Large-scale Assessments in Education

RESEARCH Open Access

Exploring the relationship between process data and contextual variables among Scandinavian students on PISA 2012 mathematics tasks

Denise Reis Costa[1] and Chia-Wen Chen[2*]

*Correspondence: five43@gmail.com

[1] Department of Research, Cancer Registry of Norway, Oslo, Norway
[2] Centre for Educational Measurement (CEMO), University of Oslo, Forskningsparken C1, Gaustadalléen 21, 1161, 0318 Oslo, Norway

Abstract

Given the ongoing development of computer-based tasks, there has been increasing interest in modelling students' behaviour indicators from log file data with contextual variables collected via questionnaires. In this work, we apply a latent regression model to analyse the relationship between latent constructs (i.e., performance, speed, and exploration behaviour) and contextual variables among Scandinavian students (i.e., Norway, Sweden, and Denmark) during their completion of three interactive mathematics tasks in the 2012 Program for International Student Assessment (PISA). The purpose of this study is threefold: (1) to investigate whether a factor model is adequate for explaining the relationship between extracted process indicators from log files and students' performance on the three items; (2) to evaluate whether measurement invariance holds for the three analysed countries; and (3) to determine how well contextual variables [i.e., a student's background, availability, use of information and communication technology (ICT) resources, and learning conditions for mathematical literacy] correlate with the model's latent constructs. Our findings reveal that a three-factor CFA model is well-suited to the analysed data and that a weak measurement invariance model performs best. By including contextual variables in the modelling, we also highlight the differences in Scandinavian students' profiles. Specifically, higher economic social cultural status (ESCS) consistently led to higher math proficiency across all three countries. Norway did not show any gender differences in math proficiency, response time, or exploration behaviours. Experience with math tasks at school led to more exploration behaviours across all three countries. Swedish and Norwegian students who had more experience with pure math tasks at school were likely to obtain higher mathematics achievement scores and tended to interact more frequently with the testing platform when answering the three math items. When Danish students had higher ICT scores (i.e., more ICT resources available at school and home), they had lower response accuracy. There was, however, no effect of ICT on the three factors among Swedish and Norwegian students. Finally, we also discuss the implications and limitations of this study.

Keywords: Computer-based assessment, Process data, ILSA, Scandinavian, Measurement invariance

Introduction

One important advantage of the transition from paper-based to computer-based assessments in educational measurement has been the possibility of using log files as sources of process data (Provasnik, 2021). In the literature, process data can be defined as part of data management of raw log files with the aim to extract any piece of information (e.g., response action or timing) from the computer-generated files (Reis Costa & Leoncio Netto, 2022) or linked to the response process as defined by Provasnik (2021): 'the empirical data that reflect the process of working on a test question—reflecting cognitive and noncognitive, particularly psychological, constructs.' There has been increasing interest in analysing such data since it is possible to gather students' information beyond response accuracy (i.e., correct or incorrect answers). For example, one can extract the amount of time students spend on each task (i.e., time on task, Chen, 2020) as well as data on their interactions with the available tools (e.g., the use of an online calculator, Jiang et al., 2023).

A latent variable framework is a common approach to analysing the relationship between student performance and indicators extracted from log files. For example, De Boeck & Scalise (2019) used a three-factor confirmatory factor analysis (CFA) model to explore the relationships between performance, time, and action variables. Reis Costa and collaborators (2021), in turn, incorporated information from log data into scoring for their analysis of the precision of ability estimates. In a Bayesian framework, Klotzke & Fox (2019) proposed covariance structure modelling for the nested and crossed dependences of data extracted from log files. Other studies have addressed joint modelling by including not only the number of actions but also even more complicated features, such as action sequences, time, response, and background variables (Han et al., 2019; Tang et al., 2020; Ulitzsch et al., 2021). Qiao et al., (2022) conducted a multigroup joint model with the response, response time, and action sequence, which focused on the group invariance of gender. However, rare studies yet showed the framework of cross-country comparison of latent variables measured by the processing indicators together with the measurement invariance examination for processing data in an international large-scale survey.

In this study, we use data from the 2012 computer-based assessment of mathematics (CBAM) from the Program for International Student Assessment (PISA), and our purpose is threefold: (1) to investigate whether a latent variable model is adequate for explaining the relationship between extracted process indicators from log files and students' performance on three items; (2) to evaluate whether measurement invariance holds for a number of selected countries (i.e., Scandinavian); and (3) to determine how well contextual variables [i.e., a student's background, availability, use of information and communication technology (ICT) resources, and learning conditions for mathematical literacy] correlate with the model's latent constructs.

The effects of student-level characteristics (such as personal backgrounds) and their context (e.g., learning environment and opportunity to learn the knowledge content) on outcomes (such as mathematics achievement) have been well studied (Schmidt et al., 2015; Senkbeil & Wittwer, 2013; Wihardini, 2016). However, research on the joint latent model, which incorporates the effects of contextual variables on the measurement model for process data (e.g., time on task or use of an online calculator), is still in its infancy. As the third step of our analysis framework (i.e., after accounting for model adequacy and measurement invariance), we propose a latent regression analysis to further explore students' cognitive processes when answering mathematic problems and differences between PISA participating countries. Before this exploration, however, an examination of invariance for the measurement model is required to ensure that the comparison of effects of contextual variables across different groups is meaningful (Nagengast & Marsh, 2014; Raykov & Marcoulides, 2006).

This study adds to the literature by analysing how data on students' performance, self-reported data from questionnaires, and extracted observed behaviour through process data are related to one another with regard to a set of tasks from the PISA. We work with data from the 2012 CBAM in three publicly released items (i.e., CM015Q01, CM015Q02D, CM015Q03D). They all belong to the CD production unit, which facilitates the interpretation of the model's results, as students are exposed to the same item features and stimulus. We focus our analysis on Scandinavian students (i.e., Norway, Sweden, and Denmark) and investigate measurement invariance for the extracted process data indicators for a meaningful comparison of the model's results among this group of countries.

This study adds to the body of knowledge by applying the De Boeck & Scalise's (2019) model for a set of mathematics items from PISA and also showed a procedure for cross-country comparison of the latent variables for process data, including (1) the measurement invariance examination, which makes the comparison among Scandinavian countries meaningful and (2) latent regression analysis which links country differences to the contextual variables. This paper suggests standard procedures for applied researchers who would like to make meaningful cross-group comparisons of process data from international large-scale assessments (ILSAs) with latent variable models. Due to its nature, however, this study is explorative and possible relationship among explanatory variables and differences in students' process behaviours among Scandinavian countries should be investigated further.

Theoretical background and research questions

Based on the item features, several process indicators can be extracted from log files that can help increase the understanding of student performance. For instance, De Boeck & Scalise (2019) investigated the relationships among performance, invested time, and students' actions in a collaborative task in PISA 2015. By using the sequence mining technique, He et al., (2019a, 2019b) identified generalized patterns of respondents' problem-solving behaviours in the Program for the International Assessment of Adult Competencies (PIAAC).

We focus our analyses on two behaviour indicators—response times and frequency of actions—using process data to explore the relationship between these indicators and students' performance during the three relevant math tasks across the three Scandinavian countries.

Our interest in the analysis of Scandinavian students is twofold. First, these countries have attracted attention in the educational context due to their successful combination of economic performance and social well-being. For example, together with other Nordic countries, they are generally perceived as having the most pronounced equality regarding educational opportunity (Frønes et al., 2020). Although this group of countries is more homogeneous than other PISA participating nations, there is a lack of studies that capture the nuances of how Scandinavian students spend their time and interact with mathematics items. Thus, our second interest is the exploration of how these countries differ in their relationships of contextual factors, students' performance, and process data in a joint framework. Using latent regression modelling, this study aims to fill this gap.

In this study, we have added three types of contextual variables—students' background, availability and use of information and communication technology (ICT) resources, and learning conditions for mathematical literacy—to provide a broader overview of the similarities and differences in students' outcomes among the analysed countries. All datasets are public and available for download on the Organization for Economic Cooperation and Development (OECD) website.

Research question 1: modelling performance and process data indicators from log files

Statistical approaches for modelling response outcomes together with process data indicators have been used to evaluate students' scientific inquiry performance in a problem-solving scenario (Greiff et al., 2015; Scalise & Clarke-Midura, 2018; Teig et al., 2020). For example, Greiff et al., (2015) identified students' mastery of an exploration strategy associated with problem-solving proficiency by analysing students' interaction with a computer-based task. Teig et al., (2020), in turn, applied latent profile analysis to frequency of action, response accuracy, and response time to identify students' profiles of inquiry performance.

Many studies applied psychometrical approaches to analyse the process data for various educational purposes, such as classifying or predicting student performances. Chen (2020) proposed a continuous-time dynamic choice measurement model where they aimed to use the sequence of response actions to predict students' final response to items and overall performance in the test. Ulitzsch et al., (2021) proposed a similarity statistic to connect response time pre-actions and sequence of actions, and based on that, they clustered students' homogeneous response process patterns. Tang et al., (2020) applied a dissimilarity measure in the multidimensional scaling framework to identify the discrepancy between response processes. Han et al., (2019) explored the action features from process data that predicted the item responses.

Among all possible process indicators that can be extracted from log file data, frequency of action has been recognized as one of the indicators associated with students' exploration behaviour and response time as supporting evidence of scientific inquiry processes (Scalise & Clarke-Midura, 2018; Teig et al., 2020). Using such indicators, De Boeck & Scalise (2019) developed a CFA model employing the indicators of frequency

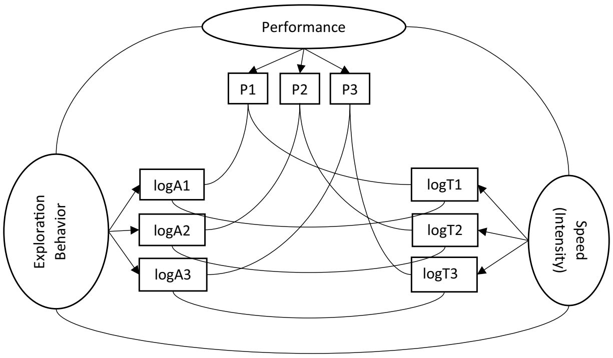


Fig. 1 Adaptation of De Boeck & Scalise's (2019) three-factor model for the analysis of three items in a test. While P1, P2, and P3 measure the response accuracy to the three items in binary outcomes (0 = incorrect and 1 = correct answer), T1, T2, and T3 refer to response times, and A1, A2, and A3 refer to the frequency of actions for each item, respectively. The single-headed dashed arrows indicate the direct effect. The curves between the observed variables or latent variables indicate the correlations

of action, response time, and response accuracy to respectively measure three latent variables: inquiry exploration behaviour, speed, and collaborative problem-solving performance. Figure 1 illustrates De Boeck & Scalise's three-factor CFA model (2019). It is assumed that the local independence condition is satisfied for all three constructs, indicating that differences in the construct fully account for the apparent correlation among the process indicators. This is especially true for assessment items since test developers must take into consideration not only the item contents but also how much time students will spend on each task and what tools are available for their successful completion.

The model developed by De Boeck & Scalise (2019) for processing data (i.e., response times and frequency of actions) together with product data (i.e., response accuracy) has demonstrated a good fit to the response data for four items measuring collaborative problem solving among a sample of students from the United States in PISA 2015. In contrast to classical factor models, the residuals of response accuracy and frequency of action within each item were added to the modelling to predict the residual of response time, and the residual of response accuracy was assumed to be correlated with the residual of frequency of action. In contrast to the usual CFA framework, extra dependencies were added to the modelling (i.e., direct effects and correlated residuals) to capture within-item relationships between process aspects and performance beyond the variance explained by the latent variables. These dependencies are particularly important since the successful completion of the tasks requires a minimum amount of time and may entail the manipulation of specific tools presented in the item stimulus.

In our study, we expanded De Boeck & Scalise's (2019) framework for modelling students' performance and process data indicators by including contextual variables and studying the relationship of these variables across the three Scandinavian countries. By assuming that speed and exploration behaviour factors can be considered educational outcomes, such as mathematical performance, this analysis opens a new avenue of research by investigating how contextual variables are associated with these latent constructs among Scandinavian students. Since group comparisons require that the meaning of constructs remains invariant across the studied countries (Nagengast & Marsh, 2014), we evaluated measurement invariance using a multiple-group approach.

Research question 2: measurement invariance

Measurement invariance evaluation is a psychometric evaluation of the equivalence of model's parameters across groups, time points, or test occasions (Brown, 2015). When measurement invariance does not exist between groups, the operationalized construct and the structure between constructs cannot be meaningfully tested across groups (e.g., hypothesis tests for the regression of one construct on another) because the measured construct or behaviour can have a different meaning for the disparate groups (Vandenberg, 2002). Therefore, measurement invariance is essentially better examined and demonstrated prior to testing the relations between constructs across groups.

Due to its importance, measurement invariance is applied in various fields. For example, Senese et al., (2012) examined measurement invariance across cultural groups before mean difference tests. In clinical psychology research, Kueh et al., (2018) tested measurement invariance between the genders regarding physical activity and the leisure motivation scale among youth. In an international large-scale analysis, Hansson & Gustafsson (2013) investigated the measurement invariance in socioeconomic status between immigrant and nonimmigrant backgrounds in Sweden using data from TIMSS 2003.

Measurement invariance can be tested using a CFA framework. For example, Nagengast & Marsh (2014) studied the invariance in motivation and engagement constructs using data from 57 countries via PISA 2006 through a multiple-group approach. He et al., (2019a, 2019b) also applied multigroup CFA modelling to evaluate the measurement invariance in noncognitive constructs via international large-scale assessments.

Using process data from Scandinavian countries, we tested measurement invariance in latent factors, such as speed and exploration behaviour, using a multiple group CFA. Few studies have explored the measurement invariance in latent constructs associated with process data from large-scale international assessments. Reis Costa et al., (2021), for example, investigated the measurement invariance in time-related variables across countries using PISA 2012 data. Concerning exploration behaviour, however, there is still a lack of studies examining measurement invariance for such constructs across countries participating in international surveys.

Research question 3: effects of contextual variables on modelling

In addition to comparing student achievement, another purpose of international large-scale assessments is to compare student responses to background questionnaires (Kyllonen & Bertling, 2013). These questionnaires include a set of items to gather contextual information, such as students' background, context of instruction, or learning aspects (e.g., students' interests, motivation, and engagement).

In the PISA, a large selection of items from student questionnaires is devoted to contextual factors linked to cognitive and/or noncognitive outcomes (OECD, 2013). Regression analyses are usually the standard for statistically evaluating the effects of these contextual variables on performance. For instance, Senkbeil & Wittwer (2013) ran a multivariate linear regression to evaluate the relationship between mathematical achievement and computer use at home in PISA 2006 while accounting for social background variables.

To the best of our knowledge, this study is one of the first to incorporate contextual variables into the modelling of students' performance and behaviour indicators across different PISA countries. We include such variables to advance the understanding of how contextual factors can influence educational outcomes. Specific to the analysis of process data in ILSA, several factors can play a role in how students respond to an assessment item and how it can be translated into process data indicators. For example, Reis Costa & Leoncio Netto (2022) presented a six-layer ecological framework for capturing the nuances of intrinsic and extrinsic characteristics that can help explain test score variability. With this framework in mind, we limited our study to three layers in our analysis: (1) item characteristics and associated process data, (2) personal characteristics, and (3) school and family context. Figure 2 illustrates our approach.

Specifically, we analysed three types of contextual variables from the PISA 2012 questionnaire: (1) personal background (i.e., gender, socioeconomic status index, language at home, and immigration background); (2) availability and use of ICT resources; and (3) learning conditions for mathematical literacy (i.e., opportunity to learn content). By exploring the effect of these contextual factors on the latent constructs (i.e., mathematic performance, speed, and exploration behaviour), this study may provide insights into how effective and equitable school practices are in Scandinavian countries.

Personal characteristics and family background play an important role in students' academic success. For example, Wihardini (2016) showed that the PISA socioeconomic status (SES) index has a significant effect on students' math performance estimates after controlling for other covariates. Gender is another important control variable that specifically for mathematics tasks, has shown that boys consistently outperform girls on standardized tests (Liu et al., 2008). Together with these two variables, students' migration status and language at home were also analysed, since many studies using PISA data

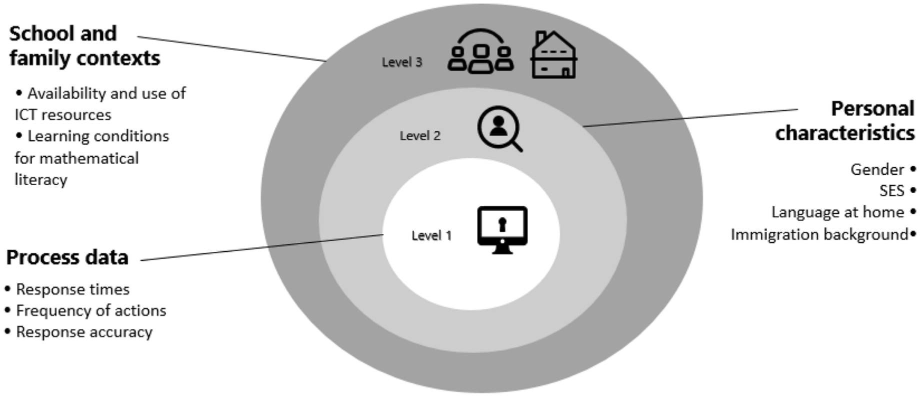


Fig. 2 Framework for the analysis of process data and contextual factors

strive to disentangle the effect of these variables from the socioeconomic effect (OECD, 2013).

In the learning environment, new digital resources have entered students' homes and classrooms in the last decade. The availability of such ICT devices at school or home can influence a student's performance on assessments (Senkbeil & Wittwer, 2013). Given the increase in computer-based tests, we also anticipated that the availability and use of ICT might have a positive effect on exploration behaviour and might influence the speed construct as well. To evaluate this, we explored differences in the latent constructs among Scandinavian students using five PISA scales: ICT availability at home, ICT availability at school, ICT use at home for school-related tasks, use of ICT in mathematics lessons, and use of ICT at school.

Using PISA's opportunity to learn (OTL) indices relating to student-perceived experiences and familiarity with mathematical tasks, we also explored the effect of learning conditions on the extracted process indicators among Scandinavian countries. Since formal mathematics OTL has a strong relationship with student achievement (Schmidt et al., 2015), we anticipated that such variables would also play an important role in the time students spend on the test and in how intensely students interact with test items.

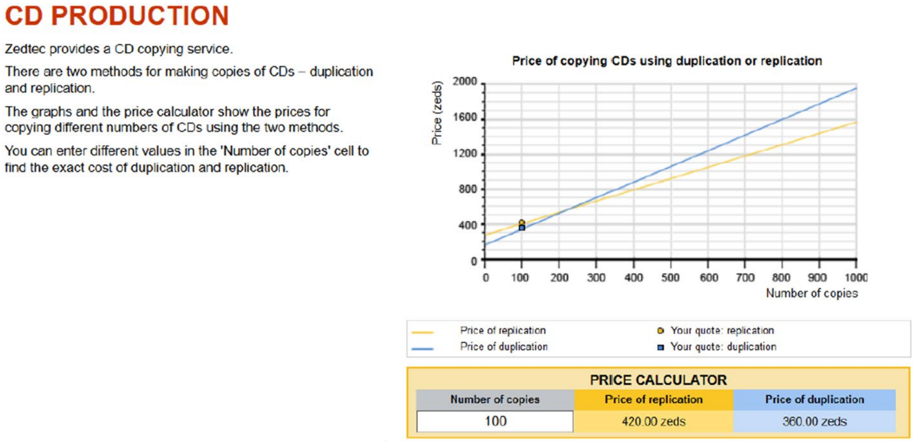
In summary, the purpose of this study is to explore the potential of extracted process data and questionnaire data to advance our understanding of students' outcomes from an international survey. Using three computer-based mathematics tasks from the PISA 2012 cycle, we first investigated the relationship between two specific behavioural indicators (i.e., response times and frequency of actions) and students' item-level performance (i.e., response accuracy) in a three-factor CFA model, following the framework developed by De Boeck & Scalise (2019). We further investigated measurement invariance using multiple-group CFA modelling across the three Scandinavian countries (i.e., Denmark, Sweden, and Norway). Using the best model from the invariance evaluation, we present a latent regression model to evaluate the effect of contextual variables on the latent constructs across the analysed countries. Accordingly, we address the three following research questions, which build upon each other:

1. Does a factor model of the observed measures of performance and process data support evidence for latent constructs (i.e., latent performance, latent speed, and latent exploration behaviour) among the three PISA 2012 math tasks?
2. How do Scandinavian countries differ in the model's parameters for time and frequency of action and the relationship between the latent constructs for math tasks?
3. How do contextual variables (i.e., a student's background, availability, use of ICT resources, and learning conditions for mathematical literacy) relate to the latent factors and across the three countries?

The remaining sections are organized as follows: First, we describe the three computer-based mathematics tasks in PISA 2012, analyse the samples, and provide our analytical framework. In the Results and Discussion sections, we elaborate model

Table 1 Sample size information for each Scandinavian country

Country	n	Female	Native	Language at home is the same as the language of assessment
Denmark	615	322 (52%)	451 (73%)	493 (80%)
Sweden	408	215 (53%)	333 (82%)	322 (79%)
Norway	403	196 (49%)	352 (87%)	354 (88%)
Total	1426	733 (51%)	1136 (80%)	1169 (82%)



CD PRODUCTION

Zedtec provides a CD copying service.

There are two methods for making copies of CDs – duplication and replication.

The graphs and the price calculator show the prices for copying different numbers of CDs using the two methods.

You can enter different values in the 'Number of copies' cell to find the exact cost of duplication and replication.

PRICE CALCULATOR

Number of copies	Price of replication	Price of duplication
100	420.00 zeds	360.00 zeds

Fig. 3 Shared stimuli and interfaces among the three studied items

data fit, measurement invariance, and latent regression coefficient results and discuss the implications and insights of the findings for the field.

Data and methods

Sample

In this study, we adopted public data from PISA 2012, which was administered by the OECD. This PISA data was collected from a representative sample of students aged 15 in each participating education stem. For the purpose of this study, we analysed students' data from Scandinavian countries (i.e., Denmark, Sweden, and Norway). In total, there were 1614 students from this group of countries with available log file data on the three analysed computer-based math tasks. After removing 188 students who had missing values, the final sample comprised 1426 students. Table 1 shows the demographic characteristics of the analysed sample of students by country.

Instrument

From the PISA 2012 computer-based assessment of mathematics, we selected three items from the unit "CD production" for this study. These items were administered on the same test forms and were grouped together in a bundle; they shared the same reading passage, figure, and interaction interface (see Fig. 3). To answer the questions, students were provided with two linear curves. In the textbox, students were allowed to input the value of the x-axis (i.e., number of copies), and the values of the y-axis for the two linear curves were then generated. While working on these tasks, the computer-generated log files collected students' information, such as response accuracy, response time, and interactions between student and computer (e.g., actions of keying words in the textbox and clicking the screen).

These math tasks varied in content and format. While the first item measured mathematical quantity, the other two were related to change and relationship content. The format of the first item (item code: CM015Q01) was multiple choice, and it involved calculating the difference between two linear curves for a specific x-value. The following two items were of the constructed response type. For the second item (item code: CM015Q02D), students were asked to write the equation of the regression function for one linear curve from the figure. For the third item (item code: CM015Q03D), students had to find the value of the x-axis where the two linear curves intersected.

Process indicators

We used the LOGAN R package (Reis Costa & Leoncio, 2019) to extract students' process data (i.e., response time, and frequency of actions) for the three analysed items from the PISA 2012. Table 2 presents a description of each extracted process indicator. More details on the structure of log file data from the PISA and preprocessing analysis can be found in Reis Costa & Leoncio Netto (2022).

In this study, we used binary and ordinal variables for response accuracy. While students only received correct or incorrect answers for the first item (CM015Q01), partial credit was given for the remaining two items (CM015Q02D and CM015Q03D). The response time was the total time of solving the item in minutes. To transform the positively skewed distribution of response time into a symmetric shape, we used the logarithm of response time as the analysed indicator of response speed (van der

Table 2 Description of the extracted process data

Process indicator	Description
Accuracy 1	Binary variable for CM015Q01 item with correct = 1 and incorrect = 0
Accuracy 2	Ordinal variable for CM015Q02D item with correct = 2, partial correct = 1, and incorrect = 0
Accuracy 3	Ordinal variable for CM015Q03D item with correct = 2, partial correct = 1, and incorrect = 0
Response time 1	Total amount of time (in min) student spent on the CM015Q01 item
Response time 2	Total amount of time (in min) student spent on the CM015Q02D item
Response time 3	Total amount of time (in min) student spent on the CM015Q03D item
Frequency of action 1	Number of valid values that the student typed in the "number of copies" box when answering CM015Q01 item
Frequency of action 2	Number of valid values that the student typed in the "number of copies" box when answering CM015Q02D item
Frequency of action 3	Number of valid values that the student typed in the "number of copies" box when answering CM015Q03D item

(1) The computation of the response time is the difference on the amount of time for the "START" and "END" trace log events.
(2) Example of possible valid values in the "number of copies" box are: 1; 20; 500; or 1000

Linden, 2006). The frequency of actions in the tasks was the number of actions of inputting and submitting the values in the "number of copies" textbox. Since the available log file data presented information for each keystroke as a "keyup" event, not as a whole number, we considered each valid value in the textbox a valid interaction. For example, when a student entered a number with three digits (e.g., "100"), three indicators of interactive behaviour were considered for this student (i.e., one for each event: "1," "10," and "100"). The log of the frequency of action indicators was used in the analysis based on De Boeck & Scalise (2019). The latent variable behind frequency of actions was named "Exploration behaviour" because while students can use the textbox to facilitate their understanding regarding the task, it was not a necessary action for task completion. In other words, an item could be solved by other means (e.g., by paper and pencil, since the use of these tools was allowed during the 2012 PISA), for which the computation of the "frequency of actions" would be equal to zero However, when students decided to use the "number of copies" textbox, we

Table 3 Contextual variables extracted from PISA 2012 students' questionnaires

Category	Variable	Type	Measure
Personal background	Gender (ST04Q01)	Categorical	Dummy variable with male = 0 and female = 1
	Index of economic, social, and cultural status (ESCS)	Numeric	PISA 2012 index derived from five indices: highest occupational status of parents, highest educational level of parents, family wealth, cultural possessions, and home educational resources (OECD, 2014)
	Language at home (ST25Q01)	Categorical	An internationally comparable variable computed in PISA 2012 with two categories: (1) language at home is the same as the language of assessment for that student; and (2) language at home is another language (OECD, 2014)
	Immigration background (IMMIG)	Categorical	This index has three categories: (1) native students (those students who had at least one parent born in the country); (2) second generation students (those born in the country of assessment but whose parent(s) were born in another country); and (3) first-generation students (those students born outside the country of assessment and whose parents were also born in another country) (OECD, 2014)
Availability and use of ICT resources	ICT availability at home (ICTHOME)	Numeric	PISA 2012 IRT scales based on the weighted likelihood estimates (WLEs). (OECD, 2014)
	ICT availability at school (ICTSCH)	Numeric	
	ICT use at home for school-related tasks (HOMSCH)	Numeric	
	Use of ICT in mathematics lessons (USEMATH)	Numeric	
	Use of ICT at school (USESCH)	Numeric	
Learning conditions	Experience with applied mathematics tasks at school (EXAPPLM)	Numeric	PISA 2012 IRT scales on OTL content based on the WLEs. (OECD, 2014)
	Experience with pure mathematics tasks at school (EXPUREM)	Numeric	

categorized this as an exploration solution behaviour because it may have helped students in their mathematical thinking.

Contextual variables

Table 3 presents a description of the contextual indicators we selected for our study. Although other variables in the PISA may also be relevant (e.g., attitude towards mathematics or familiarity with mathematical concepts), not all items from the PISA 2012 questionnaire were presented to all students due to the rotated scheme adopted in this edition of the assessment (OECD, 2014). In this study, scales of the availability and use of ICT resources were included in the modelling as a single construct, while other variables were considered single measures. The rates of missing values of contextual variables ranged from 2 to 34.7%. We did not use a multiple imputation approach in this study because we adopted the full-information maximum likelihood in the parameter estimation, which performed well in recovering parameters when the missing rate was under 50% (Lee & Shi, 2021).

Analytic strategy

To answer our research questions, we conducted our analyses in four steps. First, a descriptive analysis of the performance, process, and contextual indicators was performed. Then, a CFA model, following Boeck's framework, was used to analyse the relationship between all observed measures and latent factors (RQ1). For the third step, we evaluated the measurement invariance in the time and action model parameters across the three Scandinavian countries (RQ2). Finally, latent regression models with covariates were analysed to evaluate how contextual factors are related to the latent factors (RQ3).

Descriptive statistics

For each math task, we computed the average, standard deviation, range (i.e., minimum and maximum), skewness, and kurtosis statistics for extracted measures from the process data (i.e., response times and frequency of action), as well as each student's final outcome (i.e., response accuracy). These statistics were also calculated for each of the contextual variables extracted from the PISA 2012 data.

CFA

For research question 1, CFA modelling was conducted to ensure that the processing data in PISA 2012 fit De Boeck & Scalise's (2019) framework (see Fig. 1) by using lavaan (Rosseel, 2012) in R version 4.0.4 (R Core Team, 2021). The marginal maximum likelihood estimator with robust standard errors (i.e., MLR option in lavaan) using a numerical integration algorithm was implemented in the parameter estimation. We evaluated the model data fit indices, comparative fit index (CFI; Bentler, 1990), Tucker–Lewis index (TLI; Bentler & Bonett, 1980; Tucker & Lewis, 1973), root mean square error of approximation (RMSEA; Steiger, 1990), and standardized root mean squared residual (SRMR; Hu & Bentler, 1999). We used two fit thresholds: good (CFI & TLI > 0.95; SRMR < 0.08; RMSEA < 0.06) and moderate (0.90 < CFI & TLI < 0.95; 0.08 < SRMR < 0.10; 0.06 < RMSEA < 0.10).

Based on De Boeck's framework, the mathematical formulation of the three-factor CFA model was as follows: P_{ni}, T_{ni}, and A_{ni} are the observed response correctness, response time, and frequency of actions of a person n ($n = 1, \ldots, N$) to Item i ($i = 1, \ldots, I$). The binary response outcome to the items is scored 1 for correct answers and 0 otherwise. The response time is the total time students spent on each item. The frequency of actions is the number of attempts that a student performed when inserting a valid number of copies in the textbox located in the "Price calculator" box (see Fig. 2) to answer each item. The logarithm function for the transformation of response times and frequency of actions is used to make the positive skewed distributions symmetric. The latent cognitive factors (i.e., performance, exploration behaviour, and speed) were measured by response accuracy, frequency of actions, and response time, respectively.

Thus, we represent person n's response accuracy P_{ni} to Item i measuring cognitive performance θ_n as:

$$P_{ni} = \delta_i + \alpha_i \mathrm{q}_n + \epsilon_{ni}$$

where δ_i and α_i are the intercept and factor loading for Item i, respectively, and ϵ_{ni} is the residual term. Likewise, person n's frequency of actions measuring exploring behaviour ξ_n is defined as:

$$A_{ni} = \nu_i + \lambda_i \xi_n + \varepsilon_{ni}$$

where ν_i and λ_i are the intercept and factor loading for Item i, respectively, and ε_{ni} is the residual term. The logarithm transformed response time $\log\mathrm{T}_{ni}$ is used to measure person n's speed. The three factors correlate with each other, and we also assumed correlated residuals to model the within-item relationship among the process indicators. The residual of response accuracy and that of frequency of actions correlates with the residual of response time, and the residual of response accuracy correlates with the residual of frequency of actions. Notably, for model flexibility, we released De Boeck's model's hypothesis of the causal relationship between indicators. This means that the residuals of indicators within items were correlated rather than having a direct effect.

Response accuracy, logarithm frequency of actions, and logarithm response time were regarded as continuous in this study. Although categorical indicators, the measurement model for response accuracy using a traditional CFA for continuous indicators. We acknowledge that a traditional CFA model is not the most appropriate model in terms of the misspecification for the categorical and count variables since the potential consequence is that the factor loadings are imprecisely estimated; however, fortunately, the intercorrelation is still estimated with good parameter coverage (Li, 2016) when using the marginal maximum likelihood estimator with robust standard errors (MLR). Specific to research question 3, this approach is useful on the modelling of process data and contextual variables which can reveal meaningful correlations between the three latent factors (which are recovered well under MLR, even for categorical and count data) and well accommodate how the latent factors are explained by the covariates. Although De Boeck & Scalise (2019) used weighted least squares (WLSMV) estimation method which does not make assumptions about observed data, WLSMV still assumes the normal latent distribution underlying each categorical indicator and performs worse intercorrelation recovery than MLR when the latent distributions are nonnormal (Li, 2016). Therefore, to obtain accurate intercorrelation and answer the research questions, we adopted MLR and considered response accuracy continuous indicators. Additionally, we used a sandwich estimator in MLR (i.e., an estimator to make standard errors robust for the data not normal; Maydeu-Olivares, 2017) for the standard error computation.

Because the students were sampled with unequal probabilities in PISA's two-stage sampling design (OECD, 2009), we incorporated the sampling weights to establish unbiased estimations by using lavaan.survey (Oberski, 2014) in R version 4.0.4.

In addition to De Boeck and Scalise's joint CFA model for response accuracy, response time, and frequency of actions, we examined the three separate CFA models for response accuracy, response time, and frequency of actions to demonstrate the unidimensionality of measurement models for response accuracy, response time, and frequency of actions, respectively.

Measurement invariance via CFA

For research question 2, we examined measurement invariance by using a multiple-group approach, considering each of the three Nordic countries a group in the CFA model with three latent factors (i.e., mathematic performance, speed, and exploration behaviour). We followed the procedure presented by Marsh et al., (2009) to examine measurement invariance. We compared five models: the configural invariance model, weak invariance model, strong invariance model, strict invariance model, and structural invariance model. The configural invariance model had the same measurement framework (e.g., the same number of factors, indicators, and relationships among factors) but freely estimated all the parameters. The weak invariance model then similarly fixed the factor loadings across the three countries. The strong invariance model similarly fixed the intercepts and factor loadings across the three countries, and the strict invariance model similarly fixed the intercepts, factor loadings, and residual variances across the three countries. The structural invariance model is a baseline model where not only the measurement parameters but also the variance–covariance between factors and residual correlation between factors are invariant. Table 4 provides the status of the parameter constraints for the above models.

Table 4 Constraints of parameters in the measurement invariance models

Model	Covariance between factors	Residual correlation	Residual variance	Intercept	Loading
Configural	Free	Free	Free	Free	Free
Weak	Free	Free	Free	Free	Fixed
Strong	Free	Free	Free	Fixed	Fixed
Strict	Free	Fixed	Fixed	Fixed	Fixed
Structural	Fixed	Fixed	Fixed	Fixed	Fixed

Free indicates that the parameters were freely estimated for each country. Fixed indicates that the parameters were fixed to be the same between countries

Model comparison evaluation The fit indices for the model comparison in the measurement invariance examination were CFI, REMSA, SRMR, Akaike information criterion (AIC; Akaike, 1998), and Bayesian information criterion (BIC; Schwarz, 1978). We conducted a chi-square difference test (Brown, 2015) and compared it to the adjacent complex model to determine what level of invariance was achieved. The adjacent complex model is a model that releases the parameter invariance constraints more than the invariance model. We adopted a conservative significance level of 0.05 for hypothesis testing. When the chi-square test showed that the invariant model did not significantly differ from the adjacent complex model, we selected the invariant model because it is more compact than the adjacent complex model.

Since the chi-square test is influenced by the sample size and complexity of the model (Yuan & Chan, 2016), we also accounted for the difference in the fit indices as additional information to complete the model comparison. In this study, the criteria of this difference were based on Chen's (2007) suggestion; hence, if the invariant model had ΔCFI less than − 0.010, ΔRMSEA less than 0.015, and ΔSRMR less than 0.030 compared to the adjacent complex model, it was selected, even if the chi-square different tests produced significant results.

When the weak invariance model fits better than the configural model, factor variance and covariance can be compared meaningfully between countries. When strong invariance holds, the means of the factors can be meaningfully compared. When strict invariance holds, the reliability of the measures is consistent between countries (Raykov & Marcoulides, 2006). To answer research question 3, at least weak invariance was required because research question 2 entails comparing the factor covariance between Denmark, Sweden, and Norway.

Based on the model that had the best model data fit in the measurement invariance step, we compared the intercorrelation (i.e., correlation between response, time intensity, and action frequency) among Denmark, Sweden, and Norway. A Wald test was

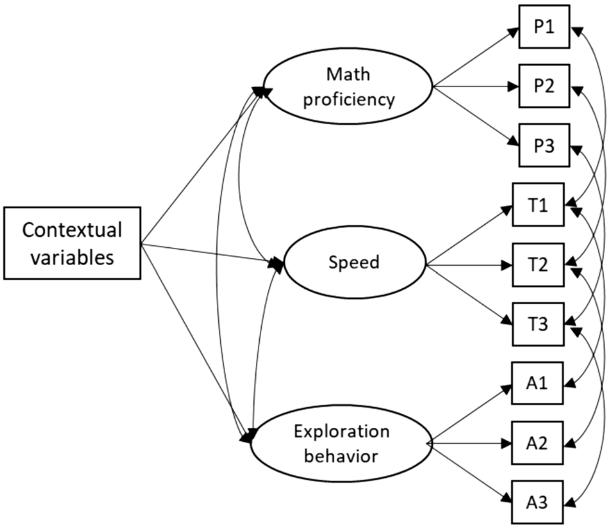


Fig. 4 Latent regression framework for the analysis of students' performance, response times, and frequency of actions in a joint framework with context indicators

conducted to examine the difference in intercorrelation among these three Nordic countries.

Latent regression model with covariates

As long as the variance–covariance among the three factors is meaningful (i.e., at least weak measurement invariance is reached), the covariates can meaningfully explain the variance in the three factors. To answer research question 3, we fitted the invariant CFA model again, but this time, the covariates of a latent variable ICT and two manifest variables concerning OLT were used to predict the three factors (i.e., math performance, speed, and exploration behaviour) while considering each Scandinavian country a group. This is called a latent regression model, where the measured latent variables are predicted by the covariates.

Figure 4 illustrates the proposed model. In this framework, we allowed the three countries to have different structural coefficients; thus, the regression coefficients for Denmark, Sweden, and Norway were freely estimated separately. For the contextual variables, we used the following: (1) a latent ICT variable measured by five items (i.e., ICT available at home, ICT available at school, ICT used at home for school-related tasks, use of ICT in mathematics lessons, and use of ICT at school); and (2) two manifest variables concerning OLT—Experience with Applied Mathematics Tasks at School (ExApplM) and Experience with Pure Mathematics Tasks at School (ExPureM). In addition to the ICT and OLT variables, the covariates included four controlling variables: immigration status, ESCS, language at home, and gender. We tested the significance of the regression coefficients for each contextual variable (e.g., ICT, ExApplM, and ExPureM) by computing the p values to answer research question 3.

Table 5 Descriptive analysis of the indicators in their original metrics

Variable	Mean	SD	Min	Max	Skew	Kurtosis
Accuracy 1	0.61	0.49	0	1	− 0.43	− 1.82
Accuracy 2	0.13	0.46	0	2	3.41	10.34
Accuracy 3	0.5	0.74	0	2	1.08	− 0.32
Response time 1 (min)	1.22	0.75	0.04	9.06	2.54	14.68
Response time 2 (min)	1.61	1.16	0.05	11.01	1.93	7.28
Response time 3 (min)	1.71	1.1	0.03	10.54	1.47	5.31
Frequency of action 1	2.74	6.25	0	105	7.36	83.57
Frequency of action 2	4.92	12.9	0	240	6.76	88.98
Frequency of action 3	28.21	49.41	0	424	2.3	6.81
ICT availability at home	0.28	0.83	− 4.02	2.78	0.77	1.72
ICT availability at school	0.61	0.74	− 2.8	2.83	− 0.29	2.13
ICT use at home for school-related tasks	0.23	0.91	− 2.44	3.73	0.13	3.64
Use of ICT in mathematics lessons	0.43	1.11	− 0.77	2.8	0.53	− 0.61
Use of ICT at school	0.67	0.73	− 1.61	4.11	− 0.05	3.1
Experience with applied mathematics tasks at school	0.32	0.96	− 2.99	3.2	0.46	2.29
Experience with pure mathematics tasks at school	− 0.14	0.99	− 2.73	0.8	− 0.81	− 0.08
Economic, social, and cultural status	0.31	0.85	− 2.71	2.28	− 0.54	0.17

Table 6 Factor loadings in the three-factor CFA model

Variable	Estimate	Standardized estimate	SE	*p*-value
Performance factor				
Accuracy 1 (Ac1)	1	0.507	–	–
Accuracy 2 (Ac2)	0.861	0.444	0.101	0***
Accuracy 3 (Ac3)	2.247	0.751	0.160	0***
Speed factor				
Log response time 1 (Tm1)	1	0.513	–	–
Log response time 2 (Tm2)	1.936	0.787	0.166	0***
Log response time 3 (Tm3)	1.949	0.773	0.171	0***
Exploration behavior factor				
Log frequency of actions 1	1	0.697	–	–
Log frequency of actions 2	1.178	0.656	0.064	0***
Log frequency of actions 3	2.411	0.881	0.119	0***

***$p < 0.001$

Table 7 Extra dependencies in the CFA model

Variable	Estimate	Standardized estimate	SE	*p*-value
Ac1 ↔ Tm1	0.009	0.038	0.008	0.297
Ac2 ↔ Tm2	0.006	0.028	0.008	0.442
Ac3 ↔ Tm3	0.071	0.282	0.012	0***
Fq1 ↔ Tm1	0.076	0.164	0.015	0***
Fq2 ↔ Tm2	0.225	0.404	0.027	0***
Fq3 ↔ Tm3	0.222	0.397	0.038	0***
Ac1 ↔ Fq1	0.058	0.160	0.013	0***
Ac2 ↔ Fq2	0.169	0.345	0.023	0***
Ac3 ↔ Fq3	0.089	0.169	0.051	0.078

Ac1, Ac2, and Ac3 are indicators for response accuracy to items 1, 2, and 3, respectively. Tm1, Tm2, and Tm3 are indicators for log response time to items 1, 2, and 3, respectively. Fq1, Fq2, and Fq3 are indicators for log frequency of actions to items 1, 2, and 3, respectively. The symbol ↔ indicates correlated residuals

***$p < 0.001$

Results

Descriptive analysis

Table 5 presents the means, standard deviations, minimal values, and maximum values for the indicators and covariates. Item 1 was the easiest task, whereas Item 3 demanded the longest response time and required the greatest number of actions in solving the math problem. The skewness of response time to the three items was 2.54, 1.93, and 1.47, respectively, and that of frequency of actions was 7.36, 6.76, and 2.30, respectively. We used log transformation of response times and frequency of actions as indicators of latent speediness and latent exploration behaviour. The skewness of log response time to the three items was − 0.82, − 0.80, and − 1.19, respectively, and that of log frequency of actions was 0.77, 1.31, and 0.55, respectively. All skewness after log transformation ranges from − 2 to + 2, which is considered acceptable for the normal distribution assumption (George & Mallery, 2010). The correlations among indicators and manifest variables can be found in Appendix A. The descriptive statistics of all indicators and covariates for Denmark, Sweden, and Norway, separately, can be found in Appendix B.

CFA Modelling

The model data fit for the CFA model showed CFI = 0.982 (> 0.95); TLI = 0.957 (> 0.95); RMSEA = 0.052 (< 0.08); and SRMR = 0.028 (< 0.08). All the indices demonstrate that De Boeck & Scalise's (2019) framework fit the response data to the computer-based items when solving the math problem in PISA 2012. Table 6 shows the factor loadings of the indicators. All factor loadings were significantly positive for the corresponding factors. The standardized estimates show that the loadings of response accuracy for Item 2 were lower than those for other items. The correlation between math performance and response speed was 0.615, between math performance and exploring behaviour was 0.871, and between response speed and exploring behaviour was 0.573.

Table 7 shows the relations between the residuals of the indicators. All the residual relations were significantly larger than zero, except those relations between response accuracy and response time for Items 1 and 2. When answering Item 3, higher response accuracy required a longer response time. Similar to De Boeck & Scalise's (2019) conclusions, response times and frequency of actions were highly correlated with each other within the items.

In addition to the joint CFA modelling, we have implemented separate CFA models for response accuracy, response time, and frequency of actions. The results indicate that all the model data fit indices showed a perfect fit (i.e., CFI = 1.000; TLI = 1.000; RMSEA = 0.000; and SRMR = 0.000); i.e., the degree of freedom for those measurement models containing only three items is equal to zero. The standardized factor loadings to the three items were 0.721, 0.871, and 0.775 for response accuracy; 0.535, 0.802, and 0.750 for response time; and 0.688, 0.626, and 0.876 for frequency of actions. All standardized factor loadings larger than 0.535 showed sufficient factor loadings for measuring the unidimensional latent variables.

Measurement invariance for the confirmatory factor model

Measurement invariance was examined to answer research question 2. Table 8 shows the comparisons between the configural, weak, strong, strict, and structural invariance models. Although the configural model had the smallest SRMR and the structural model had the smallest BIC, the AIC suggests that the weak invariance model fits the data

Table 8 Model fit indices of the measurement invariance models

Model	CFI	RMSEA	SRMR	AIC	BIC	$\Delta\chi^2$	*p*-value
Configural	**0.97**	0.081	**0.039**	28234	28850		–
Weak	0.968	0.075	0.042	**28231**	28784	16.269	0.18
Strong	0.958	0.078	0.049	28268	28757	60.273	0.000***
Strict	0.952	0.068	0.054	28262	28562	44.847	0.15
Structural	0.949	**0.066**	0.062	28261	**28498**	23.57	0.02*

$\Delta\chi^2$ is the change in chi-square from the one model above to the current reduced model. The bold values indicate the best fit among models under the fit indices.

***$p \leq 0.001$

Table 9 Intercorrelations among math proficiency, speed, and exploration behavior for Denmark, Sweden, and Norway

Correlation	Denmark (N = 615)	Sweden (N = 403)	Norway (N = 408)
Math proficiency ↔ Speed	0.568	0.646	0.612
	[0.503 0.633]	[0.571 0.721]	[0.535 0.689]
Math proficiency ↔ Exploration behavior	0.936	0.861	0.869
	[0.908 0.964]	[0.811 0.911]	[0.821 0.917]
Speed ↔ Exploration behavior	0.566	0.578	0.575
	[0.501 0.631]	[0.498 0.658]	[0.495 0.655]

The numbers in the square brackets are the lower bound and higher bound of 95% confidence interval for the correlation coefficients

best. This indicates that the factor loadings were invariant between Denmark, Sweden, and Norway. The $\Delta\chi^2$ test showed that the weak invariance and configural invariance models did not significantly differ from each other in terms of model data fit. The same conclusion can be derived from the changes in CFI, RMSEA, and SRMR between the weak and configural invariance models. Based on Chen's suggestion (2007), ΔCFI > -0.01, ΔRMSEA < 0.015, and ΔSRMR < 0.03 between the two nested models could be the same performance identified in model data fit. The strong invariance model had a worse model data fit than the weak invariance model, since ΔCFI < − 0.01 and the test of $\Delta\chi^2$ reached the significance of a nominal alpha level of 0.05. Generally, the weak and configural invariance models fit the data best. When comparing the weak and configural invariance models, the weak invariance model is less complex (fewer free parameters) than the configural invariance model. As a result, the weak invariance model was preferred, and it was concluded that the three Nordic countries have consistent factor loadings according to De Boeck & Scalise's (2019) measurement model. That is, response accuracy, response time, and frequency of actions had the same measurement construct across the three Nordic countries.

This result implies that factor covariance among math proficiency, speed, and exploration behaviour can be compared meaningfully between countries. Additionally, the intercorrelations among the three factors for Denmark, Sweden, and Norway (see Table 9) show that all the intercorrelations are positive for the Nordic countries. The higher the math proficiency performance is, the longer the response time and the greater the frequency of actions to solve computer-based math problems. Denmark had a significantly higher correlation between math proficiency and exploration behaviour than Sweden and Norway but a significantly lower correlation between math proficiency and speed than Sweden. Sweden and Norway did not show significant differences in any of the intercorrelations. The correlations between speed and exploration behaviour for the three countries were not significantly different. In summary, Danish students had a larger extent of the positive relation between math proficiency and exploration behaviour than Swedish and Norwegian students but a smaller extent of the positive relation between math proficiency and speed than Swedish students.

Table 10 Coefficients of ICT, OTL, and controlling variables on latent factors measured by response accuracy, response time, and frequency of actions

Effect	Denmark (N = 615)	Sweden (N = 403)	Norway (N = 408)
Regressing mathematic achievement (response accuracy)			
ExApplM	0.039 (0.021)	− 0.031 (0.02)	− 0.019 (0.025)
ExPureM	− 0.009 (0.018)	0.062** (0.021)	0.064** (0.024)
ICT	− 0.243* (0.106)	− 0.094 (0.071)	− 0.089 (0.132)
ESCS	0.075*** (0.02)	0.06* (0.023)	0.098*** (0.028)
Gender	− 0.131*** (0.035)	− 0.12** (0.039)	− 0.018 (0.041)
Language at home	0.03 (0.076)	0.106 (0.138)	− 0.091 (0.095)
Immigration status	− 0.067 (0.049)	-0.063 (0.088)	0.056 (0.059)
Regressing speed (response time)			
ExApplM	0.031 (0.023)	− 0.015 (0.033)	0.002 (0.046)
ExPureM	0.016 (0.02)	0.044 (0.031)	0.112** (0.039)
ICT	− 0.048 (0.147)	− 0.09 (0.12)	0.177 (0.233)
ESCS	0.092*** (0.022)	0.036 (0.035)	0.064 (0.042)
Gender	0.01 (0.037)	− 0.016 (0.048)	0.049 (0.051)
Language at home	− 0.044 (0.102)	− 0.011 (0.13)	0.001 (0.147)
Immigration status	− 0.077 (0.068)	0.054 (0.088)	0.024 (0.105)
Regressing exploratory behavior (frequency of actions)			
ExApplM	0.127* (0.064)	− 0.083 (0.073)	− 0.005 (0.076)
ExPureM	− 0.027 (0.06)	0.173* (0.071)	0.18* (0.073)
ICT	− 0.688 (0.351)	− 0.281 (0.239)	− 0.086 (0.464)
ESCS	0.187** (0.06)	0.037 (0.083)	0.137 (0.082)
Gender	− 0.192 (0.107)	− 0.514*** (0.132)	0.009 (0.119)
Language at home	− 0.238 (0.23)	0.081 (0.39)	0.031 (0.29)
Immigration status	0.001 (0.147)	0.027 (0.249)	− 0.208 (0.203)

The standard errors of the coefficients are in parentheses

ExApplM experience with applied mathematics tasks at school, *ExPureM* experience with pure mathematics tasks at school, *ICT* information communication technology, *ESCS* economic, social, and cultural status

*$p \leq 0.05$

**$p \leq 0.01$

***$p \leq 0.001$

Latent regression with covariates

Since a minimum level of measurement invariance was reached, it was possible to analyse the effects of ICT and OTL after controlling for immigration status, international language at home, SES, and gender across the Scandinavian countries. In a multigroup latent regression framework, as depicted in Fig. 4, the model fit statistics were CFI = 0.926, TLI = 0.904, RMSEA = 0.048, and SRMR = 0.058. This indicates that the multigroup model where the three factors (i.e., math proficiency, speed, and exploration behaviour) were predicted by the selected contextual variables fit the data satisfactorily.

The effects of ICT and OTL on the three factors are shown in Table 10. On math proficiency (measured by response accuracy), ESCS had a positive effect across the three countries. Gender can predict math proficiency (males had higher math proficiency than females) in Denmark and Sweden but not in Norway. Experience with pure math at school (ExPureM) positively predicted math achievement in Sweden and Norway but not in Denmark. The ICT score negatively predicted math achievement in only

Denmark. This means that the greater the ICT use at school and at home in Denmark was, the lower the math achievement on the three analysed items.

Regarding response time, the higher the ESCS was, the greater the response time required to answer items in Denmark. There was no gender difference in response time across the three countries. For Norwegian students, more experience with pure math at school led to longer response times when answering computer-based math problems. However, this positive relationship did not exist in Denmark or Sweden.

Regarding exploration behaviours, measured by frequency of actions, higher ESCS led to more exploration behaviours in Denmark, whereas there was no SES effect in Sweden or Norway. Gender differences in exploration behaviours existed only in Sweden. Experience with pure math at school positively predicted exploration behaviours for Swedish and Norwegian students. For Danish students, experience with application math at school positively predicted exploration behaviours as well. Overall, the more experience the students had with pure or application math tasks at school, the more frequently they interacted with the computer program when solving math problems.

In summary, we found that a higher ESCS consistently led to higher math proficiency. Norway did not have any gender differences in any math proficiency, speed, or exploration behaviours. Experience with math tasks at school led to more exploration behaviours across the three countries. Swedish and Norwegian students who had more experience with pure math tasks at school received higher mathematics achievement scores and tended to interact more frequently with the testing platform when responding to the three math items. When Danish students had higher ICT scores, that is, more available ICT at their school and home, they had lower response accuracy. There was no effect of ICT on the three factors among Swedish and Norwegian students.

Discussion

This study is among the first to employ joint modelling of students' performance, process data indicators, and contextual variables to analyse data from Scandinavian students via the PISA. Focusing on an explorative framework, we have aimed to gain insights into students' characteristics and overt behaviours when answering mathematic tasks. This study may also provide insights into how effective and equitable school practices are in Scandinavia countries in regards to allowing the exploration of the effect of contextual factors on latent constructs (i.e., mathematic performance, speed, and exploration behaviour).

The results from exploring our first research question show that De Boeck & Scalise's (2019) factor model for the observed measures of performance and process data is supported by the evidence for latent constructs (i.e., latent performance, latent speed, and latent exploration behaviour) on the three PISA 2012 math tasks. The results addressing the second research question, built upon the results for the first, show that the weak invariance model fits the data best. Because the weak invariance model allows meaningful comparisons between factor variance and covariance across the different countries, the results for our third research question show that contextual variables may also help explain the variability among the latent factors and across countries. For instance, students from Sweden and Norway who had more exposure to pure math tasks at school were more likely to receive higher mathematics scores and tended to interact more with the item features when completing the three math tasks.

Our findings provide important insights for research on CBAM. For example, they demonstrate the potentialities of extracting process data indicators from log files to better understand how intensively students interact with test features and how they relate to students' performance. We believe our analysis and log file data management for the extraction of the process indicators (i.e., response times and frequency of actions) can be generalized to other interactive items beyond our chosen unit (i.e., CD production) and in different domains. This is also of great interest in the context of international surveys; it can open new avenues for a better understanding of the similarities and differences in the outcomes of these indicators among students from different countries.

Our primary interest in this study, however, was not to advance theory in the field of CBAM but to showcase how different sources of data (i.e., process data and self-reported data from questionnaires) can be exploited to produce new information and knowledge regarding educational outcomes when comparing the PISA data of participating countries. By analysing Scandinavian data, this study adds to the understanding of the similarities and differences in students' performance and behaviour indicators on an international survey. Findings from the OECD (2013) indicate that students' performances on the PISA 2012 computer-based assessment of mathematics across these three countries were not significantly different. Our study, however, has highlighted the differences in how students from Scandinavia approach PISA test items by moving beyond response accuracy.

Concerning the interpretation of our results, we see broad implications for researchers and stakeholders interested in this field. At an assessment level, the increased availability of fine-grained log files from computer-based tests can facilitate a major step towards the measurement of latent constructs in addition to students' performance. At a methodological level, in turn, the use of latent regression modelling can address the relationships among observed measures and these latent factors; this can be attractive to and of great potential for researchers in various contexts.

The findings of our work should be considered in light of several limitations that can be considered opportunities for future research. First, we focused our analyses on two types of process indicators (i.e., response times and frequency of actions), but the number of potential behaviour variables that can be extracted from log files is unlimited. We also defined the number of keyups available in log files from the analysed items as single actions, not the students' final choices (i.e., numerical value included in the textbox). In this case, the absolute number of actions may have been inflated due to this limitation from the PISA 2012 log files. However, we believe that our findings were not affected by this issue, since students received the same data management (i.e., more keyup actions indicated more interaction with test items). Second, the 2012 PISA's rotated scheme for students' questionnaire makes it impossible to incorporate some context variables into the model because of missing patterns. For example, a contextual variable of "Familiarity with Mathematical Concepts" was missing, by design, for all Norwegian students.

Thus, incorporating it into the model was not possible with the available data. Third, response accuracy was analysed as a continuous indicator in this study because the maximum likelihood estimator was not supported for the ordinal data in lavaan. The measurement model for response accuracy should be considered with caution. Adopting the WLSMV estimator is a general way to model ordinal data and allowed us to establish an accurate measurement model for response accuracy. However, WLSMV met the convergence problem of a nonpositive definite covariance matrix between latent variables for Norway when we applied a latent regression model. This problem mirrors previous studies that have reported how the covariance matrix between latent variables is biased when the sample size is relatively small (Li, 2016). A larger sample size can enable the convergence of the models with WLSMV, but such a sample size was not feasible in this study. Although the consequence of using MLR that misspecifies ordinal data is underestimating factor loadings, Li has suggested that MLR outperforms WLSMV in small sample size conditions for the estimation of interfactor correlations. Since the main focus of this study is on the correlations between the three latent factors (which are recovered well under MLR, even for categorical and count data) and how the latent factors are explained by the covariates across the analysed countries, the CFA with MLR estimator has proven to be useful for model mixed indicators. Finally, we tried to capture as many relationships among indicators and factors as possible in our modelling, but we acknowledge that possibly equivalent models may exist. We also acknowledge that there is a possible nonlinear relationship between the three latent constructs and that the intraindividual relationship between time and performance factors might be different for students at lower or higher proficiency levels. These nuances were not captured in our modelling, but they may be an interesting topic for further research.

Other than response time and frequency of actions, the future study should consider the sequence of actions in responding items into the framework of cross-country comparison and measurement invariance examination. The sequence of actions to respond to items has been utilized in methodological studies (Chen, 2020; Han et al., 2019; He et al., 2022; Tang et al., 2020; Ulitzsch et al., 2021). The cross-country comparisons based on the approaches above could be valuable for future studies in the large-scale assessment field.

Due to the exploratory nature of this work, evaluating why some contextual variables significantly relate to specific latent factors (e.g., why Danish students with more ICT available at school and at home had lower response accuracy) is beyond its scope. Notably, however, the PISA data are cross-sectional, which prevents causal assertions based on this study's findings. Further work is needed to better explain the effects of contextual variables among Scandinavian countries.

Appendix A

Correlation between indicators and manifest variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Accuracy 1	1															
2. Accuracy 2	0.214	1														
3. Accuracy 3	0.350	0.386	1													
4. Response time 1 (min)	0.148	0.024	0.129	1												
5. Response time 2 (min)	0.250	0.226	0.405	0.429	1											
6. Response time 3 (min)	0.252	0.212	0.484	0.401	0.602	1										
7. Frequency of action 1	0.170	0.087	0.133	0.241	0.113	0.128	1									
8. Frequency of action 2	0.210	0.340	0.369	0.071	0.391	0.213	0.151	1								
9. Frequency of action 3	0.326	0.247	0.596	0.151	0.312	0.473	0.209	0.322	1							
10. ICTHOME	− 0.073	− 0.079	− 0.039	− 0.102	− 0.089	− 0.082	− 0.028	− 0.072	− 0.060	1						
11. ICTSCH	− 0.009	− 0.069	− 0.056	− 0.033	− 0.042	− 0.015	− 0.011	− 0.057	− 0.022	0.230	1					
12. HOMSCH	0.021	0.010	0.011	0.022	0.035	0.079	− 0.010	− 0.042	0.046	0.191	0.241	1				
13. USEMATH	− 0.048	0.032	− 0.018	− 0.049	− 0.004	0.007	0.031	− 0.011	− 0.026	0.167	0.212	0.206	1			
14. USESCH	− 0.029	− 0.033	− 0.033	− 0.075	− 0.046	− 0.048	− 0.003	− 0.019	− 0.005	0.172	0.271	0.530	0.224	1		
15. EXAPPLM	− 0.001	− 0.004	0.027	0.027	0.040	0.059	0.040	0.019	0.015	0.098	0.088	0.145	0.035	0.070	1	
16. EXPUREM	0.025	0.059	0.104	0.093	0.093	0.101	0.007	0.044	0.078	− 0.011	− 0.023	0.013	0.014	− 0.009	0.478	1
17. ESCS	0.096	0.156	0.193	0.016	0.131	0.160	0.075	0.078	0.090	0.222	0.056	0.126	0.047	0.087	0.013	0.073

ICT information and communication technology, *ICTHOME* ICT Availability at Home, *ICTSCH* ICT Availability at School, *HOMSCH* ICT Use at Home for School-related Tasks, *USEMATH* Use of ICT in Mathematic Lessons, *USESCH* Use of ICT at School, *EXAPPLM* Experience with Applied Mathematics Tasks at School, *EXPUREM* Experience with Pure Mathematics Tasks at School, *ESCS* Economic, social and cultural status

Appendix B

Descriptive analysis of the indicators in their original metrics for Denmark, Sweden, and Norway, separately

See Tables 11, 12, 13

Table 11 Descriptive analysis of the indicators in their original metrics for Denmark

Variable	Mean	SD	Min	Max	Skew	Kurtosis
Accuracy 1	0.61	0.49	0	1	− 0.45	− 1.8
Accuracy 2	0.13	0.45	0	2	3.57	11.44
Accuracy 3	0.48	0.74	0	2	1.17	− 0.2
Response time 1 (min)	1.11	0.62	0.04	5.15	1.77	6.48
Response time 2 (min)	1.6	1.17	0.07	11.01	2.32	10.82
Response time 3 (min)	1.72	1.02	0.05	6.86	1.06	2.21
Frequency of action 1	3.01	6.43	0	105	9.01	117.96
Frequency of action 2	5.67	14.67	0	240	8.18	109.83
Frequency of action 3	29.7	49.58	0	308	2.04	4.2
ICT availability at home	0.35	0.82	− 1.4	2.78	1.04	1.27
ICT availability at school	0.81	0.74	− 2.8	2.83	− 0.02	1.69
ICT use at home for school-related tasks	0.45	0.83	− 2.44	3.73	0.12	5.24
Use of ICT in mathematics lessons	0.69	1.15	− 0.77	2.8	0.25	− 0.87
Use of ICT at school	0.83	0.71	− 1.61	4.11	− 0.04	4.15
Experience with applied mathematics tasks at school	0.31	1.02	− 2.99	3.2	− 0.02	1.69
Experience with pure mathematics tasks at school	− 0.32	1.07	− 2.73	0.8	− 0.61	− 0.59
Economic, social, and cultural status	0.28	0.92	− 2.71	2.28	− 0.49	− 0.05

Table 12 Descriptive analysis of the indicators in their original metrics for Sweden

Variable	Mean	SD	Min	Max	Skew	Kurtosis
Accuracy 1	0.6	0.49	0	1	− 0.39	− 1.85
Accuracy 2	0.12	0.42	0	2	3.6	12.1
Accuracy 3	0.5	0.72	0	2	1.07	− 0.3
Response time 1 (min)	1.3	0.78	0.07	4.76	1.43	2.6
Response time 2 (min)	1.67	1.18	0.05	8.62	1.85	5.86
Response time 3 (min)	1.67	1.09	0.05	6.83	1.1	1.9
Frequency of action 1	2.6	5.59	0	73	7.35	75.12
Frequency of action 2	5.41	11.76	0	77	3.19	11.18
Frequency of action 3	24.91	45.97	0	286	2.16	4.59
ICT availability at home	0.16	0.83	− 2.79	2.78	0.73	1.25
ICT availability at school	0.35	0.77	− 2.8	2.83	− 0.16	1.76
ICT use at home for school-related tasks	0.04	1	− 2.44	3.73	0.71	3.35
Use of ICT in mathematics lessons	− 0.23	0.9	− 0.77	2.8	1.58	1.7
Use of ICT at school	0.45	0.76	− 1.61	4.11	0.4	2.5
Experience with applied mathematics tasks at school	0.46	0.98	− 2.06	3.2	1.18	1.86
Experience with pure mathematics tasks at school	− 0.11	0.95	− 2.73	0.8	− 0.86	0.18
Economic, social, and cultural status	0.25	0.83	− 2.71	2.26	− 0.53	0.2

Table 13 Descriptive analysis of the indicators in their original metrics for Norway

Variable	Mean	SD	Min	Max	Skew	Kurtosis
Accuracy 1	0.61	0.49	0	1	− 0.44	− 1.81
Accuracy 2	0.16	0.51	0	2	3.01	7.59
Accuracy 3	0.54	0.74	0	2	0.97	− 0.53
Response time 1 (min)	1.31	0.88	0.09	9.06	3.44	21.98
Response time 2 (min)	1.56	1.12	0.07	6.77	1.34	2.59
Response time 3 (min)	1.73	1.24	0.03	10.54	2.03	8.96
Frequency of action 1	3.51	6.16	0	53	4.42	24.72
Frequency of action 2	4.5	10.39	0	101	4.34	25.89
Frequency of action 3	30.31	51.81	0	424	2.7	11.12
ICT availability at home	0.28	0.83	− 4.02	2.78	0.46	2.75
ICT availability at school	0.6	0.62	− 2.8	2.83	− 1.15	4.38
ICT use at home for school-related tasks	0.1	0.86	− 2.44	3.73	− 0.42	3.36
Use of ICT in mathematics lessons	0.72	0.97	− 0.77	2.8	0.39	0.04
Use of ICT at school	0.67	0.66	− 1.61	4.11	− 0.56	3.97
Experience with applied mathematics tasks at school	0.18	0.84	− 2.99	3.2	0.56	3.67
Experience with pure mathematics tasks at school	0.08	0.84	− 2.73	0.8	− 0.95	0.47
Economic, social, and cultural status	0.44	0.71	− 2.02	2.27	− 0.42	− 0.21

Acknowledgements
We have to express our appreciation to our colleagues from the Frontier Research in Educational Measurement (FREMO) research group at the Center for Educational Measurement, University of Oslo (CEMO) for sharing their pearls of wisdom with us during the course of this research.

Author contributions
DRC conceptualized the topic, and C-WC conducted the statistical analyses. All authors contributed equally to the development of the ideas and the writing and revising of the manuscript. Both authors read and approved the final manuscript.

Author informations
Dr. Denise Reis Costa is a Postdoctoral research fellow at the Department of Research, Cancer Registry of Norway. Her research interests focus on statistical modeling, psychometrics, machine learning techniques, large-scale assessments, and register data.
Dr. Chia-Wen Chen is Postdoctoral research fellow at the Center for Educational Measurement, University of Oslo. His research interests focus on the item response theory model, computerized adaptive testing, forced-choice items, and Rasch analysis.

Funding
This research received no specific grant from any funding agency in the public, commercial, or organizations for the submitted work.

Availability of data and materials
The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Declarations

Ethics approval and consent to participate
This research discharges its duty imposed by the European Economic Area (EEA)'s general data protection regulation (GDPR) by following Norwegian Centre for Research Data (NSD)'s notification. The Program for International Student Assessment (PISA) data provided by Organization for Economic Co-operation and Development (OECD) contains only aggregated and de-personalized datasets with no possibility of back-tracing to any particular participant. Resultantly, no identifiable personal data were collected or used at any stage of this research.

Consent for publication
We, Denise Reis Costa and Chia-Wen Chen, give our consent for the publication of identifiable details, which can include photograph(s) and/or videos and/or case history and/or details within the text ("Material") to be published in the Large-scale Assessments in Education. We confirm that we have seen and been given the opportunity to read both the Material and the Article to be published by Large-scale Assessments in Education. We understand that Large-scale Assessments in Education may be available in both print and on the internet, and will be available to a broader audience through marketing channels and other third parties. Therefore, anyone can read material published in the Journal. I

understand that readers may include not only educational assessment professionals and scholarly researchers but also journalists and general members of the public.

Competing interests

The authors declared no potential conflicts of financial interest or non-financial interest with respect to the research, authorship, and/or publication of this article.

Received: 31 January 2022 Accepted: 24 January 2023
Published online: 08 February 2023

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In K. T. E. Parzen & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika, 85*(4), 1052–1075. https://doi.org/10.1007/s11336-020-09734-1

De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology, 10*, 1280. https://doi.org/10.3389/fpsyg.2019.01280

Frønes, T. S., Pettersen, A., Radišić, J., & Buchholtz, N. (2020). Equity, equality and diversity in the Nordic model of education—Contributions from large-scale studies. In T. S. Frønes, A. Pettersen, J. Radišić, & N. Buchholtz (Eds.), *Equity, Equality and diversity in the nordic model of education* (pp. 1–10). Springer. https://doi.org/10.1007/978-3-030-61648-9

George, D., & Mallery, M. (2010). *SPSS for windows step by step: A simple guide and reference, 17.0 update* (10th ed.). Pearson.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105.

Han, Z., He, Q., & Von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology, 10*, 2461.

Hansson, Å., & Gustafsson, J. E. (2013). Measurement invariance of socioeconomic status across migrational background. *Scandinavian Journal of Educational Research, 57*(2), 148–166.

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019a). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 369–385.

He, Q., Borgonovi, F., & Paccagnella, M. (2019b). *Using process data to understand adults' problem-solving behaviours in PIAAC: Identifying generalised patterns across multiple tasks with sequence mining*. (No. 205) OECD Education Working Papers

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2022). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12748

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jiang, Y., Cayton-Hodges, G. A., Oláh, L. N., & Minchuk, I. (2023). Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the National Assessment of Educational Progress (NAEP). *Computers & Education, 193*, 104680. https://doi.org/10.1016/j.compedu.2022.104680

Klotzke, K., & Fox, J. P. (2019). Bayesian covariance structure modeling of responses and process data. *Frontiers in psychology, 10*, 1675. https://doi.org/10.3389/fpsyg.2019.01675

Kueh, Y. C., Abdullah, N., Kuan, G., Morris, T., & Naing, N. N. (2018). Testing measurement and factor structure invariance of the physical activity and leisure motivation scale for youth across gender. *Frontiers in psychology, 9*, 1096. https://doi.org/10.3389/fpsyg.2018.01096

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). CRC Press. https://doi.org/10.1201/b16061

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods, 26*(4), 466–485. https://doi.org/10.1037/met0000381

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949.

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*(1), 18–35.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 439–476.

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(3), 383–394.

Nagengast, B., & Marsh, H. W. (2014). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In L. Rutkowski, M.

von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 317–344). CRC Press. https://doi.org/10.1201/b16061

Oberski, D. (2014). lavaan. Survey: An R package for complex survey analysis of structural equation models. *Journal of statistical software, 57*(1), 1–27.

OECD. (2009). *PISA data analysis manual* (SPSS 2). OECD Publishing. https://doi.org/10.1787/9789264056275-en

OECD. (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. https://doi.org/10.1787/9789264190511-en

OECD. (2014). *PISA 2012 technical report*. OECD Publisher. Retrieved June 23, 2021, from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education, 9*(1), 1–17. https://doi.org/10.1186/s40536-020-00092-z

Qiao, X., Jiao, H., & He, Q. (2022). Multiple-group joint modeling of item responses, response times, and action counts with the Conway-Maxwell-Poisson distribution. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12349

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved June 23, 2021, from http://www.R-project.org/

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Reis Costa, D., Bolsinova, M., Tijmstra, J., & Andersson, B. (2021). Improving the precision of ability estimates using time-on-task variables: Insights from the PISA 2012 computer-based assessment of mathematics. *Frontiers in Psychology, 12*, 579128.

Reis Costa, D., & Leoncio, W. (2019). LOGAN: An R package for log file analysis in international large-scale assessments. *R package version 1.0.0*. Retrieved June 23, 2021, from https://cran.r-project.org/web/packages/LOGAN/index.html

Reis Costa, D., & Leoncio Netto, W. (2022). Process data analysis in ILSAs. In T. Nilsen, A. Stancel-Piątak, & J. E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education*. Springer International Handbooks of Education. https://doi.org/10.1007/978-3-030-38298-8_60-1

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Scalise, K., & Clarke-Midura, J. (2018). The many faces of scientific inquiry: Effectively measuring what students do and not only what they say. *Journal of Research in Science Teaching, 55*(10), 1469–1496.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*(7), 371–386. https://doi.org/10.3102/0013189X15603982

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Senese, V. P., Bornstein, M. H., Haynes, O. M., Rossi, G., & Venuti, P. (2012). A cross-cultural comparison of mothers' beliefs about their parenting very young children. *Infant Behavior and Development, 35*(3), 479–488.

Senkbeil, M., & Wittwer, J. (2013). The relationship between computer use and educational achievement. In D. Rutkowski, L. Rutkowski, & M. von Davier (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 287–298). CRC Press. https://doi.org/10.1201/b16061

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*(2), 378–397.

Teig, N., Scherer, R., & Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. *Journal of Research in Science Teaching, 57*(9), 1400–1429.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10. https://doi.org/10.1007/BF02291170

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika, 86*(1), 190–214.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158. https://doi.org/10.1177/1094428102005002001

Wihardini, D. (2016). *An investigation of the relationship of student performance to their opportunity-to-learn in PISA 2012 mathematics: The case of Indonesia*. Berkeley: University of California.

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods, 21*(3), 405–426. https://doi.org/10.1037/met0000080

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.