (the year of education reform) and 2020 (most recent year with available data). GDPR registration is lodged through the NSD Portal and the UiO ethics approval is also obtained. All data import, storage, and analyses are to be conducted within the secured infrastructure TSD provided by the UiO Central IT Division. TSD logs all activities and no data or results can be copied out of the restricted system without prior approval from project leaders.

Under the advisory of He and Stockford (2015), subjects with fewer than 1,000 candidates and students taking fewer than two GPA subjects will be excluded from subsequent analyses. Each year's record (score matrix) will contain N rows representing the number of valid candidates and L columns reflecting the usable number of GPA subjects in that year. Since no student took all the GPA subjects, a large proportion of the score matrices will remain missing by design. The existence of missing data does not pose any problems for using the Rasch model as the model functions at the individual subject and as long as there is sufficient overlap across subjects in the score matrix. The ability to deal with incomplete data is one major advantage of using the Rasch model for studying inter-subject comparability.

Rasch Model

The Rasch model was developed in the 1960s for establishing measurement scales and for improving test development (Rasch, 1980). In a simple Rasch model, the underlying ability or latent trait of the person (θ) and the item characteristics (δ_j) are specified; a logistic function (Λ) is then used to describe the probability that the person will successfully pass a subject ($x_j = 1$) given their ability θ and the item characteristics δ_j (de Ayala, 2009):

$$\mathbb{P}\left(x_j = 1|\theta, \delta_j\right) = \Lambda(\theta - \delta_j) = \frac{1}{1 + e^{-(\theta - \delta_j)}}. \tag{1}$$

Equation (1) is suitable for modelling subjects with dichotomous (pass/fail) outcomes. GPAs on the other hand are polytomous in nature (§3-5, *Forskrift til opplæringslova*), therefore requires models capable of accommodating more than two achievement outcomes. Multiple extensions have been put forward over the decades such as rating scale models (Rasch, 1980), partial credit models (Masters, 1982), and generalised partial credit models (Muraki, 1992). Master's (1982) partial credit models (PCM) are particularly parsimonious and flexible for studying the structure of GPA data. A PCM states that, for a subject with $m + 1$ available grades ($m = 5$ for Norway's GPA system), the probability of a candidate with ability θ receiving grade x can be expressed as:

$$\mathbb{P}\left(\theta, x\right) = \begin{cases} \dfrac{1}{1 + \sum_{l=1}^{m} \exp\left\{\sum_{k=1}^{l}(\theta - \delta_k)\right\}} & \text{for } x_j = 0 \\ \dfrac{\exp\left\{\sum_{k=1}^{x}(\theta - \delta_k)\right\}}{1 + \sum_{l=1}^{m} \exp\left\{\sum_{k=1}^{l}(\theta - \delta_k)\right\}} & \text{for } x_j = 1, 2, \ldots, m \end{cases} \tag{2}$$

where δ_k is the location of the k-th step on the latent trait continuum, often referred to as the item step parameter associated with a grade category, step difficulty or threshold. $\mathbb{P}\left(\theta, x\right)$ is commonly known as the category response function or the item category probability curve (CPC). Model parameters in Equation (2) can be solved using conditional maximum

likelihood estimation (Andersen, 1972).

Difficulty Parameters

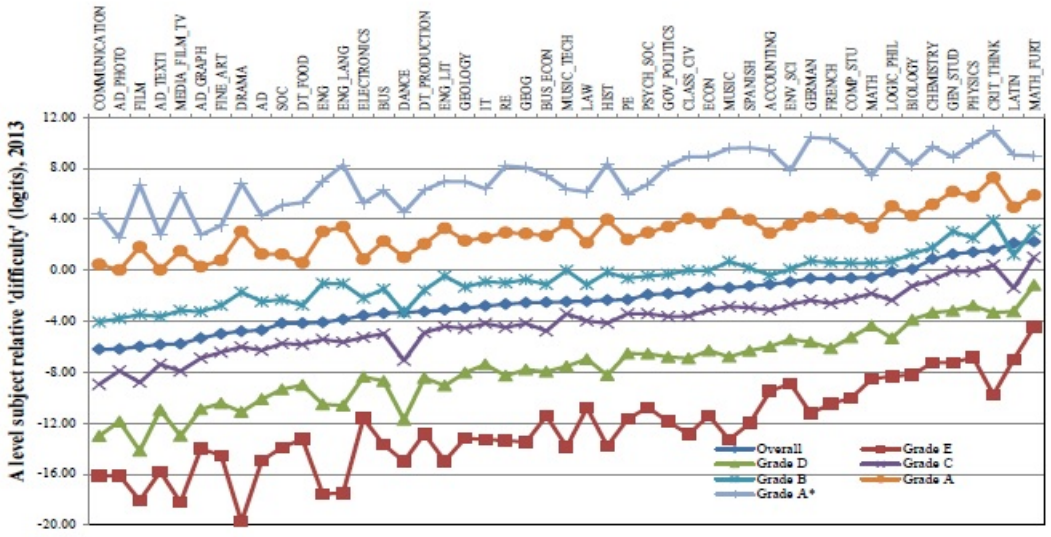
It is important to highlight that δ_k cannot be interpreted as the difficulty of scoring k in a subject. Wu and Adams (2007) proposed measures of subject difficulty based on expected scores by first defining the item characteristic curve (ICC): $\mathbb{E}(\theta) = \sum_{x=0}^{m} x\mathbb{P}(\theta, x)$, then the difficulty of scoring k (d_k) as the ability at which the expected score on the ICC is $k - 0.5$:

$$d_k = \theta|_{\mathbb{E}(\theta)=k-0.5}. \tag{3}$$

The overall difficulty of a subject (D) can be obtained by averaging all step parameters:

$$D = \frac{1}{m}\sum_{k=1}^{m}\delta_k. \tag{4}$$

This study aims to ascertain the grade difficulties (d_k) as well as the overall difficulty of each subject, similar to Figure 13 of He and Stockford (2015):



Data Analyses

This study will make extensive use of STATA 17's IRT module. Based on recommendations from recent simulation studies (Christensen, 2013; Steinfeld & Robitzsch,

2021), CML will be used as the primary estimation procedure. Missing data require no additional treatment since one major advantage of IRT is its ability to handle dataset containing "planned missingness". Analyses will be repeated for each year, with the results forming a pooled cross sectional data for temporal comparisons (subject $\times$ grade $\times$ time).

References

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(1), 42–54. https://doi.org/10.1111/j.2517-6161.1972.tb00887.x

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–460). Addison-Sesley.

Bock, R. D., & 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

Christensen, K. B. (2013). Conditional maximum likelihood estimation in polytomous Rasch models using SAS. *ISRN Computational Mathematics*, *2013*(617475), 1–8. https://doi.org/10.1155/2013/617475

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

He, Q., & Stockford, I. (2015). *Inter-subject comparability of exam standards in GCSE and A Level* (ISC Working Paper 3). Office of Qualifications and Examinations Regulation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606044/3-inter-subject-comparability-of-exam-standards-in-gcse-and-a-level.pdf

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

Lamprianou, I. (2009). Comparability of examination standards between subjects: An international perspective. *Oxford Review of Education*, *35*(2), 205–226. https://doi.org/10.1080/03054980802649360

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mislevy, R. J., & Stocking, M. L. (1987). A consumer's guide to LOGIST and BILOG. *ETS Research Report Series*, *1987*(2), 1–73. https://doi.org/10.1002/j.2330-8516.1987.tb00247.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

SSB. (2021). *Completion rates of pupils in upper secondary education*. Statistisk sentralbyrå [Statistics Norway]. https://www.ssb.no/en/utdanning/videregaende-utdanning/statistikk/gjennomforing-i-videregaende-opplaering

Steinfeld, J., & Robitzsch, A. (2021). Item parameter estimation in multistage designs: A comparison of different estimation approaches for the Rasch model. *Psych*, *2021*(3), 279–307. https://doi.org/10.3390/psych3030022

Tognolini, J., & Andrich, D. (1996). Analysis of profiles of students applying for entrance to universities. *Applied Measurement in Education*, *9*(4), 323–353. https://doi.org/10.1207/s15324818ame0904_3

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical aproach*. Educational Measurement Solutions.

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-content/15-irt/2007-Wu-Apply-Rasch-Measurement.pdf