# 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. Since different elective subjects are treated *equally* in its calculation, GPA implicitly assumes that grades across different specialised subjects are *equivalent* indicators of students' preparedness for higher education—an assumption that remains untested and questioned by descriptive statistics (Utdanningsdirektoratet, 2022). This paper focuses on the comparability of difficulty levels across subjects to provide a test of the hidden assumption in the current procedure for producing the GPA.

### Theoretical Framework

Fairness is both an essential and an elusive integral of educational assessment. Following Gipps and Stobart's (2009) social-cultural framing of assessment fairness and Tierny's (2017) democratic–measurement–pedagogical construction, the current study models GPA as a selection device (Kane, 2013) for accessing privileged social resources (Bourdieu, 1973). It addresses the construct validity of GPAs by examining any construct-irrelevant variance (Messick, 1989) related to students' subject choices.

### Methodology

Item response theory 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. Registry data containing Norwegian students' GPA performance in 2019 are first regularised by removing subjects with fewer than 1,000 candidates 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, group invariance tests are applied 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. We expect Norway's GPA subjects to differ in difficulties (He et al., 2018) and to exhibit significant selection effects (Korobko et al., 2008).

**Relevance to Nordic Educational Research**

Given that university entries in Europe is largely based on the final grades from secondary schooling, the presented analysis is likely to be relevant to other countries using grades as the selection criteria into tertiary education. The issue of potential unequal treatment of students with different specialisation in upper secondary school applies beyond the Norwegian context. By testing the assumption that grades from different specialities support GPA's selection purpose equally well, this study lends statistical support 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


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

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

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

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

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

# Analysis Code, Additional Tables and Figures

## Register Data Re-shaping

```r
###### ADMIN INFO ######
# Date: 26 April 2022
# Author: Tony Tan
# Email: tctan@uio.no
# Position: PhD candidate
# Organisation: CEMO, UV, UiO
# Script purpose: Re-format teacher-assigned marks into student-by-subject shape

###### DATA PROTECTION ######
# Nature: An R script sourcing Norwegian registry data leading to files
# containing equally sensitive personal info
# Security level (input-script-output): black-green-black
# Computer environment (store-view-edit-execute): any-any-any-TSD

#####      Begin script      #####
 ###                          ###
  #                            #

# Point working directory to the location of all registry datasets,
# depending on OS
if (Sys.info()["sysname"] == "Windows") {
    setwd("N:/durable/data/registers")
} else {
    setwd("/tsd/p1708/data/durable/data/registers")
}
if (interactive()) {getwd()} else {cat(paste0(
    "Working directory is now set to ", getwd(), "\n"
))}

# Read in W21_4952_TAB_KAR_GRS.csv
if (!interactive()) {print("Start data loading...")}
gpa <- data.table::fread("W21_4952_TAB_KAR_GRS.csv")
if (interactive()) {names(gpa)} else {print("Data loading complete.")}

# Only keep 2019 data
# STP (Teacher assigned marks)
teacher_mk <- gpa[which(gpa$AVGDATO == 201906), c(1:4, 7)]
# Save the total number of students
n_student <- dim(teacher_mk)[1] # Should be 1,073,204 obs
if (interactive()) {n_student}

# Inspect unusual marks in the "STP" column
if (interactive()) {
    table(unlist(teacher_mk$STP))
}
# These marks are not usable:
#   '' empty [n = 20,042],
#   7 [n = 33],
#   D [n = 58,182],
#   F [n = 37,273],
#   GK [n = 55],
#   IM [n = 2],
#   IV [n = 12,576].

# Recode un-usable STP into NA
teacher_mk$STP <- car::recode(teacher_mk$STP, "
    c('', '7', 'D', 'F', 'GK', 'IM', 'IV') = NA
")



# Part 1: Re-shape teacher-assigned marks columns: one subject per column

# How many subjects there are? (Answer: 200 different subjects in total)
# How many times each subject name appeared (with or without valid score)?
subject_frequency <- sort(table(unlist(teacher_mk$FAGKODE)), decreasing = T)
if (interactive()) {subject_frequency}
```

```r
# Save subject list
subject_list <- as.character(data.frame(subject_frequency)[, 1])
# Save total number of subjects
n_subject <- length(subject_list)
if (interactive()) {n_subject} # Should be 200 subjects in total

# Create a placeholder spreadsheet
stp_spreadsheet <- data.frame(matrix(NA, nrow = n_student, ncol = n_subject))
colnames(stp_spreadsheet) <- subject_list

# Stitch STP and this empty placeholder spreadsheet together
teacher_reshape <- cbind(teacher_mk, stp_spreadsheet)
if (interactive()) {names(teacher_reshape)}

# Set up a progress bar
n_iter <- dim(teacher_reshape)[2] # Set the progress bar's end point
pb <- progress::progress_bar$new( # Refresh progress bar's internal definition
    format = "(:spin) [:bar] :percent [Elapsed time: :elapsedfull || Estimated time remaining: :
    ↪ eta]",
    total = n_iter,
    complete = "=",
    incomplete = "-",
    current = ">",
    clear = F,
    width = 100
)

for (j in 6:n_iter) { # 200 cycles
    # Insert progress bar here
    progress::pb$tick() # Update progress bar

    # Create a placeholder list
    temp <- rep(names(teacher_reshape)[j], n_student)
    # Test whether subject names match
    equal_test <- temp == teacher_reshape[, 4]
    # Turn FALSE/TRUE to 0/1
    equal_test <- equal_test + 0

    # If subject name matches, copy-paste teacher-assign marks
    # into the temp_subject column
    temp_subject <- equal_test * teacher_reshape[, 5]
    # Turn off list property (in order to recode)
    temp_subject <- as.numeric(unlist(temp_subject))
    # Recode 0 to NA
    teacher_reshape[, j] <- car::recode(temp_subject, "0 = NA")
}
cat("\n") # Start a new line once progress bar is full

# Remove subject name and STP columns
teacher_reshaped <- teacher_reshape[, -c(4, 5)]
# Inspect the newly shaped data set
if (interactive()) {head(teacher_reshaped, 20)}

# Save to external file.
if (Sys.info()["sysname"] == "Windows") {
    data.table::fwrite(teacher_reshaped,
        "M:/p1708-tctan/Documents/teacher0.csv",
        row.names = F
    )
} else {
    data.table::fwrite(teacher_reshaped,
        "/tsd/p1708/home/p1708-tctan/Documents/teacher0.csv",
        row.names = F
    )
}
# Should be 239,329 KB in size



# Part 2: Re-shape rows: one student per row

# How many (unique) students there are? (Answer: 64,918 unique students)
# How many times each student ID appeared (with or without valid score)?
```

```r
student_frequency <- data.frame(sort(
    table(unlist(teacher_reshaped$w21_4952_lopenr_person)),
    decreasing = T
))
# Display the top 20 students who took the most number of subjects
head(student_frequency, 20)
# Display the bottom 20 students who took the least number of subjects
tail(student_frequency, 20)
# Save student list
student_list <- as.character(student_frequency[, 1])
# Save total number of unique students
(n_unique_student <- length(student_list)) # 64,918 unique students

# Set up a placeholder spreadsheet
teacher_reshaped_final <- matrix(
    nrow = n_unique_student, ncol = dim(teacher_reshaped)[2]
)
colnames(teacher_reshaped_final) <- names(teacher_reshaped)
teacher_reshaped_final <- data.frame(teacher_reshaped_final)

# Prepare multi-core processing
if (Sys.info()["sysname"] == "Windows") { # Windows can only use single core
    n_cores <- 1
} else { # Both Linux and Mac can implement multicore
    n_cores <- parallel::detectCores() # Count the total number of CPU cores
    n_cores <- n_cores - 1 # Reserve one core for system admin
}

# Set up a progress bar
n_iter <- n_unique_student  # Set the progress bar's end point
pb <- progress::progress_bar$new( # Refresh progress bar's internal definition
    format = "(:spin) [:bar] :percent [Elapsed time: :elapsedfull || Estimated time remaining: :
    ↪ eta]",
    total = n_iter,
    complete = "=",
    incomplete = "-",
    current = ">",
    clear = F,
    width = 100
)

for(i in 1:n_iter) {
    # Insert progress bar here
    progress::pb$tick()

    # Pull out lines that share the same Student ID
    student_temp <- teacher_reshaped[which(
        teacher_reshaped[, 1] == student_list[i]
    ), ]
    # Collapse multiple lines into one line
    student_temp_teacher <- parallel::mclapply(student_temp[, -c(1:3)],
    function(x) max(x, na.rm = T), mc.cores = n_cores)
    # In cases where, same person, same subject, but multiple marks,
    # take the maximum, because I do not know which score was given first.
    # When I asked R to compute max from a column containing NA only,
    # R produced -Inf and a warning.
    # Safe to ignore these warnings and turn -Inf to NA.
    # Recode 0 and -Inf to NA
    student_temp_teacher <- car::recode(student_temp_teacher, "
        c('0', '-Inf') = NA
    ")
    # Stitch admin variables to student_temp_teacher (need transpose)
    teacher_reshaped_final[i, ] <- data.frame(cbind(
        student_temp[1, c(1:3)], t(student_temp_teacher)
    ))
}
cat("\n") # Start a new line once progress bar is full

# Save the standard Student ID list for subsequent work
if (Sys.info()["sysname"] == "Windows") {
    write.table(teacher_reshaped_final[, 1],
        "M:/p1708-tctan/Documents/student_id.csv",
        row.names = F, col.names = c("student_id")
```

```
    )
} else {
    write.table(teacher_reshaped_final[, 1],
        "/tsd/p1708/home/p1708-tctan/Documents/student_id.csv",
        row.names = F, col.names = "student_id"
    )
}
# Should be 888 KB in size

# Save teacher-assigned marks
if (Sys.info()["sysname"] == "Windows") {
    data.table::fwrite(teacher_reshaped_final,
        "M:/p1708-tctan/Documents/teacher1.csv",
        row.names = F
    )
} else {
    data.table::fwrite(teacher_reshaped_final,
        "/tsd/p1708/home/p1708-tctan/Documents/teacher1.csv",
        row.names = F
    )
}
# Should be 15,345 KB in size

  #                           #
 ###                         ###
#####       End script       #####
```

## IRT Analyses Output

### Table 1

*Generalised Partial Credit Model (GPCM) Parameter Estimates*

| Subject Code | Subject Name | $a$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ |
|---|---|---|---|---|---|---|---|
| NORW | Written Norwegian | 3.021 (0.025) | −2.882 (0.026) | −1.535 (0.010) | −0.403 (0.006) | 0.627 (0.007) | 1.789 (0.011) |
| NORO | Oral Norwegiani | 3.346 (0.028) | −3.024 (0.031) | −1.845 (0.011) | −0.817 (0.007) | 0.154 (0.006) | 1.300 (0.008) |
| ENGW | Written English | 1.790 (0.015) | −2.875 (0.029) | −1.638 (0.013) | −0.559 (0.009) | 0.534 (0.008) | 1.741 (0.013) |
| ENGO | Oral English | 1.689 (0.014) | −3.185 (0.041) | −2.024 (0.016) | −1.001 (0.010) | 0.175 (0.008) | 1.473 (0.011) |
| MATH | Mathematics | 1.715 (0.014) | −2.773 (0.024) | −1.027 (0.010) | −0.113 (0.009) | 0619 (0.009) | 1.694 (0.013) |
| NATS | Natural Sciences | 2.656 (0.022) | −2.877 (0.026) | −1.602 (0.010) | −0.640 (0.007) | 0.206 (0.007) | 1.266 (0.009) |
| SOCS | Social Sciences | 3.397 (0.028) | −2.920 (0.027) | −1.766 (0.010) | −0.826 (0.007) | 0.074 (0.006) | 1.170 (0.008) |
| RELI | Religion and Ethics | 3.154 (0.026) | −2.850 (0.025) | −1.715 (0.011) | −0.800 (0.007) | 0.108 (0.006) | 1.181 (0.008) |
| MUSI | Music | 1.331 (0.011) | −3.832 (0.077) | −2.558 (0.026) | −1.493 (0.014) | −0.140 (0.009) | 1.559 (0.013) |
| HAND | Arts and Handcraft | 1.138 (0.010) | −4.129 (0.101) | −2.924 (0.032) | −1.616 (0.016) | −0.124 (0.010) | 1.751 (0.016) |
| FOOD | Food and Health | 1.429 (0.012) | −4.565 (0.173) | −3.263 (0.037) | −1.644 (0.014) | −0.208 (0.009) | 1.479 (0.012) |
| PHED | Physical Education | 0.804 (0.008) | −4.483 (0.138) | −3.264 (0.050) | −2.295 (0.026) | −0.658 (0.015) | 1.750 (0.019) |

*Note.* A generalised partial credit model (GPCM) computes the discrimination ($a$) and difficulty ($b$) parameters for each subject. Standard errors are enclosed in parenthesis below point estimates. All estimates are significant at .001 level.