An extended abstract on grading fairness in Norwegian school system

Tony C. A. Tan
Centre for Educational Measurement
University of Oslo

CEMO PhD Application (205778)
Prof Rolf V. Olsen
14 June 2021

Overview of Research Topic

Grading fairness plays a key role in upholding public's confidence in any assessment process. Although consensus remains high for the ideal of test fairness, empirical evidence suggested substantial variations in grading practices in Norwegian secondary schools (Tveit, 2014). Attempts to standardise assessments were received by intense politics; decision makers, in response, demanded urgent research on theory- and evidence-based assessment framework.

Educational assessment has a unique history in Norway. Formal marking in primary schools was brought to an end by an act of Parliament in 1973 driven by, among others, concerns about the negative impact on low achievers (NOU (Green Paper) (1974), as cited in Tveit (2014)). It was not until the "PISA shock" in the early 2000s that social debate resumed over the applicability and best practice of measuring the quantity and quality of educational outcomes (Lie et al. (2001) as cited in Tveit (2014)). Curriculum reforms soon followed amidst the 2005 election cycle, with the new government admitting problems ranging from poor understanding of regulations to inadequate assessment literacy shared by teachers and teacher educators (Stortingsmelding nr. 16, 2006–2007, as cited in Tveit (2014)). Despite the intervening decade, a recent study found that the clarification of the purposes of exams remained insufficient, with adverse consequences of inappropriate use of examination data by various stakeholders (Tveit & Olsen, 2018). It is therefore imperative to continue educational research into the art and science of assessments, with fairness among the chief concerns.

Key Concepts

Grading fairness can be understood along three axes: fairness across modes of administration, fairness across subjects, and fairness across schools.

Is *trekkordningen* a truly randomised process?

Since Norway's high schools remain the responsibility of local municipalities, national tests are conducted using a *trekkordningen* system whose participants were drawn from a random process. External examiners then mark the national samples while teachers evaluate the remaining students' educational outcomes. If *trekkordningen* were truly random, one should expect no systematic differences between the national and local performance. In practice, there were sufficient concerns over the logistics and even incentives behind such random sampling procedure (Olson, 2021) and empirical study by Hovdhaugen et al. (2018) reported sustained gaps between external-marked exams (*eksamen*) and teacher-marked scores (*standpunkt*) over the period between 2007 and 2011. Such differences present potential treats to fairness because achievement measures were shown to be dependent not only on learners' input but also on the modes of test administration.

Uneven practice of grading cross subjects

A second line of research addresses grading fairness between subjects. There appeared to be silent acceptance by teachers and education administrators that some

subjects (e.g., mathematics) naturally produce lower average scores than others (e.g., English) (Olson, 2021). Such differences create distortion in students' study choices such that learners self-select away from hard subjects (*realfag*), weakening Norway's long-run competitiveness in science and technological innovation. It is also insightful to investigate the shared mentality amongst teachers and school leaders that sustained the grade differences across subjects and to enquire the possible ramification to professional practices should grading across subjects be better calibrated.

Additionally, the policy of awarding top-up scores to hard sciences (*realfagspoeng*) remains under-researched. It is unclear, for example, whether such policy was introduced to counter-balance the grading penalties or as an incentive to attract more youth into STEM subjects. An archival study into Parliamentary debates may shed light on the policy intent and would help evaluate *realfagspoeng*'s contribution to promoting assessment fairness (Olson, 2021).

Unequal grading practices across schools

A third key concept relates to teacher practices. Norwegian teachers have long been crediting students' effort when awarding overall achievement marks (Dale & Wærness (2006), as cited in Tveit (2014)). Despite the explicit policy declaration in the 2006 reform requiring teachers to only consider students' achievement, later study observed that such policy was better implemented for high achieving students while low achievers' marks remained entangled with their effort and attitude (Prøitz & Spord Borgen (2010), as cited in Tveit (2014)). Mixing effort and academic achievement in *standpunkt* may partially explain its divergence from *eksamen*, undermining assessment fairness due to lack of construct agreement (Olson, 2021). As the migration from criterion- to norm-based assessment remains incomplete, students from schools with lower average grades would be expected to receive upward biases in *standpunkt*. A multilevel modelling approach would be most suited for verifying such effect.

Potential Research Questions

Trekkordningen

RQ1.1 Whether there exists statistically significant differences between external-marked exams and teacher-marked assessments.

RQ1.2 Whether such differences, if any, remained stable over time.

RQ1.3 Whether the sizes of examiner-teacher differences remain comparable between subjects such as mathematics and Norwegian language.

RQ1.4 Which school characters covary with examiner-teacher differences?

Grading practices between subjects

RQ2.1 Do grading differences (e.g., between mathematics and English) also exist in other countries?

RQ2.2 Do all subjects share the same difficulties and powers to discriminate learners' capabilities?

RQ2.3 How would IRT calibrations modify grade distributions?

RQ2.4 Do the IRT calibrations experience differential item functioning (DIF) for boys and girls?

Grading practices between schools

RQ3.1 Do students from disadvantaged schools receive upward biases in *standpunkt*?

RQ3.2 Which school characters covary most strongly with *standpunkt* biases?

Methodological Approaches

ANOVA

Analysis of variance (ANOVA) remains an effective statistical method for investigating RQ1.1 to RQ1.3. Specifically, repeated measures ANOVA (RM-ANOVA) shall be applied to RQ1.2 since the phenomena under investigation are related in time. Analysis of covariance (ANCOVA) is particularly suitable for studying RQ1.4 where covariate such as school sizes can be controlled for while examining *eksamen-standpunkt* differences.

IRT and DIF

A two-parameter logistic model (2PL) may be applied for item calibration purposes (RQ2.3), where all students' grades in all subjects can be placed into one scale representing their general ability. This approach would work particularly well in the presence of missing data (Olson, 2021) leading to two potential outcomes: (a) school subjects can be assigned to a continuum from easy to hard with approximate interval properties, and (b) the resulting IRT model would produce an alternative grade point average (GPA) estimate under a unidimensional general ability scale. DIF can be subsequently investigated to detect any gender differences in grading.

Multilevel modelling

A multilevel model may be proposed for detecting and verifying the upward biases in *standpunkt* (RQ3.2). If students from disadvantaged schools were more likely to receive lenient marking, one would expect a negative contextual effect. Recent development in multilevel modelling literature made improved corrections for both measurement errors and sampling errors as lower-level constructed being aggregated to higher-levels (e.g., the multilevel latent covariate approach by Lüdtke et al. (2008) and doubly-latent models by Marsh et al. (2009)).

References

Hovdhaugen, E., Prøitz, T. S., & Seland, I. (2018). National examination grades and final classroom grades—two of a kind? *Acta Didactica Norge*, *12*(4), Art. 0. https://doi.org/10.5617/adno.6276

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*(6), 764–802. https://doi.org/10.1080/00273170903333665

Olson, R. V. (2021). *Fairness-issues related to exams and teacher-made grade in the Norwegian context*. Personal communication, 9 June 2021.

Tveit, S. (2014). Profiles of educational assessment systems worldwide: Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 221–237. https://doi.org/10.1080/0969594x.2013.830079

Tveit, S., & Olsen, R. V. (2018). The multiple roles of national exams in the certification, governing and support of learning and instruction in Norwegian secondary education. *Acta Didactica Norge*, *12*(4), Art. 18. https://doi.org/10.5617/adno.6381