Identifying Inter-subject Difficulties in Norwegian GPA Data

Tony C. A. Tan

Centre for Educational Measurement, University of Oslo

Continuous Draft

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

Vår 2023

Identifying Inter-subject Difficulties in Norwegian GPA Data

Conceptual Framework

The Norwegian Education and Assessment System

The Norwegian education system is organised into three levels: primary school (Year 1–7) where formal grading is not practised, lower secondary school (Year 8–10) and upper secondary school (Year 11–13). During the first ten years of schooling (*grunnskole*), students follow centralised national curricula with largely compulsory subjects plus some electives. Upon successful completion of Year 10, students may choose between vocational and academic tracks for their upper secondary schools. The former is a two-year program that prepares students for employment in a specific field, whereas the latter is a three-year program (*videregående opplæring*, VG1–3) that prepares students for university studies.

The grade point average (GPA) aims to provide a sum-score measure of a student's overall competency. For *grunnskole* graduation purposes, the GPA is calculated as the unweighted average of students' grades from all Year 10 subjects. Both teacher-assigned grades and exam grades are included in the GPA calculation, with each subject ranging from 1 (low competency) to 6 (outstanding). While every compulsory subject receives a teacher-assigned grade, Year 10 students are randomly assigned into participating in *one* of the three written exams (mathematics, Norwegian, and English), as well as *one* oral exam (same as written exams, plus many electives). A candidate's GPA is then computed by averaging the grades they have obtained, multiplying by 10, and rounding to two decimal places.

The *manu*–*mente* Clusters

Not all GPA subjects target the same cognitive domain. While all subject demand students' cognitive input, some courses are undoubtedly more hand-on and practice-based such as physical education and food and health. We label the more hand-on subject in Latin as "*manu* subjects" while the cognitive-demanding ones as "*mente* subjects". [insert references suggested by Jose].

Methods

Student Population and Exam Subjects

This study retained the entire cohort of Year 10 graduates from Norway's lower secondary education (*grunnskole*) in 2019 ($N_0 = 64,918$). Students' teacher-assigned grades, written-, and oral-exam grades were extracted from the national registers. This data source is unique because it is the *population*, rather than samples, that forms our bases of analysis. Next, 4,300 students without valid GPA records were excluded from subsequent analyses [due to . . .], representing a loss rate of 6.62%.

Year 10 students in Norway should complete 13 compulsory subjects as well as electives. The compulsory subjects are: mathematics (MATH), written Norwegian *hovedmål* (main, NORW), written Norwegian *sidemål* (secondary), oral Norwegian (NORO), written English (ENGW), oral English (ENGO), natural sciences (NATS), social sciences (SOCS), religion (RELI), music (MUSI), arts and handcraft (HAND), physical education (PHED), and food and health (FOOD). This study included all compulsory subjects except the secondary written Norwegian language *sidemål* due to non-random missingness. Norwegian has two written forms and students can be exempt from the one that is not their main language based on considerations such as bilingualism [re-write the hovedmål/sidemål complexity using Astrid's text]. We also merged courses instructed in Norwegian and in Sami language.

Written exams at the Year 10-level involve equal-probability sampling. A lottery system randomly assigns students into *one* of the three written exams: Norwegian (NOR_W), English (ENG_W), and mathematics (MAT_W). This planned missingness enables written exam grades to be modelled under the missing completely at random (MCAR) assumption (Little & Rubin, 2019). Even if the lottery is less than perfectly random, Rasch models are still valid under the weaker assumption of missing at random (MAR), hence "ignorable" (Molenaar, 1995), as long as the missing propensities are unrelated to either item- or person-parameters. This practice is in agreement with previous studies (e.g., He et al., 2018) that utilised Rasch models for handling missing values for score matrices with sufficient subject overlaps.

Similar to written exams, oral exams consist of the same three subjects plus a wide selection of electives (e.g., advanced natural sciences) with students being randomly assigned into *one* oral exam by lottery. In order to form teacher assigned-, written-, and oral-exam grade comparisons, we selected oral Norwegian (NOR_O), oral English (ENG_O), and oral

mathematics (MAT_O) for analyses.

In summary, this study contains a final population size of $N = 60,618$ students. Twelve teacher-assigned grades, three written- and three oral-exam grades formed 18 "Rasch items" for subsequent IRT modelling. Detailed description about each subject is available in Table 1.

Rasch Model and Difficulty Measures

A Rasch model is a unidimensional IRT model with the assumption that the probability of a student's correct response to an item is a function of the student's ability and the item's difficulty (Rasch, 1960). Rasch models are powerful tools for analysing both dichotomous and polytomous ordinal data thanks to its ability to accommodate missing values and its capability to estimate person- and item-parameters simultaneously (de Ayala, 2020). Similar to He et al. (2018), this study models the 18 Norwegian GPA subjects as Rasch items—manifest outcomes of each candidate's latent scholastic capability—and consider two difficulty measures. We operationalise each subject's *overall difficulty* as the expected grade for a candidate possessing average competency ($\mathbb{E}(x \mid \theta = 0)$). We further decompose each GPA subject's *grade-level difficulties* by examining the competencies students need to transition onto the next grade level (difficulty thresholds, δ_k).

Multiple specifications of the Rasch models have been proposed to address different analytical demands. Masters's (1982) partial credit model (PCM) is particularly suitable for the current study given GPAs in Norway are *unweighted* sum scores across all subjects. The PCM model states that for a polytomous item with a maximum available score of m, the probability $\mathbb{P}(\theta, x)$ of a candidate with latent ability θ scoring x on a subject can be expressed as

$$\mathbb{P}(\theta, x) = \begin{cases} \dfrac{1}{1 + \sum_{j=1}^{5} \exp\left\{\sum_{k=1}^{j} (\theta - \delta_k)\right\}} & \text{for } x = 0, \\ \dfrac{\exp\left\{\sum_{k=1}^{x} (\theta - \delta_k)\right\}}{1 + \sum_{j=1}^{5} \exp\left\{\sum_{k=1}^{j} (\theta - \delta_k)\right\}} & \text{for } x = 1, \ldots, 5, \end{cases} \tag{1}$$

where θ is the latent competency of the candidate, and δ_k is the location of the k-th threshold on the latent ability continuum. Since Norwegian students receive grades between 1 and 6, $m = [0, 5]$ in this study. In addition, when each two adjacent grade curves intersect, a subject with six grades would generate five thresholds $(\delta_1, \ldots, \delta_5)$.

Estimation Procedures

Although Rasch models accommodate missing values well, certain output such as infit and outfit statistics are only computable under full data matrices (Chalmers, 2022). We therefore apply multiple imputations (MI) to the score matrix under the MCAR assumption (Little & Rubin, 2019). Each of the ten MI datasets is then analysed separately using the R package mirt (Version 1.38.1, Chalmers, 2022), then pulled together following Rubin's rules (Rubin, 1987).

Results

Descriptive Statistics

Table 1 summarised key information about the 18 GPA subjects examined by this study, including the number of valid entries, grade distributions, and links to official documentation. It is firstly noticeable that data missing rates differed significantly across modes of assessment. Teacher-assigned grades carried small missing percentages most under 5 percent, hence imposing little concerns over estimation bias. Although written- and oral-exams had large missing percentages, this was the effect of the equal-probability sampling procedures. Under planned missingness, the observed grades represent unbiased estimates of true grades despite only 1/3 or 1/5 of the students were studied.

Secondly, grade distributions differed both between- and within-modes of assessment. A large number of grade counts clustered around Grade 3 and 4 for external exams, whereas teacher-assigned grades peaked at different bands depending on the subject, with MATH mainly covering Grade 2 to 4 while FOOD covering largely Grade 4 and 5.

Subject Difficulties

Overall Difficulties

GPA subjects' overall difficulties are shown in Figure 1. Using MATH as an example, the horizontal axis of Panel A represents students' latent competencies, ranging from low ($\theta = -10$) to high ($\theta = 5$), and the vertical axis represents grades randing from 1 to 6. Students with low competencies are expected to receive Grade 1 while Grade 6 is reserved to students with very high competencies. Mapping every competency level to its expected grade yields the sigmoid curve in Panel A. Furthermore, there exists a median student, who evenly divides MATH's observations into 50% below, and 50% above him/her, whose θ is defined as

zero. Tracing this median student's expected score from the curve in Panel A, one reads a grade of 3.64 as the *overall difficulty* for MATH. Repeating this procedure for all 18 GPA subjects gave rise to the scatter plot in Panel B. Subject with low expected grades such as MATH are more difficult while PHED and FOOD are easy subjects evidenced by the high expected grades from median students.

Ranked by overall difficulties, teacher-assigned grades appeared to align themselves along the *manu–mente* dichotomy. A median student is expected to receive a score one grade lower in the most difficult subject MATH than from the easiest one PHED. Written exams are more difficult than oral exams, with NOR_W being more difficult than teacher-assigned MATH. Oral English exam, in contrast, is comparable in difficulty to *mente* subjects such as teacher-assigned FOOD.

Grade-level Difficulties

This study operationalises grade-level difficulties using difficulty thresholds. For a polytomous IRT item such as MATH, a category characteristic curve (CCC) describes the likelihood a particular grade is received by students with varying competency levels. The $P1$ curve in Panel A Figure 2, for example, states that Grade 1 is awarded to students with low competencies almost surely (probability approaching 1) but to those with high competencies almost never (probability approaching 0). Similarly, the $P2$ curve suggests that Grade 2 is most likely to be awarded to students with competencies between approximately $\theta = [-6, -1]$ but low probabilities outside this domain. The intersection between $P1$ and $P2$ marks a difficulty threshold δ_1, above which the next grade is more likely. Six CCCs produce five difficulty thresholds $\delta_1, \ldots, \delta_5$, which concisely summarise each subject's *grade-level difficulties*. Repeating this procedure to all 18 GPA subjects produces Panel B.

Among teacher-assigned grades, the competency demands for receiving a particular grade differed widely depending on the low- and high-end of the grading scale. High consistency was observed at the δ_5-level where all subjects required students to have high competencies ($\theta \approx 2.5$) to transitions from Grade 5 to the top grade 6. As one moves down the grade ladder, however, the difficulty gap expanded to more than one grade between the most difficult subject and the easiest one such that a Grade 3 in MATH is more comparable to a Grade 4 in FOOD. Lastly, the lengthening 95% confidence intervals in δ_1 suggests that teachers did not fully utilise the entire grade scale, especially for the *manu* subjects—an

observation corroborated by the grade distributions in Table 1.

Model Fit Measures and Information Curves

Figure 3 visualises the Rasch model fit statistics using the 2019 Year 10 GPA data. A model with perfect fit would generate an information weighted fit (infit) and unweighted fit mean square (outfit) of 1 (Wu et al., 2016). Infit and outfit mean squares below 1 suggest overfit where the item is more discriminating than the average item discrimination. Resultantly, Wu et al. (2016) consider high quality items (*mente* subjects) to have mean squares less than 1 even though some of these items may be deemed as misfitting the model. GPA subjects with mean squares much greater than 1 are deemed poorer IRT items. Under these criteria, teacher-assigned grades for *manu* subjects HAND, and PHED showed poor model fit, as well as oral English exam grades.

Lastly, Figure 4 presents the information curves (left scale, blue) for the 18 GPA subjects. An information curve plots the information function against the latent competency. The information function is the expected information gained from a student's response to an item given their competency level. Figure 4 also displays the standard error curves (right scale, red) that communicate the precision of each Rasch item over the competency range. The information and standard error curves jointly suggest that the Rasch model used in this study provided strong explanatory power and high precision over the mid-range of the latent competency scale where most students reside.

References

de Ayala, R. J. (2020). *The theory and practice of item response theory* (2nd ed.). Guilford.

Chalmers, P. (2022). *Package 'mirt'*. https://cran.r-project.org/web/packages/mirt/mirt.pdf

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

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

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: Foundations, recent developments, and applications* (pp. 39–51). Springer-Verlag. https://doi.org/10.1007/978-1-4612-4230-7_3

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

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

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer. https://doi.org/10.1007/978-981-10-3302-5

Table 1

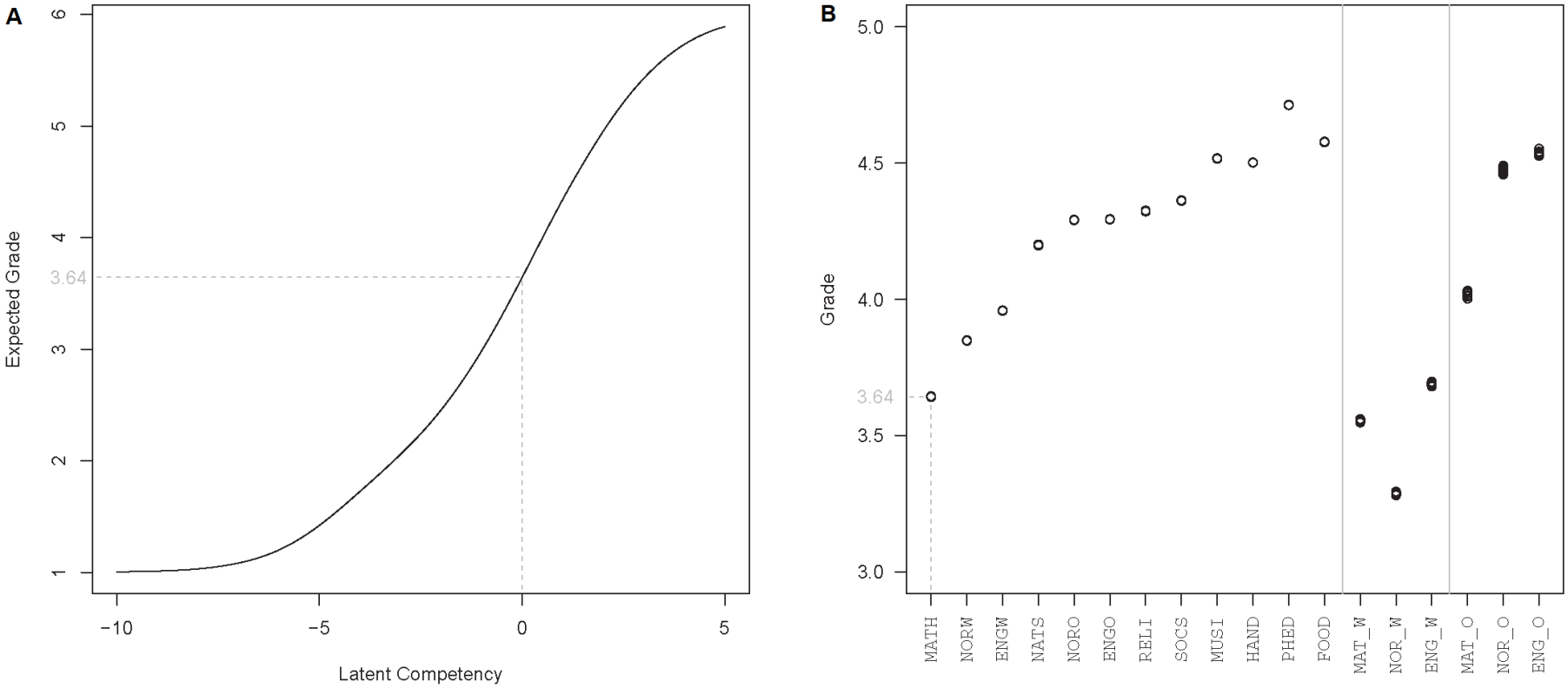
Descriptive Statistics for the 18 GPA Subjects

Subject Code	Subject Name	Valid Entries	Missing (n)	Missing (%)	Grade Frequency 1	2	3	4	5	6	UDIR Course Code
Teacher-assigned Grades (12 subjects)											
MATH	Mathematics	59,184	1,434	2.37	1,165	10,086	15,447	15,443	12,520	4,523	MAT0010
NORW	Written Norwegian	58,946	1,672	2.76	498	5,504	15,503	20,482	13,888	3,071	NOR0214 & NOR0041
ENGW	Written English	59,047	1,571	2.59	850	5,315	13,671	19,934	14,937	4,440	ENG0012
NATS	Natural Sciences	59,642	976	1.61	490	4,801	12,134	17,179	17,448	7,590	NAT0010 & NAT0020
NORO	Oral Norwegian	58,982	1,636	2.70	210	2,994	10,855	18,764	19,254	6,905	NOR0216 & NOR0042
ENGO	Oral English	59,148	1,470	2.43	441	3,200	9,938	19,725	18,947	6,897	ENG0013
RELI	Religion	56,993	3,625	5.98	316	3,386	9,928	16,966	18,293	8,104	RLE0030 & REL0040
SOCS	Social Sciences	59,715	903	1.49	307	3,466	10,489	17,527	19,538	8,388	SAF0010 & SAF0020
MUSI	Music	57,716	2,902	4.79	120	1,450	6,771	18,830	22,940	7,605	MUS0010 & MUS0020
HAND	Arts and Handcraft	58,001	2,617	4.32	93	1,142	6,876	19,744	23,140	7,006	KHV0010 & KHV0020
PHED	Physical Education	57,731	2,887	4.76	102	913	4,612	16,606	26,037	9,461	KRO0020
FOOD	Food and Health	57,683	2,935	4.84	16	561	6,060	19,098	23,907	8,041	MHE0010 & MHE0020
Written Exam Grades (3 subjects)											
MAT_W	Written Mathematics	15,252	45,366	74.84	235	2,482	4,261	4,529	2,902	843	MAT0010
NOR_W	Written Norwegian	13,851	46,767	77.15	237	2,345	5,245	4,106	1,616	302	NOR0214
ENG_W	Written English	14,723	45,895	75.71	225	1,573	4,224	4,940	2,913	848	ENG0012
Oral Exam Grades (3 subjects)											
MAT_O	Oral Mathematics	8,838	51,780	85.42	16	770	2,041	2,503	1,958	1,550	MAT0011
NOR_O	Oral Norwegian	9,310	51,308	84.64	34	459	1,651	2,520	2,342	2,304	NOR0216
ENG_O	Oral English	9,207	51,411	84.81	36	330	1,405	2,651	2,452	2,333	ENG0013

Note. Missing counts (n) and percentages (%) were computed relative to the population size $N = 60,618$. Official documentation about each subject is available from the Norwegian Ministry of Education (UDIR) database by clicking each hyperlink. Subjects offered in both Norwegian and Sami as instruction languages are merged, with both codes given in the UDIR column.

Figure 1

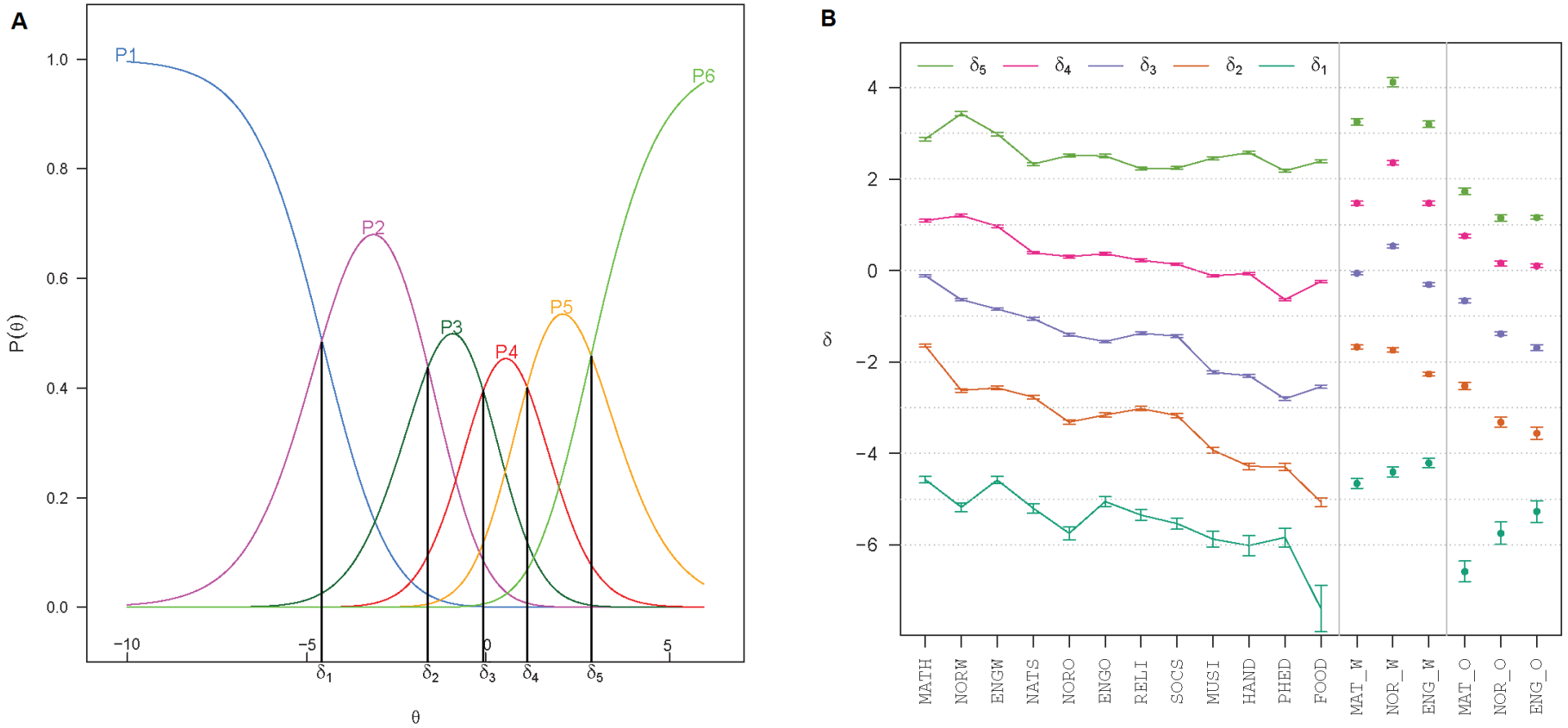
Overall Difficulties



Note. In Panel A, a median student evenly divides MATH candidates into 50% below, and 50% above him/her, whose θ is defined as zero. The expected grade of this median student 3.64 represents the *overall difficulty* for MATH. Repeating this procedure for all 18 GPA subjects produces the scatter plot in Panel B. Subject with low expected grades such as MATH are more difficult while PHED and FOOD are easy subjects evidenced by the high expected grades from median students. Written- and oral-exams' overall difficulties are also shown. Results from ten imputed datasets were superimposed, leading to jitters in exam grades resultant from slightly larger imputation variations.

Figure 2

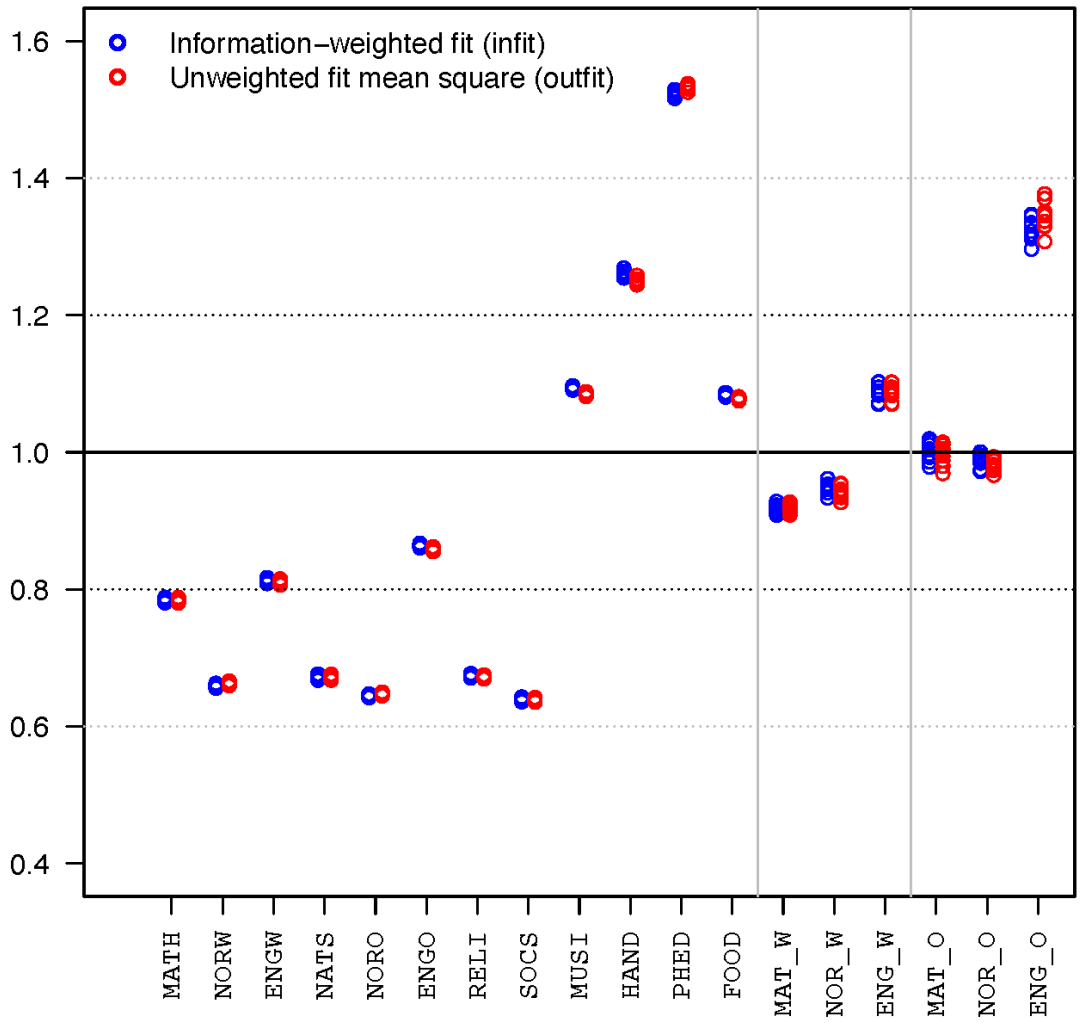
Grade-level Difficulties



Note. Panel A illustrates the category characteristic curve (CCC) for MATH. The vertical axis represents probabilities ranging from 0 to 1, and the horizontal axis represents students' latent competencies ranging from low ($\theta = -10$) to high ($\theta = 5$). The CCC $P1$, for example, describes the association between competency levels and the likelihood students posessing this competency would receive Grade 1. The intersection between $P1$ and $P2$ marks a difficulty threshold δ_1, above which Grade 2 is a more likely outcome. Six CCCs produce five thresholds $\delta_1, \ldots, \delta_5$, which concisely summarise each subject's *grade-level difficulties*. Repeating this procedure to all 18 GPA subjects produces Panel B. The 95% confidence intervals are pooled over ten imputed datasets.

Figure 3

Model Fit Measures



Note. A perfectly fit item in a Rasch model corresponds to infit and outfit statistics of 1. Fit measure below 1
is more discriminating than the average item discrimination. Overfitting is usually not a problem comparing
rules suggests close examination of items with infit and outfit statistics between 1.2 and 1.5 (Wu et al., 2016)
PHED fit poorly into "*the* general scholastic aptitude" construction of GPA. Oral English exam is also deviatin

Figure 4

Information and Standard Error Curves