Evaluating the Impact of Test Motivation on PISA and TIMSS Outcomes

Jasmijn L. Bazen[1], Remco Feskens[2], and Marieke van Onna[3]

[1]Social and Behavioural Sciences, Utrecht University
[2]CITO, University of Twente
[3]CITO

January 7, 2025

Word count: 2457

Ethical approval number: 24-1970

Prospective Journal for Publication: Large-scale Assessments in Education

1 Introduction

1.1 Societal Relevance

In the Netherlands, a lot of commotion has been made by news sources on the deteriorating skills of 15-year-old students based on the Programme for International Student Assessment (PISA) results from 2022 (NOS, 2023; RTL, 2023). The main concern seems to be the reading abilities of the students, but also their competence on mathematics and natural sciences are under fire. The Dutch minister of Primary and Secondary Education, Mariëlle Paul, labelled the situation worrying (zorgelijk) (Paul, 2023), which repeatedly got quoted by the main Dutch news sources. Due to this tumult, new educational programs have been created. Therefore, it is of importance that the correct conclusions are being drawn from the data. Some methodological doubts are being discussed whether this is being done correctly, particularly when taking into account the motivation of students.

1.2 Previous Research

Compared to the previous PISA cycle, there is a decline in the mean proficiency of students on the previously mentioned subjects in the Netherlands compared to other European countries. The report of PISA Netherlands speculates that the decline in abilities is due to COVID-19, a lower reading pleasure, online chatting and/or due to mobile phones (M. Meelissen et al., 2023). Acknowledging that the effect probably cannot be attributed to only one factor. After the PISA test, students answered two questions on a scale from one to ten: ”How much effort did you put into this test?” and ”How much effort would you have invested?”. These responses were included as a measurement of motivation in the original report. The levels were said to be comparable across countries, leading to no further adjustments.

However, it is uncertain whether the students filled out these questions honestly, raising doubt about the consistency of motivation across tests and countries. Social desirability bias, which is the tendency to provide answers perceived as socially acceptable, may have influenced responses (King & Bruner, 2000). Students might overstate their motivation on consequence-free tests compared to graded exams to avoid a feeling of judgement. Thus, could varying motivation levels contribute to the observed performance differences between countries?

1.3 The gap in the research

In the previously mentioned analysis of the PISA results, no motivation component was used to explain potential differences between the scores of different (European) countries, since motivation was deemed to be roughly the same across countries based on the previous questions. However, another indirect motivation monitor is available in the data due to the digital nature of the test:

response times. With these, one can estimate the motivation of a student for a particular question, and thus for the whole test. If a student takes not even a second or two for a question, one can assume they were not very motivated to answer the question to the best of their ability. Thus, a speed-accuracy trade-off can be made, where quick answers can still be correct. Next to this, one can look at whether responses were missing, where more questions left unanswered indicates a lower motivation.

Furthermore, one can wonder whether there is a difference between responses and motivation at different ages, since one could reason that smaller children have a lower effect of consequences on motivation. A study found that younger children are less affected by extrinsic motivation than older ones, since they have a higher academic motivation in general (Smith et al., 2023). Therefore, the Trends in International Mathematics and Science Study (TIMSS) dataset will be used to assess the same information, but for children in the Dutch group 6, which are on average ten years old (M. R. Meelissen & Punter, 2016).

Thus, the research questions that this paper aims to answer are: To what extend does motivation, measured by response time and questions answered, explain the differences in test results between countries? And, are there differences in the motivations and responses at different ages? Where it is rationalised that Dutch students do not have a statistically significant difference in skill level compared to students of other countries, when controlling for motivation. And that there is a significant difference between motivation between younger and older students taking arbitrary tests.

2 Methods

2.1 Data

Since the effect of age as well as motivation on test scores needs to be examined, both the TIMSS data as well as the PISA data will be analysed. The TIMSS data is on about ten year olds, and the PISA on fifteen year olds. Only the mathematics component will be used of the TIMSS data. Both datasets are publicly available. The grade 4 2019 version of the TIMSS data will be used (*TIMSS 2019 International Database*, 2019), and the 2018 PISA data (OECD, 2023). In large-scale assessments like PISA and TIMSS, multiple booklets are used to reduce the number of questions each student must answer. Overlapping some questions across booklets allows researchers to estimate how students might have performed on questions they did not encounter.

All analyses will be performed in the latest version of R 4.3.2 and Rstudio at the time of starting.

2.2 Motivation

Test motivation can affect performance (Wise & DeMars, 2005), although some evidence exists that younger children are less affected by extrinsic motivation

(Smith et al., 2023). This motivation effect is not necessarily a problem, but research suggests motivation may vary between countries (Gneezy et al., 2019), which could threaten the validity of cross-cultural comparisons in studies like PISA and TIMSS. Assessing test motivation is difficult, but can be measured indirectly through response speed or item non-responses. The first step in this study is to operationalise motivation in TIMSS and PISA.

Using the indirect measurement of test motivation has some clear advantages and might provide a more realistic view of test effort. However, it complicates the assessment of the relationship between test effort and performance, as both effects need to be disentangled.

A possible solution is to model both effects simultaneously using joint hierarchical modelling with the LNIRT package in R (Fox et al., 2023). These are models in which ability and speed are included simultaneously. Models by Wim van der Linde (Van Der Linden & Hambleton, 1997; Van Der Linden & van Krimpen-Stoop, 2003; Van der Linden, 2007) and Fox (Fox, 2010) were used as examples.

The indirect measure of motivation was operationalized using several variables. First, the responses to the question "How much effort did you put into this test?" rated from one to ten were included. The question "How much effort would you have invested?" was exluded from this intermediary report.

Second, response times on questions served as indicators of motivation for answering a question correctly. If a question was answered within 5 seconds, this answer was flagged. For every student a total was calculated for how many times they answered a question within 5 seconds. The same was done for questions that were answered after 5 minutes. Both totals were possible markers for an unmotivated student. Additionally, raw response times were included in the model as a separate variable, accounting for effects not modelled by the two previous binary variables.

Third, the total number of missing items per person can indicate how much effort went into answering the test questions correctly. If a student only answers half the questions, this is a heavy indication of low motivation. Thus, this is our final variable measuring our latent variable Motivation.

This could be summarized in the following model:

$$M_i = \alpha + \beta_1 RT_i + \beta_2 LittleTime_i + \beta_3 MuchTime_i + \beta_4 Effort_i + \beta_5 Miss_i + \epsilon_i$$

, where α represents the intercept of motivation, β_1 the effect of a reaction time (RT_i) on motivation, β_2 represents the effect of the total number of times a person took very little time to answer ($LittleTime_i$), and β_3 the effect of the amount of times a person took very long to answer ($MuchTime_i$) on motivation. The effect of effort ($Effort_i$) on motivation was modelled by β_4. β_5 represents the effect of the total number of items per person ($Miss_i$) on motivation. And finally, ϵ_i models the residual error term.

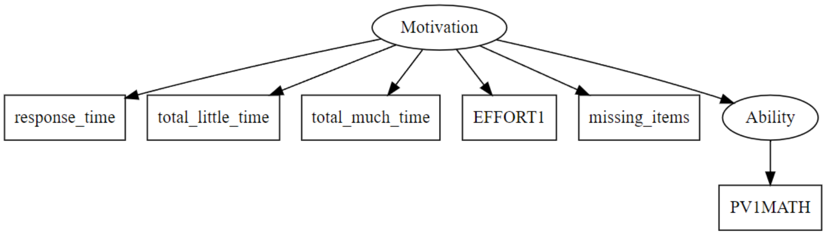


Figure 1: The SEM Model implemented in this research report.

2.2.1 SEM Model

When motivation is defined, we will examine the relationship between motivation and test results using Structural Equation Modelling (SEM) with the R package lavaan (Rosseel, 2012). Then, we will test if age affects this relationship using a multi-group SEM model, also calculated with lavaan in R.

2.3 For this Research Report

This intermediary research report aims to apply a basic SEM model to a subset of the PISA data, focusing on Dutch and German students for one booklet (booklet 7), no TIMSS data was included yet. Using only data from two countries and one booklet reduces computational complexity, making it ideal for the report. The chosen booklet 7 contains 23 items, answered by 360 students from the Netherlands and Germany. Only one plausible value per student will be used for this intermediary analysis, instead of all available ones. Plausible values are random samples taken from the posterior distribution of a latent variable and help reduce biases in analyses due to the inherent randomness (OECD, 2009).

2.3.1 The SEM Model

Both countries were analysed in separate SEM models, using identical models, see Figure 1. The latent variable Motivation was measured using response times, the variables derived from extreme response times, the responses to the motivation-related question, and the number of missing items. The latent variable Ability was modelled by the plausible value. Motivation then regressed on Ability.

3 Results

3.1 Descriptive Analyses

3.1.1 Missingness and Easiness on Items

For this study, missingness was defined as either ”no response” or ”not reached”, with no distinction between the two. The number of missing questions per student was distributed with statistically significant differences between the Netherlands and Germany. These differences were confirmed by a Wilcoxon signed-rank test for non-parametric data, $z = 10.910$, $p < .000$. This difference can be seen in Table 1.

Table 1:
Frequency counts of the number of missing items per person by country.

Number of missing items	Country	
	Germany	The Netherlands
0	81	99
1	42	27
2	23	6
3	12	10
4	15	4
5	11	1
6	11	2
7	5	1
8	2	1
9	0	1
10	3	0
11	1	0
12	1	0
13	0	0
14	0	0
15	1	0

Item easiness was calculated per question as the average score divided by the maximum possible score on an item. Thus, if the item easiness score was high the item was easy. These easiness score were plotted against the percentage of missing answers per item, see Figure 2. The plot reveals that questions with more missing responses tend to have lower easiness scores. The question arises whether missing responses caused the average score to drop, thereby increasing perceived difficulty, or whether the item was inherently difficult, leading to more missing responses.

Additionally, open-ended questions were answered less often than multiple choice questions. Among the multiple-choice questions, more complex items seemed to have a higher rate of non-responses compared to simpler ones. Suggesting motivation playing a role in answering questions.

3.2 The SEM Models

3.2.1 The Netherlands

The SEM model did not seem to fit the data on the Dutch students. There were no improper solutions, nor any obvious estimation problems. The chi-

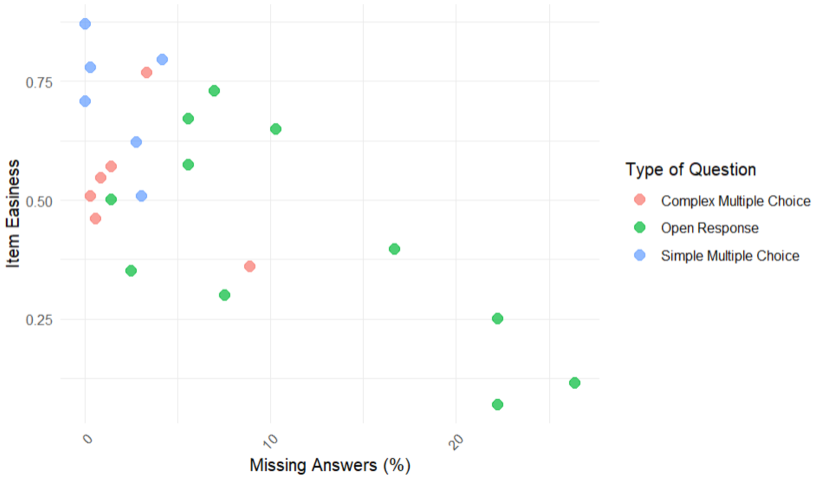


Figure 2: Scatter-plot with the percentage of missing answers per item on the x-axis, and item easiness on the y-axis. Where colour indicates the type of question.

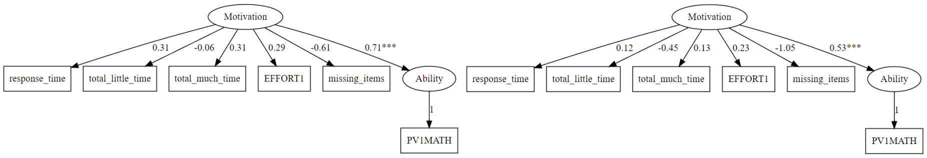
square test was highly significant, meaning that the hypothesis of the model fitting the data has to be rejected. This indicated a bad fit of the model to the data $\chi^2(9) = 295.59, p < .000$. The Root Mean Square Error of Approximation (RMSEA) indicated a bad model fit as well, with an estimate of .10. The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were .68 and .81 respectively, also indicating poor fit. The Standardized Root Mean Square Residual (SRMR) indicated a proper fit with an estimate of .06. This could be due to a large sample size or a high number of parameters.

Thus, the model does not seem to fit the data. It should be looked into further how the SEM model can be constructed, so that it fits the data well.

The only significant effect was that of the latent variable Motivation on Ability, the standardized coefficient was $b_{standardised} = .71, p < .000$, see subfigure 3a. Thus, for every one standard deviation (SD) increase in the latent variable Motivation, the latent variable of Ability went up with .71 SDs.

3.2.2 Germany

Thee SEM model did not seem to fit the data of the German students either. No improper solutions, nor obvious estimation problems were present. The chi-square test was highly significant again. Thus, the model was a bad fit for the data, $\chi^2(9) = 552.37, p < .000$. The RMSEA indicated a poor model fit, with an estimate of .12. The CFI and TLI also indicated a poor fit, with estimates of were .85 and .75 respectively. The SRMR indicated a proper fit with an estimate of .07. In conclusion, the second model also does not seem to fit the data.



(a) The Netherlands

(b) Germany

Figure 3: The standardized results of the SEM model applied to both countries separately. Where each coefficient represents the expected change in standard deviations in the outcome variable for a one-standard-deviation increase in the predictor variable.

Just like in the first model, the only significant effect was that of the latent variable Motivation on Ability, the standardized coefficient was $b_{standardised} = .53, p < .000$, and thus lower than that of the Dutch model, see sub-figure 3b.

4 Conclusion and Discussion

The model currently does not fit the data, important paths could be missing, while unnecessary paths or non-valid latent variable constructs could be included. However, it provides insights into probable relationships, suggesting a statistically significant effect of the latent variable Motivation on the latent variable Ability. This effect seems to be larger in the Netherlands than in Germany.

4.1 Next Steps

4.1.1 Compatibility of the Data

The TIMSS and PISA data differ in structure, requiring evaluation to see whether a comparison is feasible, and how much weight could be attributed to this. This can be evaluated by for example if the same countries participated in the research, especially the countries that the original report is based on. These countries were: Belgium, Denmark, Germany, Finland, France, Greece, Ireland, Italy, Netherlands, Austria, Portugal, Spain, United Kingdom and Sweden (M. Meelissen et al., 2023). Not all these countries are present in the TIMSS data, which might lead to potential issues. The structures of missing data should be compared as well. Does the PISA dataset have more, fewer or an equal number of missing answers? Finally, PISA records response times per question, whereas the TIMSS sometimes aggregates response times for questions shown on the same screen.

4.1.2 Changes to the Model

The answers to the second motivation question in the PISA dataset, ”How much effort would you have invested?” will be included in the analysis, modelled as the discrepancy from the previous ”How much effort did you put into this test?”. This aims to capture the difference between students' typical test motivation and their motivation for the PISA test.

Then, data on all countries in the original report will be included into the analysis, alongside of the TIMSS data to compare ages. All plausible values will be included instead of only the first, to better model the latent variable Ability.

The SEM model will be expanded into a multilevel structure to account for the hierarchy of students nested within schools and countries, while incorporating age as a variable. Person characteristics such as gender and socioeconomic status will be incorporated, and clearer labels used in figures.

Finally, an optional simulation study will apply different SEM models to simulated datasets, evaluating fit with statistics as the AIC, BIC, RMSEA, and coverage. The best model will then be applied to real data, with simulations varying by motivation levels and age.

References

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* Springer.

Fox, J.-P., Klotzke, K., & Simsek, A. S. (2023). R-package lnirt for joint modeling of response accuracy and times. *PeerJ Computer Science*, *9*, e1232.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, *1*(3), 291–308.

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing*, *17*(2), 79–103.

Meelissen, M., Maassen, N., Gubbels, J., van Langen, A., Valk, J., Dood, C., ... Wolbers, M. (2023). Resultaten pisa-2022 in vogelvlucht.

Meelissen, M. R., & Punter, R. A. (2016). Twintig jaar timss: ontwikkelingen in leerlingprestaties in de exacte vakken in het basisonderwijs 1995-2015.

NOS. (2023, Dec). *Leesvaardigheid nederlandse 15-jarigen verder achteruitgegaan.* NOS Nieuws. Retrieved from `https://nos.nl/artikel/2500415-leesvaardigheid-nederlandse-15-jarigen-verder-achteruitgegaan`

OECD. (2009, April). Use of proficiency levels: Plausible values. In *PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION* (pp. 133–142). Author.

OECD. (2023). *Pisa 2022 results (volume i): The state of learning and equity in education.* OECD Publishing. Retrieved from `https://www.oecd.org/en/publications/pisa-2022-results-volume-i_53f23881-en.html` doi: 10.1787/53f23881-en

Paul, M. (2023). *Kamerbrief bij rapport over resultaten pisa-2022 in vogelvlucht.* Retrieved from `https://open.overheid.nl/documenten/dpc-cc149c0314b92bff1a4e139eb3b6f726b633d2c1/pdf`

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of statistical software*, *48*, 1–36.

RTL. (2023, Dec). *Middelbare scholieren scoren ondermaats: prestaties in lezen en exacte vakken op dieptepunt.* RTL Nieuws. Retrieved from `https://www.rtl.nl/nieuws/nieuws/artikel/5422599/leesvaardigheid-nederlandse-leerlingen-exacte-vakken-achteruit`

Smith, Z. R., Flax, M., Becker, S. P., & Langberg, J. (2023). Academic motivation decreases across adolescence for youth with and without attention-deficit/hyperactivity disorder: Effects of motivation on academic success. *Journal of Child Psychology and Psychiatry*, *64*(9), 1303–1313.

Timss 2019 international database. (2019). IEA, TIMSS PIRLS International Study Center. Retrieved from `https://timss2019.org/international-database/`

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory:

Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1–28). Springer.

Van Der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, *10*(1), 1–17.