# Identifying Inter-subject Difficulties in Norwegian GPA Data Using Item Response Theory

Tony C. A. Tan

Centre for Educational Measurement, University of Oslo

Continuous Draft

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

Vår 2022

## Abstract

### Research Topic

The Grade Point Average (GPA, *skolepoeng* in Norwegian) plays a determining role in Norway's tertiary admission process. The academic track in Norwegian upper secondary education offers students a set of compulsory joint core subjects as well as a wide range of elective subjects for different specialisations in, for instance, sciences or language arts. Each subject awards students a grade ranging from 0 to 6 for low- and high-competence respectively (Education Act Regulation, 2006, §3-5). Students' GPAs are best characterised as sum scores of their subject grades. For the majority of tertiary studies, different elective subjects are treated *equally* in GPA calculations. Under this practice, GPA implicitly assumes that grades across different specialised subjects are *equivalent* indicators of students' preparedness for higher education, an assumption that remains untested. Descriptive statistics suggest that there are substantial differences in grades across subjects (Norwegian Ministry of Education, 2022). The current study is part of a larger project examining Norwegian administrative grade data using item response theory (IRT). Specifically, this paper focuses on the comparability of difficulty levels across subjects, and thus provides a test of the hidden assumption in the current procedure for producing the GPA. Examining *whether Norway's GPA subjects differ in difficulty levels* serves the dual-purpose of enhancing selection fairness (Camilli, 2006) and ensuring GPA's appropriateness as an educational measurement device (AERA et al., 2014).

### Theoretical Framework

Fairness is both an essential and an elusive integral of educational assessment. Both Gipps and Stobart's (2009) social-cultural framing of assessment fairness and Tierny's (2017) democratic–measurement–pedagogical construction acknowledged the prerequisite role statistical analyses must play for understanding the fairness issues in assessment. In fulfilling this foundational support function, the current study models GPA as a *selection* device (IUA, Kane, 2013) for accessing privileged social resources (Bourdieu, 1973) and addresses the construct validity of GPA by examining any construct-irrelevant variance (Messick, 1989). Resultantly, grading is thought to be a decontextualised measurement procedure (Kalthoff, 2013) with interchangeable instruments yielding identical results.

## Methodology

IRT is particularly suitable for extracting item difficulty information in order to study assessment's selection fairness. This study considers each GPA subject as an item and each candidate as a person. Using marginal maximum likelihood (MML) estimation, the analyses will ascertain difficulty parameters for all major subjects in Norwegian upper secondary schools. A second methodological consideration relates to self-selection bias. Freedom in subject choices in Norway's upper secondary academic track inevitably produces rather sparse data matrix once all subjects and students are included. Since the presence or absence of observations was not resulted from randomisation but self-selection, and the missing likelihood is reasonably expected to covary with the subject difficulties, and the observed GPA datasets shall be considered missing not at random (MNAR, Rubin, 1976). Leaving untreated, such non-ignorable missingness would cause over- and under-estimates of person and item parameters, respectively (Rose, 2013). In order to assess the impact of non-random missings on difficulty parameter estimates, IRT analyses will be repeated on three groups: the whole population, medical school applicants (low subject choice freedom) and language arts stream students (high freedom).

Registry data containing Norwegian students' GPA performance in 2019 are first regularised by removing subjects with fewer than 1,000 candidates (as practised by He et al., 2018) and by only including candidates who have received valid GPAs through upper secondary school completions. Next, subject difficulty parameters will be extracted using generalised partial credit models (GPCM, Muraki, 1992). Lastly, the sensitivity analysis section will contain group invariance tests to assess the extent to which selection bias had impacted on subject difficulty parameter estimates.

## Expected Results

The registry data set will be available for analysis in short time and the described analyses will be presented and discussed at the conference. Given that university entries in Europe is largely based on the final grades from secondary education, Norway's GPA system is expected to be comparable to the A Levels in the UK and the Central Examinations in Secondary Education in the Netherlands. More specifically, we expect Norway's GPA subjects to differ in difficulties (per report by He et al., 2018) and to exhibit significant selection effect (as demonstrated in Korobko et al., 2008) represented by different difficulty parameters among

the whole sample, medical school applicants, and language arts candidates.

**Relevance to Nordic Educational Research**

All Nordic countries have merit-based criteria for selection into tertiary education, although such criteria are operationalised differently across our countries. The issue of potential unequal treatment of students with different specialisation in upper secondary school applies across our countries. By testing the assumption that grades from different specialty streams support GPA's selection purpose equally well, this study lends statistical output to evidence-based policy formation process commonly practised in the Nordic community and serves to strengthen the fairness of our merit-based university admission decisions.

## References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bourdieu, P. (1973). Cultural reproduction and social reproduction. In R. Brown (Ed.), *Knowledge, education, and cultural change: Papers in the sociology of education* (pp. 71–112). Tavistock Publications. https://doi.org/10.4324/9781351018142-3

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational mearement* (4th, pp. 221–256). American Council on Education; Praeger Publishers. https://www.researchgate.net/profile/Gregory-Camilli/publication/265086461_Test_fairness/links/578e4ae908ae81b4466ec0f8/Test-fairness.pdf

Education Act Regulation. (2006). *Forskrift til opplæringslova* [FOR-2006-06-23-724]. Lovdata. https://lovdata.no/forskrift/2006-06-23-724/%C2%A73-5

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st Century: Connecting theory and practice* (pp. 105–118). Springer. https://doi.org/10.1007/978-1-4020-9964-9_6

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

Kalthoff, H. (2013). Practices of grading: An ethnographic study of educational assessment. *Ethnography and Education*, *8*(1), 89–104. https://doi.org/10.1080/17457823.2013.766436

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 10–103). American Councile on Education; Macmillan.

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

Norwegian Ministry of Education. (2022). *Karakterstatistikk for videregående skole* [Grade statistics for upper secondary school]. Utdanningsdirektoratet. https://www.udir.no/tall-og-forskning/statistikk/statistikk-videregaende-skole/karakterer-vgs/

Rose, N. (2013). *Item nonresponses in educational and psychological measurement* [PhD Thesis, Friedrich-Schiller-Universität Jena]. Open Access Thesis and Dissertations. https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00027809/Diss/NormanRose.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Tierny, R. D. (2017). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 793–798). Springer. https://doi.org/10.1007/978-981-287-588-4_400