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

Attachment 3 to UV PhD application form

Prof Rolf V. Olsen & Dr Astrid M. J. Sandsør

27 August 2021

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 a 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: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.

Purpose Statement

Research Questions

Methods

Sample

For this study, students' GPA records will be captured from the Norwegian registry covering the decade between 2010 and 2020. 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 item and items 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 joint maximum likelihood estimation (Molenaar, 1995) or the conditional maximum likelihood estimation (Andersen, 1972).

Difficulty Parameters

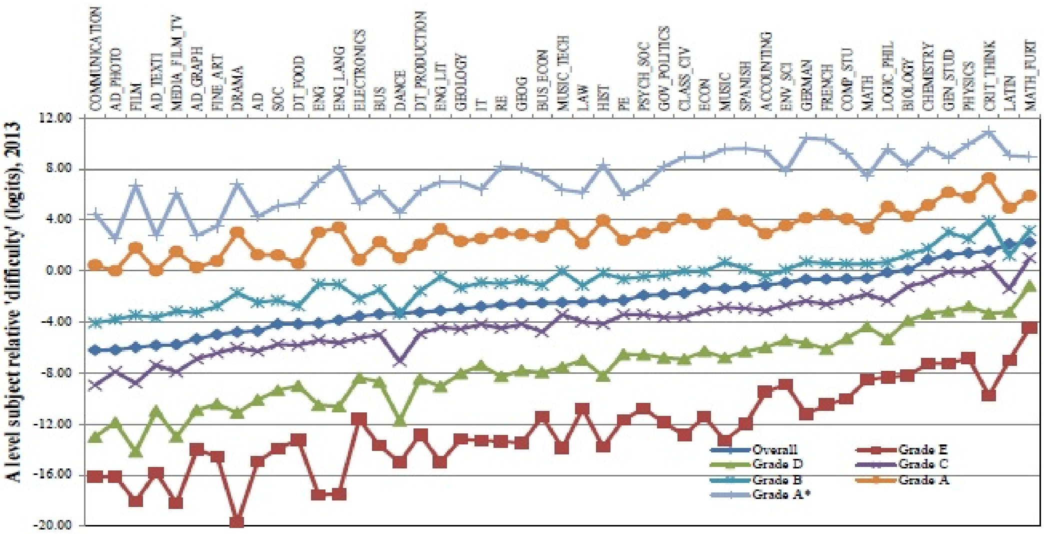
It is important to highlight that δ_k cannot be interpreted as the difficulty of scoring k in a subject. Wu and Adams (2007) proposed measures of subject difficulty based on expected scores by first defining the item characteristic curve (ICC): $\mathbb{E}\left(\theta\right) = \sum_{x=0}^{m} x\mathbb{P}\left(\theta, x\right)$, 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 heavy use of STATA 17's IRT module. Based on recent

Hypothesised Results

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.

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

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

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

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 39–51). Springer. https://doi.org/10.1007/978-1-4612-4230-7_3

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

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