Abstract

Research Topic

Grade point averages (GPA) play a determining role in Norway's tertiary admission processes. Earlier studies from the UK (He et al., 2018) and the Netherlands (Korobko et al., 2008), however, raised methodological and fairness concerns over GPA as an appropriate measure for graduates' academic competency. Violations of the unidimensionality assumption arose when different subjects contribute to the final GPA scores at different weights, undermining statistical inferences under the item response theory (IRT) framework. Additionally, misaligned subject difficulties distort upper secondary academic track candidates' incentives, leading to misallocation of youth's time and effort at a critical point in their studies. This paper aims to examine *whether Norway's GPA subjects differ in difficulty levels*, both across candidate cohorts (e.g., medical school vs general tertiary applicants) and across time. It further investigates covariates that associated strongly with any discrepancies in subject difficulties for policy considerations.

Theoretical Framework

IRT is particularly suitable in the educational measurement literature for extracting item difficulty parameters. 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 theoretical consideration relates to self-selection bias. 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 courses for different specialisations such as medicine and the language arts streams. Such freedom in subject choices led to a 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, 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). This study addresses MNAR using a multiple imputation procedure prior to IRT analyses.

Methodology

Registry data containing Norwegian students' GPA performance between 2009 and 2019 are first regularised year-by-year by removing subjects with fewer than 1,000 candidate and candidates taking fewer than two subjects following the practices in He et al. (2018). Candidates' grades are then recoded into a polytomous scale with 0 and 5 representing the low- and high-ends of the competency spectrum. Multiple imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011) are then applied to the observed data matrix to obtain ten imputed versions for MML estimations involving generalised partial credit models (GPCM, Muraki, 1992) for grade- and subject-difficulty parameter extractions. Final estimates are then pooled together using Rubin's Rule (Rubin, 1987) in order to obtain correct means and standard error statistics. Identical procedures are applied to each year to obtain a pooled cross sectional output and special cohorts such as medical school applicants are highlighted for sensitivity analyses.

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). We further expect subject difficulty parameters to increase once the missing data mechanisms had been taken into account.

Relevance to Nordic Educational Research

Researchers in Nordic countries are privileged to have access to national registry data, a gateway to nuanced information about individual-level phenomena. Consensus on a standard procedure for analysing registry data for educational research purposes, however, are yet to emerge that safeguards methodological accuracy as well as promotes social welfare at large. This study show cases a modular analysis design by explicitly addressing non-ignorable missing data issues before submitting the datasets to IRT modelling. Establishing and verifying the analytical procedures and properties of resultant estimates would directly benefit Nordic research communities using registry data.

References

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

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

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

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

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. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

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