A Baseline Statistical Model for Analyzing PISA Data in Denmark

Hans Bay, University of Copenhagen, Denmark

The Washington DC Conference on Education 2026
Official Conference Proceedings

Abstract

This paper presents a baseline statistical model for analysing student performance using PISA 2022 data from Denmark. The model draws on student-level background variables, including gender, immigrant background, and socioeconomic status (ESCS), to explain variation in mathematics achievement. Descriptive results show systematic differences across groups, with higher scores observed among boys and students without an immigrant background. A multilevel regression framework is applied to estimate the contribution of individual and school-level factors. The inclusion of school-level socioeconomic composition further improves model fit. In an extended version of the model, the mathematics anxiety index (ANXMAT) is incorporated, leading to a substantial increase in explanatory power. Although the results demonstrate strong statistical associations, they do not imply causal relationships. The paper illustrates a parsimonious modelling strategy suitable as a baseline for comparative analyses of PISA data.

Keywords: PISA, socioeconomic status, immigrant background, mathematics anxiety

iafor
The International Academic Forum
www.iafor.org

Introduction

PISA (Programme for International Student Assessment) is an international survey conducted every three years by the OECD. Its purpose is to assess how well 15 to 16-year-old students apply academic competencies in real-life situations. The test covers reading, mathematics, and science, as well as a fourth domain that varies between cycles. In addition, extensive background data are collected for analyses of well-being, learning, and development.

Methodology

The analysis is based on the Danish PISA 2022 sample, comprising students born in 2006 and enrolled in formal education. The analytical sample includes approximately 6,200 students across 347 schools, representing a population of 56,909 students.

The dependent variable is the PISA mathematics score, scaled to an OECD mean of 500 and a standard deviation of 100. Independent variables include gender, immigrant background, and socioeconomic status (ESCS), which is a composite index derived from parental education, occupation, and home resources.

In addition, a school-level socioeconomic indicator (S_ESCS) is included to account for contextual effects. The model specification allows for interaction effects between immigrant background and socioeconomic status, reflecting potential heterogeneity in how resources translate into academic outcomes.

All estimates are calculated using sampling weights and replication weights. Plausible values and BRR replication weights provided by the OECD were used in all analyses.

Variables:
Gender: Female = 1, Male = 2.

Immigrant background: Based on birthplace of the student and parents (native student = 0, student with immigrant background = 1). The indicator is derived from the OECD variable *IMMIG*.

Socioeconomic status (ESCS): A standardized index based on parental education, occupation, and home resources.

Mathematics score: Scaled with an OECD average of 500 and a standard deviation of 100.

Results

Descriptive Analysis

Table 1
Mathematics Score by Gender and Immigrant Background

Group	Sample size (n)	Score	SE	Population
All students	6,200	489.3	1.84	56,909
Females	3,066	483.3	1.90	27,893
Males	3,134	495.0	2.47	29,016
Native students	4,608	496.5	1.91	48,596
Immigrant background	1,288	442.1	3.29	5,845
Missing	304	458.0	5.59	2,468

Note. OECD scale (mean = 500, SD = 100).

The results are consistent with OECD findings: male students score higher than female students in mathematics, and native students score higher than students with an immigrant background.

Statistical Model

The baseline model specifies the mathematics score as a function of gender, immigrant background, socioeconomic status, school-level socioeconomic composition, and their interaction (1):

Baseline Model:

$$score_{ij} = \alpha + \beta_1 \cdot gender_{ij} + \beta_2 \cdot immigrant\ background_{ij} + \beta_3 \cdot ESCS_{ij} + \beta_4 \cdot S\_ESCS_j + \beta_5 \cdot (immigrant\ background_{ij} \cdot ESCS_{ij})$$

i=1…6,200 j=1…347 schools. (1)

The score varies according to the following factors:

- Gender
- Immigrant background (two groups)
- Socioeconomic background (ESCS)
- School-level socioeconomic background (S_ESCS)
- The interaction between immigrant background and socioeconomic background

Table 2
Baseline Model Estimates (Mathematics)

Parameter	Estimate	SE	t-value	p-value
Intercept	470.39	2.51	187.33	< .0001
Gender (Female)	−12.61	2.20	−5.55	< .0001
Student with immigrant background	−24.36	3.50	−6.96	< .0001
ESCS	31.83	1.75	18.15	< .0001
ESCS× Student with immigrant background	−13.95	3.12	−4.46	< .0001
S_ESCS	32.31	4.97	6.50	< .0001

Note. N = 5841 (out of 6,200), $R^2 = 0.17$. Weights and replication weights are applied in accordance with OECD's design.

The model estimates are largely consistent with the descriptive analysis. The difference between girls and boys in the model is 12.6 points, while in the descriptive analysis it was 11.7. The difference between students with native background and student with an immigrant background is estimated in the model to be 24.4; in the descriptive analysis it is larger, namely 54.4. The number of respondents in this model is 5,841, which means that 359 students have been omitted.

The effect of socioeconomic background is substantially stronger for native students than for student with an immigrant background. The estimated effect of ESCS is 31.9 for native students. While it is 18.0 (= 31.9 − 13.9) for student with an immigrant background.

Baseline Model, Reading. Denmark 2018

In the 2018 PISA study, the main domain was reading. The same model was used.

Table 3
Baseline Model Estimates (Reading 2018)

Parameter	Estimate	SE	t-value	p-value
Intercept	457.19	2.85	160.15	< .0001
Gender (Female)	27.77	2.57	10.79	< .0001
Student with immigrant background	−34.67	3.45	−10.05	< .0001
ESCS	30.79	1.96	15.74	< .0001
ESCS× Student with immigrant background	−20.19	3.52	−5.73	< .0001
S_ESCS	36.85	4.6	8.00	< .0001

Note. N = 7364 (out of 7,657), $R^2 = 0.17$.

When using the reading score as the dependent variable, the same pattern is seen as before: Here, however, it is the girls who score higher than the boys, and correspondingly the socio-economic background has a significant effect.

This pattern is observed across nearly all countries and PISA cycles, where girls tend to outperform boys in reading.

Analysis of PISA scores should at least include gender, ethnicity, and socio-economic background.

Inclusion of Indices

In each PISA round, the organization behind PISA prepares a series of indices. The indices are based on a battery of questions. The most famous is the socio-economic index called ESCS. However, in each PISA round, the OECD prepares over 20 indices that include well-being, reading enjoyment, math anxiety, etc.

Mathematics Anxiety

ANXMAT is an index based on 6 questions, all of which address the student's relationship with mathematics.

1	I often worry that it will be difficult for me in mathematics classes.
2	I get very tense when I have to do mathematics homework.
3	I get very nervous doing mathematics problems.
4	I feel helpless when doing a mathematics problem.
5	I worry that I will get poor [marks] in mathematics.
6	I feel anxious about failing in mathematics.

The ANXMAT index is based on six items capturing students' attitudes toward mathematics. Responses are recorded on a four-point Likert scale.

Scale	Coding
Strongly agree	1
Agree	2
Disagree	3
Strongly disagree	4

ANXMAT is added into the Baseline Model for mathematics (2):

$$score_{ij} = \alpha + \beta_1 \cdot gender_{ij} + \beta_2 \cdot immigrant\ background_{ij} + \beta_3 \cdot ESCS_{ij} + \beta_4 \cdot S\_ESCS_j + \beta_5 \cdot (immigrant\ background_{ij} \cdot ESCS_{ij}) + \beta_6 \cdot ANXMAT_{ij} \quad (2)$$

Table 4

Estimated Coefficients From the Baseline Model, Mathematics Including ANXMAT

Parameter	Estimate	SE	t-value	p-value
Intercept	465.50	2.65	175.85	<.0001
Gender (Female)	+4.60	2.30	2.00	0.0488
Student with immigrant background	−21.52	3.73	−5.77	<.0001
ESCS	25.33	1.58	16.07	<.0001
ESCS× Student with immigrant background	−9.72	3.86	−2.52	0.0139
S_ESCS	26.62	4.52	5.89	<.0001
ANXMAT	−27.60	0.98	−28.06	<.0001

Note. N = 4886 (missing 1,314 observations) $R^2 = 0.31$.

The index ANXMAT has a significant effect, the explained variance (R^2) has nearly doubled. Furthermore, the sign of Gender (Female) has been reversed, so it is now positive. A causal relation between ANXMAT and the score cannot be established. The observed relationship between mathematics anxiety and achievement may reflect reciprocal feedback mechanisms, in which low performance contributes to increased anxiety, while elevated anxiety further suppresses academic performance.

Discussion

The findings confirm well-established patterns in PISA research, particularly the strong role of socioeconomic background and the persistent performance gap associated with immigrant status. The inclusion of school-level ESCS highlights the importance of contextual effects, suggesting that student outcomes are shaped not only by individual characteristics but also by the composition of the school environment.

The introduction of the ANXMAT index substantially improves model fit, indicating that affective factors such as mathematics anxiety play a significant role in student performance. This result aligns with existing literature emphasising the importance of non-cognitive factors in educational achievement.

However, the cross-sectional nature of PISA data limits causal inference. The observed associations should therefore be interpreted cautiously as reverse causality or omitted variable bias cannot be excluded.

Conclusion

The baseline statistical model provides a transparent and replicable framework for analysing PISA data. By combining key background variables with school-level indicators, the model captures a substantial share of variation in student performance. The analysis demonstrates that socioeconomic status remains the most important predictor, while immigrant background and gender contribute additional explanatory power. The inclusion of mathematics anxiety further improves the model, underscoring the relevance of psychological factors. Future research should explore longitudinal data and more advanced modelling approaches to better address causal relationships and policy implications.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

AI tools (e.g., ChatGPT, Copilot) were used to support structuring, language refinement, and idea development. All generated content has been critically evaluated, revised, and verified by the author, who assumes full responsibility for the final work. AI was not used to generate empirical data or sources.

References

Carey, E., Hill, F., Devine, A., & Szűcs, D. (2016). The chicken or the egg? The direction of the relationship between mathematics anxiety and mathematics performance. *Frontiers in Psychology, 6*, 1987. https://doi.org/10.3389/fpsyg.2015.01987

Ma, X., & Xu, J. (2004). The causal ordering of mathematics anxiety and mathematics achievement: A longitudinal panel analysis. *Journal of Adolescence*, *27*(2), 165–179. https://doi.org/10.1016/j.adolescence.2003.11.003

OECD. (2024). *PISA 2022 technical report*. https://doi.org/10.1787/01820d6d-en

VIVE. (2024). *"PISA 2022 Data og metode – Delrapport"* (in Danish). The Danish Center for Social Science research. https://www.vive.dk/media/pure/5xp670pz/24426577